Causal Inference with Complex Data Structures and Non-Standard Effects

Kwangho Kim

Department of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University

This dissertation is submitted for the joint degree of
Doctor of Philosophy
in
Statistics and Machine Learning

May 2020

Thesis Committee

Prof. Edward Kennedy (co-chair)
Department of Statistics & Data Science,
Carnegie Mellon University

Prof. Larry Wasserman (co-chair)
Department of Statistics & Data Science
and
Machine Learning Department,
Carnegie Mellon University

Prof. Alessandro Rinaldo
Department of Statistics & Data Science
Carnegie Mellon University

Prof. Sivaraman Balakrishnan
Department of Statistics & Data Science
and
Machine Learning Department,
Carnegie Mellon University

Prof. Ashley Naimi
Department of Epidemiology,
University of Pittsburgh

Prof. Jose Zubizarreta
Department of Health Care Policy,
Harvard Medical School
Department of Statistics,
Harvard University

I dedicate this thesis to my loving parents and my brother
for their constant support and unconditional love . . .

Acknowledgements

Undertaking Ph.D at CMU has been a truly life-changing experience for me. It requires me to fully make a commitment to myself more than ever before. However, it would not have been possible to do without the support and guidance that I received from many people.

Firstly, I would like to express my deepest gratitude to my advisor Prof. **Edward Kennedy** for the continuous support of my Ph.D studies and related research, for his patience, motivation, and for his immense knowledge of statistics. Having him as my Ph.D advisor is the luckiest thing that ever happened during my Ph.D studies. He has been supportive since the days I began working with him, not only academically but also emotionally through the rough road to finish my Ph.D. Without his clear guidance and constant feedback, this Ph.D would not have been achievable in any sense. It is whole-heartedly appreciated that his great advice for my study proved monumental towards the success of this Ph.D degree. Edward was the reason why I decided to go to pursue a career in research.

I would like to pay my special regards to my another advisor Prof. **Larry Wasserman** for the invaluable assistance that he has provided during my Ph.D studies. He is a genuine expert in statistics and machine learning; in many cases, just a few words of his email body are enough to make a breakthrough in my research. Most importantly he is a wonderful person. I could not have imagined having a better advisor and mentor for my study in statistical machine learning.

Besides my advisors, my sincere thanks also go to the rest of my thesis committee. I wish to show my gratitude to Prof. **Alessandro Rinaldo** for his scientific advice, knowledge and many insightful discussions. I have really enjoyed our work on topological data analysis. I am indebted to Prof. **Ashley Naimi** who has been helpful in providing advice many times regarding our work on causal inference and my advanced data analysis project. I thank Prof. **Sivaraman Balakrishnan** for his insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives. I also am thankful to my future postdoctoral research advisor with whom I am really looking forward to working, Prof. **Jose Zubizarreta** for insightful discussions and all the great suggestions.

I am also very grateful to Prof. **Barnabas Poczos** for helpful advice and suggestions particularly during the early stages of my Ph.D journey. His immense knowledge of machine learning and wide spectrum of research interests really inspired me. I value his teaching and many insightful discussions for our research.

I will forever be thankful to Dr. **Jisu Kim** who is my friend as well as a great research colleague of mine. He has been my primary resource for getting my statistics questions answered and instrumental in helping me crank out all the hard problems we have been working through. I still think fondly of our time we used to spend together, foraging for some good French desserts while we were working late at night.

Dozens of people have helped and taught me immensely at both the Department of Statistics & Data Science and Machine Learning Department. All the faculty and staff members have been always supportive and caring during my time at CMU. I thank my fellow Ph.D colleagues for the stimulating discussions, for the sleepless nights we were working together before deadlines, and for all the fun we have had in the last four years. There are several people to whom I am especially grateful for their friendship and the warmth they showed to me when I was navigating through the profound personal hardship in the last year of my Ph.D studies. If you had received my personal email recently, you are the one to whom I am thankful here.

Last but not the least, I would like to say a heartfelt thank you to my parents and my little brother for always believing in me and encouraging me to follow my dreams, and supporting me spiritually throughout my Ph.D studies and my life in general.

Abstract

Many modern problems in causal inference have non-trivial complications beyond the classical settings of randomized trials, parametric models, and average treatment effects. Despite their inherent complexities, many recent questions in causal inference are still tackled via overly simplified methods and data structures. My thesis is dedicated to overcoming some of these methodological limitations of classical causal inference, aiming to bridge the gap between methodological development and practice, by effectively harness advanced machine learning tools. My work can be categorized into the following three sub-topics.

a.) *Stochastic interventions for general longitudinal data.* We generalize novel "incremental" intervention effects to accommodate subject dropout in longitudinal studies. Our methods do not require positivity or parametric assumptions, and are less sensitive to the curse of dimensionality. We present efficient nonparametric estimators, showing that they converge at $\sqrt{n}$ rates and yield uniform inferential guarantees. Importantly, we argue that incremental effects are much more efficient than conventional deterministic effects in a novel infinite time horizon setting, where the number of timepoints can grow to infinity.

b.) *Causal effects based on distributional distances.* We have proposed a novel non-standard causal effect based on the discrepancy between unobserved counterfactual distributions (i.e., L_1 distance), in order to provide more nuanced and valuable information about treatment effects than simple mean shifts. We consider single- and multi-source randomized studies, as well as observational studies, and analyze error bounds and asymptotic properties of the proposed estimators. Special difficulties arise due to the non-smoothness of the L_1 distance functional.

c.) *Causal clustering.* We give a novel adaptation of unsupervised learning methods for analyzing treatment effect heterogeneity. Specifically, we pursue an efficient way to uncover subgroup structure in conditional treatment effects by leveraging tools in clustering analysis. We find conditions under which k-means, density-based, and hierarchical clustering algorithms can be successfully adopted into our framework. For k-means causal clustering, we develop a novel estimator that attains fast convergence rates and asymptotic normality of the cluster centers, even under weak nonparametric conditions on nuisance function estimation. Unlike previous studies, our framework can be easily extended to outcome-wide studies.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

> *"There are two types of statisticians: those who do causal inference and those who lie about it."*
>
> – Larry Wasserman

1.1 Background

Statistical causal inference is about estimating what would happen to some response (outcome) when a "cause" of interest is changed or intervened upon, possibly contrary to an observed fact. This is fundamentally distinct from associational questions that are commonly found in standard statistical and machine learning analysis. Associational questions only care about how things are - they do not require us to imagine intervening upon or changing the system we are observing [63]. Consequently, they aim to learn parameters of a distribution from samples drawn of that distribution.

On the other hand, causal analysis goes one step further. They ask how things would have been if something fundamental had changed (or intervened). Therefore, causal questions are inherently *counterfactual*. Thus, its aim is to infer not only beliefs or probabilities under static conditions, but also the dynamics of beliefs under changing conditions where the changes are induced by treatments or external interventions [100].

Causal inference is essential for answering many important questions in health, public policy, economics, and has been increasingly being recognized as a crucial part of science. Some typical examples of important causal questions which cannot be answered with the associational framework alone include: how would survival change under medical treatment A vs. B, or what would be the economic effects of policy X vs. Y? However, it is a common fallacy to conflate association and causation. To mathematically frame such causal problems and distinguish causal inference from associational statistics , we need a counterfactual causal language. In this thesis, we use *potential outcomes* which are the dominant causal language in statistics [118]. For example, suppose that we have data on binary treatment $A \in \{0,1\}$ and outcome Y on units $i = 1,...,n$ where we observe i.i.d samples from $Z = (A,Y) \sim \mathbb{P}$. We concern what might have happened on Y if the treatment A changed, possibly contrary to an observed fact. The potential or counterfactual outcome we would have observed had they received treatment $A = a$ is denoted Y^a for $a \in \{0,1\}$.

It should be stressed that Y represents what we actually observed, while Y^a represents what we would have observed under treatment a. Standard associational studies (i.e., estimating correlation between Y and A) do not cope with dynamics on Y^a unless we are able to travel to a parallel universe. In our example, we never get to observe all potential outcomes in reality; we only can observe either Y^0 or Y^1 at best. This is called the *fundamental problem of causal inference* since we want to contrast potential outcomes, but only see outcomes from actual world, not counterfactual worlds [54].

Despite of this fundamental obstacle, under certain conditions still it is possible to obtain accurate estimates of some important causal parameters. For example, as in our example, to compare population-average outcomes between two treatment levels (i.e., control vs. treated), we formulate the population-level *average treatment effect* (ATE) as

$$\mathbb{E}(Y^1 - Y^0). \tag{1.1}$$

This (the ATE, or the population average effect) represents how the mean outcome in the population would have differed if all versus none were treated, and is arguably one of the most popular causal parameters [63]. Nonetheless, here we remark that other causal parameters (which contrast effects of the treated versus the control in different ways) may instead be of interest in the analysis. For example, researchers may use the risk ratio $\mathbb{E}(Y^1)/\mathbb{E}(Y^0)$ or the odd ratio $\{\mathbb{P}(Y^1 = 1)/\mathbb{P}(Y^1 = 0)\}/\{\mathbb{P}(Y^0 = 1)/\mathbb{P}(Y^0 = 0)\}$ as their target causal parameters, depending on a goal of the scientific investigation (see [63, 64, 71] or the lecture note of [68] for more examples).

1.2 Efficient influence function and nonparametric efficiency bound

Once a causal parameter of interest has been precisely defined and identified, we are ready to develop an estimator and the corresponding inference procedure for that parameter. In this section, we will give a brief introduction about efficient functional estimation based on influence function and nonparametric efficiency bound, which will serve as the core theoretical ingredient in developing estimators across different functionals in this thesis. Before we go on, I declare that the primary sources of this subsection are [64, 65, 63, 71] and the lecture notes used in CMU 36-731, 36-732 [68, 69], and that all the terms, definitions and results are directly borrowed from the resources specified here.

We begin with introducing a doubly robust estimator for the ATE in observational studies. Under the standard identification assumptions [1] the ATE $\psi \equiv \mathbb{E}[Y^1 - Y^0]$ in (4.1) is identified by

$$\psi = \mathbb{E}[\mu_1 - \mu_0] \equiv \mathbb{E}\left\{\mathbb{E}[Y \mid X, A = 1] - \mathbb{E}[Y \mid X, A = 0]\right\},$$

where we define the outcome regression function $\mu_a = \mathbb{E}[Y \mid X, A = a]$.

In what follows, we define a *doubly robust estimator* by

$$\widehat{\psi}_{\text{dr}} = \mathbb{P}_n \left\{ \left[\frac{A}{\widehat{\pi}(X)} - \frac{1-A}{1-\widehat{\pi}(X)} \right] [Y - \widehat{\mu}_A(X)] + [\widehat{\mu}_1(X) - \widehat{\mu}_0(X)] \right\}. \tag{1.2}$$

The doubly robust estimator (1.2) is known to be an efficient, model-free estimator for the identified target parameter (the ATE) compared to other estimators (e.g., regression plug-in and inverse probability weighting estimators) in using nonparametric models when we do not have any substantive information on both of the exposure and outcome processes, as it can be $\sqrt{n}$-consistent and asymptotically normal even when the nuisance functions μ_a, π are estimated flexibly at slower than $\sqrt{n}$ rates [e.g., 63, Theorem 4.5]. Given the superiority of this doubly robust estimator, the following questions naturally arise: 1) would it be possible to assess optimality of the doubly robust estimator in any way, as we did for parametric models using the Cramér-Rao bound argument? 2) what is the general way to construct such efficient, model-free estimators for a given parameter (beyond the ATE)? To address these questions, we shall introduce the influence function, which is a foundational object of statistical theory that allows us to characterize a wide range of estimators with favorable theoretical properties and their efficiency. There are two notions of the influence function: one for estimators and the other for parameters. To distinguish these two cases we will call the latter, which corresponds to parameters, influence curves as in for example, [14, 65] [2].

First, we give a definition of influence curves. It was first introduced by [45] and studied to provide a general solution to find approximation-by-averages representation for a functional statistic. We only consider nonparametric models here.

[1] By standard identification assumptions, here I refer to *consistency*, *no unmeasured confounding*, and *positivity* assumptions. See, for example, [63, 52] for more details and full definitions. We acknowledge that there exist other identification strategies one might consider for causal inference in non-experimental settings as well.

[2] However, the terms ‘influence curve’ and ‘influence function’ are used interchangeably in many cases.

Suppose that we are given a target functional ψ. For a nonparametric model $\mathbb{P}$, let $\{\mathbb{P}_\varepsilon, \varepsilon \in \mathbb{R}\}$ denote a smooth parametric submodel for $\mathbb{P}$ with $\mathbb{P}_{\varepsilon=0} = \mathbb{P}$. A typical example of a parametric submodel is given by $\{\mathbb{P}_\varepsilon : p_\varepsilon(z) = p(z)(1+\varepsilon s(z))\}$ for some mean-zero, uniformly bounded function s. Then the *influence curve* for parameter $\psi(\mathbb{P})$ is defined by any mean-zero, finite-variance function $\phi(\mathbb{P})$ that satisfies the following *pathwise differentiability*,

$$\frac{\partial}{\partial\varepsilon}\psi(\mathbb{P}_\varepsilon)\bigg|_{\varepsilon=0} = \int \phi(\mathbb{P})\left(\frac{\partial}{\partial\varepsilon}\log d\mathbb{P}_\varepsilon\right)\bigg|_{\varepsilon=0} d\mathbb{P}. \tag{1.3}$$

The above pathwise differentiability implies that our target parameter ψ is smooth enough to admit a von Mises expansion: for two distribution $\mathbb{P}, \mathbb{Q}$

$$\psi(\mathbb{Q}) - \psi(\mathbb{P}) = \int \phi(\mathbb{Q}) d(\mathbb{Q}-\mathbb{P}) + R_2(\mathbb{Q},\mathbb{P}) \tag{1.4}$$

where R_2 is a second-order remainder. Therefore, the influence curve also corresponds to the functional derivative in a Von Mises expansion of ψ.

One can obtain the classical Cramér-Rao lower bound for each parametric submodel $\mathbb{P}_\varepsilon$; the Cramér-Rao lower bound for $\mathbb{P}_\varepsilon$ is $\psi'(\mathbb{P}_\varepsilon)^2/\mathbb{E}(s_\varepsilon^2)$ where $\psi'(\mathbb{P}_\varepsilon) = \frac{\partial}{\partial\varepsilon}\psi(\mathbb{P}_\varepsilon)\big|_{\varepsilon=0}$ and $s_\varepsilon = s_\varepsilon(z) = \frac{\partial}{\partial\varepsilon}\log d\mathbb{P}_\varepsilon\big|_{\varepsilon=0}$. The asymptotic variance of any nonparametric estimator is no smaller than the supremum of the Cramér-Rao lower bounds for all parametric submodel, and it is known that under the above pathwise differentiability condition the greatest such lower bound is given by

$$\sup_{\mathbb{P}_\varepsilon} \frac{\psi'(\mathbb{P}_\varepsilon)^2}{\mathbb{E}(s_\varepsilon^2)} \le \mathbb{E}(\phi^2).$$

Therefore, $\mathbb{E}(\phi^2) = \text{var}(\phi)$ is the nonparametric analog of the Cramér-Rao lower bound, and we call the influence curve that attains the above bound the *efficient influence curve*. The efficient influence curve gives the efficiency bound for estimating ψ. In parametric models, more than one influence curves may exist. On the other hand in nonparametric model, the influence curve is unique. However, the efficient influence curve is always unique in any cases.

Once the efficient influence curve is known, no estimator can be more efficient than $\hat{\psi}(\mathbb{P})$ such that

$$\sqrt{n}(\hat{\psi} - \psi) \rightsquigarrow N(0, var(\phi)) \tag{1.5}$$

as var(ϕ) serves to be our nonparametric efficiency bound. In (1.5), we call ϕ the (efficient) *influence function* for the estimator $\hat{\psi}$ [3]. For each nonparametric estimator, the efficient

[3] In fact, influence curves themselves are the putative influence functions.

influence function, if exists, is almost surely unique, so in this sense the influence function contains all information about an estimator's asymptotic behavior. In other words, if we know the influence function for an estimator, we know its asymptotic distribution and can easily construct confidence intervals and hypothesis tests.

Characterizing the influence curves is crucial not only to give the efficiency bound for estimating ψ, thus providing a benchmark against which estimators can be compared, but probably more importantly, to construct estimators with very favorable properties, such as double robustness or general second-order bias which amounts to be a consequence of the pathwise differentiability (1.4). In fact, we can find an (asymptotically linear) estimator that satisfies (1.5) by solving appropriate estimating equations using the influence curves. Particularly Chapter 2 and Chapter 4 of the thesis contains examples developing a novel efficient, model-free estimator based on the efficient influence curve of the target parameter.

Back to the doubly robust estimator (1.2), let us consider $\widehat{\psi}^1_{\text{dr}} = \mathbb{P}_n\left\{{\varphi^1}_{\text{dr}}(Z;\hat{\eta})\right\}$ where

$$ {\varphi^1}_{\text{dr}}(Z;\eta) = \frac{A}{\pi(X)}\left[Y-\mu_A(X)\right]+\mu_1(X) $$

and $\eta = (\pi,\mu)$. Then it can be shown that

$$ \widehat{\psi^1}_{\text{dr}} - \psi^1 = \mathbb{P}_n({\varphi^1}_{\text{dr}} - \psi^1) + o_{\mathbb{P}}(1/\sqrt{n}) $$

where $\psi^1 = \mathbb{E}[\mu_1]$ (Section 3, [65]). Hence, the efficient influence function of the estimator $\widehat{\psi}^1_{\text{dr}}$ is ${\varphi^1}_{\text{dr}} - \psi^1$ by definition.

On the other hand, one may also show that the efficient influence curve for the parameter ψ^1 is given by

$$ \frac{A}{\pi(X)}\left[Y-\mu_A(X)\right]+\mu_1(X)-\psi^1 $$

which is exactly the same quantity with ${\varphi^1}_{\text{dr}} - \psi^1$. Hence, for the ordinary doubly robust estimator, the efficient influence curve for the target parameter ψ^1 coincides with the efficient influence function for the estimator. In this case we can see that indeed there is a deep connection between the estimator for a given target functional, and the corresponding influence function.

Finally we remark that for complicated functionals pretending discrete space on Z can facilitate our procedure to characterize influence curves. For example, assuming that our unit space is discrete, the *influence curve* $\phi(\mathbb{P})$ for the functional $\psi(\mathbb{P})$ can be defined by

$$ \phi(\mathbb{P}) = \frac{\partial}{\partial\varepsilon}\psi((1-\varepsilon)\mathbb{P}+\varepsilon\delta_z)\Big|_{\varepsilon=0^+} = \lim_{\varepsilon\to 0^+}\frac{\psi((1-\varepsilon)\mathbb{P}+\varepsilon\delta_z)-\psi(\mathbb{P})}{\varepsilon} \qquad (1.6) $$

where we let δ_z be the Dirac measure at $Z = z$. This definition is equivalent to the *Gateaux derivative* of ψ at $\mathbb{P}$ in direction of point mass $(\delta_z - \mathbb{P})$ (see, for example, Chapter 5 in [14]).

For more details for nonparametric efficiency theory and influence functions, we refer to [64, 65, 71, 137, 130] and references therein.

1.3 Challenges for Modern Causal Inference

Despite the great importance of causal inference in modern science, still a lot of work in causal inference rely on randomized trials, parametric models, average effects, and overly simplified data structures. Many modern problems have non-trivial complications outside of these classical settings. For example, in observational studies developing fully nonparametric estimators beyond simple data structure $Z = (X,A,Y)$ is very challenging (e.g., with time-varying exposure) and typically positivity conditions which require everyone to have some nonzero probability of receiving each treatment are unlikely to hold. Moreover, in some cases the standard ATE is not enough to convey valuable information to policy-makers (e.g., in the presence of substantial effect heterogeneity).

There has been a growing interest in novel methods of causal inference for complex data structure and non-standard effects, hoping to cope with these and other challenges that arise in modern problems. We enumerate some of important challenges in modern causal inference and delineate our approach to each them in subsequent subsections.

1.3.1 Causal inference for complex longitudinal data

Modern longitudinal data, where individuals are exposed to varying treatment levels over time, is more complex than point exposure studies with a single timepoint. The simplest data structure in longitudinal studies can be described by

$$Z = (X_1, A_1, Y_1, X_2, A_2, Y_2, ..., X_T, A_T, Y_T) \tag{1.7}$$

, where T is the number of timepoints in the study. By virtue of the recent advancement in technology a capability of collecting data has been enormously enhanced, and consequently many longitudinal studies have been proposed, typically with very large T (sometimes of the same order of sample size) (see, for example, [81, 34, 77]).

However, such studies introduce numerous statistical challenges that remain largely unaddressed. First, conventional deterministic interventions (fixed or dynamic) that are often invoked for longitudinal causal studies reply on untenable positivity assumptions, which require every subject to have a nonzero chance of receiving each available treatment at every time point. Even if positivity is only nearly violated, the finite-sample behavior of many common estimators can be severely damaged. Second, even under positivity, longitudinal studies are especially prone to the curse of dimensionality, since exponentially many samples are needed to learn about all treatment trajectories. These issues only worsen when the number of timepoints or covariates increases. Third, it is very common to have multiple right-censored outcomes in longitudinal data. The right censoring (i.e. dropout) may happen during data collection or during the experiment, especially when human subjects are involved. In the presence of dropout, the tuple (X,A,Y) is no longer observable after a certain timepoint. Incorporating dropout events can add much complexity to the data structure. These and other issues have brought new attention to the development of novel methodological framework with which we can effectively perform causal analysis for complex modern data structure.

1.3.2 Non-standard causal effects

Non-trivial mean-zero effects. Consider the case where the ATE may be less useful. For a binary treatment $A \in 0,1$ and outcome $Y \in \mathbb{R}$, suppose that $(A,Y) \sim \mathbb{P}$ and that $Y^0 = 0$ but $\mathbb{P}(Y^1 = 1) = \mathbb{P}(Y^1 = -1) = 1/2$. Then the ATE is exactly zero. Should policy makers conclude that treatment really has no impact? This may be misleading, since the treatment yields extreme harms and extreme benefits to half the population. As seen in this illustration, there are often times when mere average effects reveal potentially less valuable information about how treatment works on outcomes. Therefore, more nuanced measure of treatment effects can be needed.

Heterogeneity in treatment effects. The ATE is a measure used to compare population-average outcomes. However, subgroups of units often show considerable heterogeneity of response toward the same treatment, which can be masked by the ATE. Identifying treatment effect heterogeneity and corresponding subgroups is of great importance in policy evaluation, drug development, and health care service, and has generated growing recent interest.

The most popular approach for studying effect heterogeneity targets the conditional average treatment effects. Various methods have been proposed for this task. However, most existing methods are afflicted with some common limitations. First, many methods are limited to specific supervised learning techniques to derive a partition (subgroup structure) of unit space. Second, many methods rely on unrealistic parametric assumptions. Finally and perhaps most importantly, existing methods are not easily extendable to outcome-wide studies where treatment effects are assessed over numerous outcomes. This conflicts with the fact that a growing number of recent studies seek to adopt outcome-wide approaches, possibly with very many treatment options.

1.3.3 Contribution of the thesis

In this thesis, we develop novel methodological approaches to address each of the above challenges. Specifically,

1. For causal inference with longitudinal data, we generalize incremental interventions to accommodate subject dropout. We provide an identifying expression for incremental effects when dropout is conditionally ignorable (i.e., under a time-varying missing-at-random assumption), still without requiring (treatment) positivity, and derive the nonparametric efficiency bound for estimating such effects. Then we present efficient nonparametric estimators, showing that they converge at fast parametric rates and yield uniform inferential guarantees, even when nuisance functions are estimated flexibly at slower rates. Importantly, in this work we also study the relative efficiency of incremental effects to more conventional deterministic effects in a novel infinite time horizon setting, where the number of timepoints can grow to infinity with sample size. Specifically, we show that our incremental effects can yield near-exponential efficiency gains in this setup. Finally, we apply our methods to study the effect of low-dose aspirin on pregnancy outcomes. Chapter 2 is devoted to this work.

2. In order to provide more nuanced and valuable information about treatment effects than the ATE, we consider estimating causal effects based on the discrepancy between unobserved counterfactual distributions. Continuing the illustrating example with binary treatments, we let Q^0, Q^1 be the two counterfactual outcome distributions for the binary treatments. Then in our setting, a causal effect can be defined in terms of the L_1 distance D_1 between Q^0, Q^1, i.e. $D_1(Q^0, Q^1)$. We provide a novel way to estimate each of the counterfactual outcome distributions for efficient estimation of our target functional $D_1(Q^0, Q^1)$. We consider single- and multi-source randomized studies, as well as observational studies, and analyze error bounds and asymptotic properties of the proposed estimators. We further propose methods to construct confidence intervals for the unknown mean distribution distance. Our proposed method can be always used jointly with the ATE, as a first step in assessing whether there is effect modification beyond a mean shift; for instance, when the ATE is nearly zero but $D_1(Q^0, Q^1)$ is large, we should be cautious before making a decision based on the former. Chapter 3 is devoted to this work.

3. As to analysis of the treatment effect heterogeneity, we give a novel adaptation of unsupervised learning methods. Specifically, we pursue an efficient way to uncover subgroup structure in conditional treatment effects by leveraging tools in clustering

analysis. We find conditions under which k-means, density-based, and hierarchical clustering algorithms can be successfully adopted into our framework. Particularly for k-means causal clustering, we develop an estimator based on nonparametric efficiency theory that attains fast convergence rates to the true cluster centers, under weak nonparametric conditions on nuisance function estimation. This requires novel techniques due to the non-smoothness of the minimizer of the k-means risk. Surprisingly, we give conditions for asymptotic normality of the cluster centers. Chapter 4 is devoted to this work.

1.4 Thesis Organization

The thesis is organized as follows. In Chapter 2, we propose a more comprehensive form of stochastic dynamic intervention effects to accommodate subject dropout, and study the relative efficiency of incremental effects to more conventional deterministic effects in a novel infinite time horizon setting. In Chapter 3 we study a novel non-standrad causal effect based on the discrepancy between unobserved counterfactual distributions using the non-smooth L_1 distance. In Chapter 4 we propose Causal Clustering, a novel framework for the heterogeneous treatment effect analysis, where we pursue an efficient way to uncover subgroup structure in conditional treatment effects by leveraging tools in clustering analysis. Chapter 5 concludes with further remarks on future work.

Chapter 2

Incremental Intervention Effects in Studies with Dropout and Many Timepoints

2.1 Introduction

Causal inference has long been an important scientific pursuit, and understanding causal relationships is essential across many disciplines. However, for practical and ethical reasons, causal questions cannot always be evaluated via experimental methods (i.e., randomized trials), making observational studies the only viable alternative. Further, when individuals can be exposed to varying treatment levels over time, collecting appropriate longitudinal data is important. To that end, recent technological advancements that facilitate data collection are making longitudinal studies with a very large number of time points (sometimes of the same order of sample size) increasingly common [e.g., 81, 34, 77].

The increase in observational studies with detailed longitudinal data has also introduced numerous statistical challenges that remain unaddressed. For longitudinal causal studies, two analytic frameworks are often invoked: *deterministic fixed interventions* [108, 112, 51], in which all individuals are assigned to a fixed exposure level over all time-points; and *deterministic dynamic interventions* [98, 110] in which, at each time, treatment is assigned according to a fixed rule that depends on past history. In the real world, the fixed deterministic interventions might not be of practical interest since the treatment is typically not applied uniformly [67].

Generally, deterministic interventions (fixed or dynamic) rely on the positivity assumption which requires every unit to have a nonzero chance of receiving each of the available treatments at every time point. If the positivity assumption is violated, the causal effect defined under deterministic (fixed or dynamic) interventions will be no longer identifiable. Even under positivity, longitudinal studies are especially prone to the curse of dimensionality, since exponentially many samples are needed to learn about all treatment trajectories. These issues only worsen when the number of timepoints or covariates increases. Thus, due to a lack of analytic methods for such longitudinal data, researchers are often forced to either rely on strong parametric assumptions, or forego the estimation of causal effects altogether [e.g. 81].

Recently, [67] has proposed a novel *incremental intervention effects* which quantify the effect of shifting treatment propensities, rather than effects of setting treatment to fixed values. An incremental intervention is a stochastic intervention in that it depends on unit characteristics and is random at each timepoint [see 150, 27, 46, 96, as prior works on stochastic interventions whose setup is relevant to our study]. Importantly, incremental effect estimators do not require positivity, and can still achieve $\sqrt{n}$ rates regardless of the number of timepoints, even when flexible nonparametric methods are used. Despite these strengths, the method has not been adapted to general longitudinal studies, where multiple right-censored outcomes are common (particularly for human subjects).

In this paper we propose a more comprehensive form of incremental intervention effects that accommodate not only time-varying treatments, but time-varying outcomes subject to right censoring (i.e., dropout). We provide an identifying expression for incremental effects when dropout is conditionally ignorable, still without requiring (treatment) positivity, and derive the nonparametric efficiency bound for estimating such effects. We go on to present efficient nonparametric estimators, showing that they converge at fast rates and give uniform inferential guarantees, even when nuisance functions are estimated flexibly at much slower rates with flexible machine learning tools under weak conditions. Importantly, we study the relative efficiency of incremental effects to more conventional deterministic effects in a novel infinite time horizon setting, where the number of timepoints can grow with sample size to infinity. We specifically show that incremental effects can yield near-exponential gains in this setup. Finally we conclude with a simulation study and apply our methods to a longitudinal study of the effect of low-dose aspirin on pregnancy outcomes to demonstrate the effectiveness of our method.

2.2 Setup

We consider a study where for each subject we observe covariates $X_t \in \mathbb{R}^d$, treatment $A_t \in \mathbb{R}$, and outcome $Y_t \in \mathbb{R}$, with all variables allowed to vary over time, but where subjects can drop out or be lost to follow-up. In particular, we observe a set of i.i.d samples $(Z_1,...,Z_n)$ from a probability distribution $\mathbb{P}$ where, for those subjects who remain in the study up to the final timepoint $t=T$, we observe

$$Z=(X_1,A_1,Y_1,X_2,A_2,Y_2,...,X_T,A_T,Y_T).$$

But in general we only get to observe

$$Z=(X_1,A_1,R_2,R_2(Y_1,X_2,A_2),...,R_T,R_T(Y_{T-1},X_T,A_T),R_{T+1},R_{T+1}Y_T) \tag{2.1}$$

with $R_t=\mathbb{1}\{$ still in the study at time t $\}$ an indicator for whether the subject contributes data at time t. We write $R_t(Y_{t-1},X_t,A_t)$ as a shorthand notation of $(R_tY_{t-1},R_tX_t,R_tA_t)$, so in the missingness process that we consider subjects can drop out at each time after the measurement of covariates/treatment. This is motivated by the fact that this is likely the most common type of dropout, since outcomes Y_t at time t are often measured together with or just prior to covariates X_{t+1} at time $t+1$. As we consider a monotone dropout (i.e., right-censoring) process, R_t is non-increasing in time t, i.e.,

$$\begin{cases} R_t=1 \Rightarrow & (R_1,...,R_{t-1})=\mathbf{1} \\ R_t=0 \Rightarrow & (R_{t+1},...,R_T)=\mathbf{0}, \end{cases}$$

where $\mathbf{0},\mathbf{1}$ are vectors of zeros and ones respectively. Thus our data structure Z is a chain with t-th component

$$\{R_t,R_t(Y_{t-1},X_t,A_t)\}$$

for $t=1,...,T+1$ where $R_1=1$ and we do not use Y_0 or X_{T+1},A_{T+1}. Although we suppose each subject's dropout will occur before the t-th stage, our data structure also covers the case when the dropout will occur after the t-th stage because in that case we can write

$$\{R_t(Y_{t-1},X_t,A_t),R_{t+1}\}$$

as the t-th component of our chain, and the general structure remains the same.

For simplicity, we consider binary treatment in this paper, so that the support of each A_t is $\mathscr{A}=\{0,1\}$. We use overbars and underbars to denote all the past history and future event of a variable respectively, so that $\overline{X}_t=(X_1,...,X_t)$ and $\underline{A}_t=(A_t,...,A_T)$ for example. We also write $H_t=(\overline{X}_t,\overline{A}_{t-1},\overline{Y}_{t-1})$ to denote all the observed past history just prior to treatment at time t, with support $\mathscr{H}_t$. Finally, we use lower-case letters a_t,h_t,x_t to represent realized values for A_t,H_t,X_t, unless stated otherwise.

Now that we have defined our data structure we turn to our estimation goal, i.e., which treatment effect we aim to estimate. We use $Y_t^{\overline{a}_t}$ to denote the potential (counterfactual) outcome at time t that would have been observed under a treatment sequence $\overline{a}_t = (a_1, ..., a_t)$ (note we have $Y_t^{\overline{a}_T} = Y_t^{\overline{a}_t}$ as long as the future cannot cause the past). In longitudinal causal problems it is common to pursue quantities such as $\mathbb{E}(Y_t^{\overline{a}_t})$, i.e., the mean outcome at a given time under particular treatment sequences $\overline{a}_t$; for example one might compare the mean outcome under $\overline{a}_t = \mathbf{1}$ versus $\overline{a}_t = \mathbf{0}$, which represents how outcomes would change if all versus none were treated at all times. However identifying these effects requires strong positivity assumptions (i.e., that all have some chance at receiving every treatment at every time), and estimating these effects often requires untenable parametric assumptions when there are more than a few timepoints.

Following [67] we instead consider incremental intervention effects, which represent how mean outcomes would change if the odds of treatment at each time were multiplied by a factor δ (e.g., $\delta = 2$ means odds of treatment are doubled). Incremental interventions shift propensity scores rather than impose treatments themselves; they represent what would happen if treatment were slightly more or less likely to be assigned, relative to the natural/observational treatment. There are a number of benefits of studying incremental intervention effects: for example, positivity assumptions can be entirely and naturally avoided; complex effects under a wide range of intensities can be summarized with a single curve in δ, no matter how many timepoints T there are; and they more closely align with actual intervention effects than their fixed treatment regime counterparts. We refer to [67] for more discussion and details.

Formally, incremental interventions are dynamic stochastic interventions where treatment is not assigned based on the observational propensity scores $\pi_t(h_t) = \mathbb{P}(A_t = 1 \mid H_t = h_t)$; instead these propensity scores are replaced by new interventional propensity scores given by

$$q_t(h_t; \delta, \pi_t) = \frac{\delta \pi_t(h_t)}{\delta \pi_t(h_t) + 1 - \pi_t(h_t)} \tag{2.2}$$

to ensure the odds of treatment are multiplied by δ. We denote potential outcomes under the above intervention as $Y_t^{\overline{Q}_t(\delta)}$ where $\overline{Q}_t(\delta) = \{Q_1(\delta), ..., Q_t(\delta)\}$ represents draws from the conditional distributions $Q_s(\delta) \mid H_s = h_s \sim \text{Bernoulli}\{q_s(h_s; \delta, \pi_s)\}$, $s = 1, ..., t$. We often drop δ and write $Q_t = Q_t(\delta)$ when the dependence is clear from the context. Note here we use capital letters for the intervention indices since they are random, as opposed to $Y_t^{\overline{a}_t}$ where the intervention is deterministic. Therefore in this paper we aim to estimate the mean counterfactual outcome

$$\psi_t(\delta) = \mathbb{E}\left(Y_t^{\overline{Q}_t(\delta)}\right)$$

for any $t \leq T$. This goal is different from [67] in that we allow varying outcomes over time and dropout/right-censoring. Thus in the next section we describe the necessary conditions for identifying $\psi_t(\delta)$ in the presence of dropout.

2.3 Identification

In this section, we will give assumptions under which the entire marginal distribution of the resulting counterfactual outcome $Y_t^{\overline{Q}_t(\delta)}$ is identified. Specifically, we require the following assumptions for all $t \leq T$.

Assumption A1. $Y = Y^{\overline{a}_T}$ *if* $\overline{A}_T = \overline{a}_T$

Assumption A2-E. $A_t \perp\!\!\!\perp Y^{\overline{a}_T} \mid H_t$

Assumption A2-M. $R_t \perp\!\!\!\perp (\underline{X}_t, \underline{A}_t, Y) \mid H_{t-1}, A_{t-1}, R_{t-1} = 1$

Assumption A3. $\mathbb{P}(R_t = 1 \mid H_{t-1}, A_{t-1}, R_{t-1} = 1)$ *is bounded away from 0 a.e.* $[\mathbb{P}]$

Assumptions (A1) and (A2-E) correspond to consistency and exchangeability conditions respectively, which are commonly adopted in causal inference problems. Consistency means that the observed outcomes are equal to the corresponding potential outcomes under the observed treatment sequence, and would be violated in settings with interference, for example. Exchangeability means that the treatment and counterfactual outcome are independent, conditional on the observed past (if there were no dropout), i.e., that treatment is as good as randomized at each time conditional on the past. Experiments ensure exchangeability hold by construction, but in observational studies it cannot be justified so in general we require sufficiently many relevant adjustment covariates (H_t in our case) to be collected.

In this paper, we additionally require assumptions (A2-M) and (A3) because of the missingness/dropout. (A2-M) is a time-varying missing-at-random assumption, ensuring that dropout is independent of the future (and underlying missing data values), conditioned on the observed history up to the current time point. This would be a reasonable assumption if we can collect enough data to explain the dropout process, so we can ensure that those who dropout look like those who do not, given all past observed data. (A3) is a positivity assumption for missingness, meaning that each subject in the study has some non-zero chance at staying in the study at the next timepoint. This would be expected to hold in many studies, but may not if some subjects are 'doomed' to drop out based on their specific measured characteristics. Note that assumptions (A2-M) and (A3) also appear in more classical works on dealing with missing data [e.g. 115, 114].

Importantly, we do not need any positivity conditions on the propensity scores, since we are targeting incremental effects as defined in (2.2) rather than more common deterministic effects. The next result gives an identifying expression for the incremental effect under the above assumptions.

Theorem 2.3.1. *Suppose identification assumptions (A1) - (A3) hold. Then the incremental effect on outcome Y at time t with given value of* $\delta \in [\delta_l, \delta_u]$ *for* $0 < \delta_l \leq \delta_u < \infty$ *equals*

$$\psi_t(\delta) \;=\; \int\limits_{\overline{\mathscr{X}}_t \times \overline{\mathscr{A}}_t} \mu(h_t, a_t, R_{t+1} = 1) \prod_{s=1}^{t} q_s(a_s \mid h_s, R_s = 1) d\nu(a_s)\; d\mathbb{P}(x_s \mid h_{s-1}, a_{s-1}, R_s = 1) \tag{2.3}$$

for $t \leq T$*, where* $\overline{\mathscr{X}}_t = \mathscr{X}_1 \times \cdots \times \mathscr{X}_t$*,* $\overline{\mathscr{A}}_t = \mathscr{A}_1 \times \cdots \times \mathscr{A}_t$*,*
$\mu(h_t, a_t, R_{t+1} = 1) = \mathbb{E}(Y_t \mid H_t = h_t, A_t = a_t, R_{t+1} = 1)$*, and*

$$q_s(a_s \mid h_s, R_s = 1) = \frac{a_s \delta \pi_s(h_s, R_s = 1) + (1 - a_s)\{1 - \pi_s(h_s, R_s = 1)\}}{\delta \pi_s(h_s, R_s = 1) + 1 - \pi_s(h_s, R_s = 1)}. \tag{2.4}$$

with $\pi_s(h_s, R_s = 1) = \mathbb{P}(A_s = 1 \mid H_s = h_s, R_s = 1)$ *and a dominating measure* ν *for the distribution of* A_s*.*

Theorem 2.3.1 follows by Theorem 1 in [67] and Lemma A.4.1 given in the appendix. Note that $q_s(a_s \mid h_s)$ is the propensity score under the incremental intervention. The identifying expression (2.3) shows that the mean counterfactual outcome $\psi_t(\delta)$ is identified and can be expressed in terms of the observed data distribution $\mathbb{P}$.

As mentioned earlier, without the additional assumptions (A2-M) and (A3) together with the result of Lemma A.4.1, the intervention effect $\psi_t(\delta)$ would in general not be identifiable under the setting considered by Kennedy [67], due to the dropout. It is also worth noting that here we do not make any parametric assumptions and the censorship process is also allowed to be model-free. Theorem 2.3.1 therefore extends previous results on incremental interventions to studies with arbitrary time-varying outcomes and missing-at-random style dropout.

To illustrate, the next corollary shows what the identification result gives in the simple setting where there is only one timepoint, so dropout amounts to mere missing outcomes.

Corollary 2.3.1. *When* $T = 1$*, the data structure reduces to*

$$Z = (X, A, R, RY)$$

where $R = 1$ means the outcome was not missing. Then the identifying expression for $\psi(\delta)$ simplifies to

$$\psi(\delta) = \mathbb{E}\left[\frac{\delta\pi(X)\mu(X,1,1) + \{1-\pi(X)\}\mu(X,0,1)}{\delta\pi(X) + \{1-\pi(X)\}}\right]$$

where $\pi(X) = \mathbb{P}(A = 1 \mid X)$ and $\mu(x,a,1) = \mathbb{E}(Y \mid X = x, A = a, R = 1)$.

Therefore when $T = 1$ the effect $\psi(\delta)$ is simply a weighted average of the regression functions $\mu(X,1,1)$ and $\mu(X,0,1)$ among those with observed outcomes, with weights depending on the observational propensity scores and δ.

2.4 Efficiency Theory

In the previous section, we showed the incremental intervention effect adjusted for right-censoring and repeated outcomes can be identified under weak nonparametric assumptions, without requiring any positivity conditions on the treatment process. Our main goal in this section is to develop a nonparametric efficiency theory for the incremental effect, via the efficient influence function for $\psi_t(\delta)$.

The efficient influence function is a crucial object in non/semiparametric efficiency theory because 1) its variance gives an asymptotic efficiency bound that cannot be improved upon without adding assumptions, and 2) its form indicates how to do appropriate bias correction in order to construct estimators that attain the efficiency bound under weak conditions. Mathematically, given a target parameter ψ an *influence function* ϕ acts as the derivative term in a distributional Taylor expansion of the functional of interest, which can be seen to imply

$$\frac{\partial\psi(\mathbb{P}_\varepsilon)}{\partial\varepsilon}\Big|_{\varepsilon=0} = \int \phi(z;\mathbb{P})\left(\frac{\partial\log d\mathbb{P}_\varepsilon(z)}{\partial\varepsilon}\right)\Big|_{\varepsilon=0} d\mathbb{P}(z) \tag{2.5}$$

for all smooth parametric submodels $\mathbb{P}_\varepsilon$ containing the true distribution so that $\mathbb{P}_{\varepsilon=0} = \mathbb{P}$. Of all the influence functions, the *efficient influence function* is defined as the one which gives the greatest lower bound of all parametric submodel $\mathbb{P}_\varepsilon$, so giving the efficiency bound for estimating ψ. For more details we refer to Bickel et al. [11], Vaart [132], van der Laan & James M Robins [133], Tsiatis [128], Kennedy [65], as well as Section 1.2.

The next theorem gives an expression for the efficient influence function for the incremental effect $\psi_t(\delta)$ at arbitrary time $t \le T$ under a nonparametric model, which is the main result in this section.

Theorem 2.4.1. *The efficient influence function for the intervention effect $\psi_t(\delta)$ under a nonparametric model is given by*

$$\sum_{s=1}^{t}\left\{\frac{\{A_s-\pi_s(H_s)\}(1-\delta)}{\delta A_s+1-A_s}\right\}\left[\frac{m_s(H_s,1,R_{s+1}=1)\delta\pi_s(H_s)+m_s(H_s,0,R_{s+1}=1)\{1-\pi_s(H_s)\}}{\delta\pi_s(H_s)+1-\pi_s(H_s)}\right]$$

$$\times\,\omega_s(H_s,A_s)\left(\prod_{k=1}^{s}\frac{\delta A_k+1-A_k}{\delta\pi_k(H_k)+1-\pi_k(H_k)}\cdot\frac{\mathbb{1}\,(R_{s+1}=1)}{\omega_s(H_s,A_s)}\right)$$

$$+\prod_{s=1}^{t}\left\{\frac{\delta A_s+1-A_s}{\delta\pi_s(H_s)+1-\pi_s(H_s)}\cdot\frac{\mathbb{1}\,(R_{s+1}=1)}{\omega_s(H_s,A_s)}\right\}Y_t-\psi_t(\delta)$$

where $\pi_s(h_s)=\mathbb{P}(A_s=1\mid H_s=h_s,R_s=1)$, $\omega_s(H_s,A_s)=d\mathbb{P}(R_{s+1}=1\mid H_s,A_s,R_s=1)$, and

$$\begin{aligned}&m_s(h_s,a_s,R_{s+1}=1)\\&=\int_{\mathscr{R}_s}\mu(h_t,a_t,R_{t+1}=1)\prod_{k=s+1}^{t}q_k(a_k\mid h_k,R_k=1)d\nu(a_k)d\mathbb{P}(x_k|h_{k-1},a_{k-1},R_k=1)\end{aligned}$$

for $\forall s\leq t$, where $\mathscr{R}_s=(\overline{\mathscr{X}}_t\times\overline{\mathscr{A}}_t)\setminus(\overline{\mathscr{X}}_s\times\overline{\mathscr{A}}_s)$, $\mu(h_t,a_t,R_{t+1}=1)=\mathbb{E}(Y_t\mid H_t=h_t,A_t=a_t,R_{t+1}=1)$, and ν is a dominating measure for the distribution of A_k.

A proof can be found in the appendix A.4.2. This result will be used to construct efficient, model-free estimators for our new incremental effects in the next section. Note that in Theorem 2.4.1, all terms can be estimated via regression tools or simply obtained from the observed data. As one may expect, if there is no censoring (i.e., $\mathbb{P}[R_t=0]=1$ a.e $[\mathbb{P}]$ for all $t\leq T$) then both the identifying expression and the efficient influence function reduce to the expressions presented in Kennedy [67].

The efficient influence function in Theorem 2.4.1 consists of an augmentation term and an product term, both of which are quite different from those that appear in estimators for more standard causal effects. The structure of quotient terms is rooted in the form of our incremental interventional score defined in (2.4). It is worth noting that every quotient term is multiplied by $\frac{\mathbb{1}(R_{s+1}=1)}{\omega_s(H_s,A_s)}$ to adjust dropout effects at each stage s.

The above efficient influence function involves three types of nuisance functions: the treatment propensity scores $\pi_s(H_s)$, the missingness propensity scores $\omega_s(H_s,A_s)$ and the psuedo outcome regression functions $m_s(H_s,A_s,R_{s+1}=1)$ for $s\leq t$. The propensity scores $\pi_s(H_s)$ and $\omega_s(H_s,A_s)$ can be directly estimated. The psuedo outcome regression functions m_s can be estimated through sequential regressions without resorting to complicated conditional density estimation, since they are marginalized versions of the full regression function $\mu(h_s,a_s,R_{s+1}=1)$ that condition on all of the past. In the appendix A.4.3 we give a sequential regression formulation for these regression functions m_s.

The efficient influence function in the $T = 1$ case follows a relatively simple and intuitive form, equaling a weighted average of the efficient influence functions for $\mathbb{E}(Y^1)$ and $\mathbb{E}(Y^0)$ plus some contribution from the propensity scores ω_s, π_s. We give this result in the appendix A.4.4.

2.5 Estimation and Inference

2.5.1 Proposed Estimator

In this section we develop an estimator that can attain fast$\sqrt{n}$ convergence rates, even when other nuisance functions are modeled nonparametrically and estimated at rates slower than$\sqrt{n}$.

To begin, let $\varphi(Z;\boldsymbol{\eta},\delta,t)$ denote the uncentered efficient influence function from Theorem 2.4.1, which is a function of the observations Z, indexed by a set of nuisance functions

$$\boldsymbol{\eta} = (\boldsymbol{\pi},\boldsymbol{m},\boldsymbol{\omega}) = (\pi_1,...,\pi_t,m_1,...,m_t,\omega_1,...,\omega_t),$$

δ, and $t \leq T$, where π_t, m_t, ω_t are the same nuisance functions defined in Theorem 2.4.1. Thus $\mathbb{E}[\varphi(Z;\boldsymbol{\eta},\delta,t)] = \psi_t(\delta)$.

A natural estimator for $\phi(Z;\boldsymbol{\eta},\delta)$ would be given by the solution to the efficient influence function estimating equation, i.e., the naive plug-in Z-estimator

$$\hat{\psi}_{inc.pi}(t;\delta) = \mathbb{P}_n\{\varphi(Z;\hat{\boldsymbol{\eta}},\delta,t)\}$$

where $\hat{\boldsymbol{\eta}}$ represents a set of nuisance functions estimates, and $\mathbb{P}_n$ denotes the empirical measure so that sample averages can be written as $\frac{1}{n}\sum_i f(Z_i) = \mathbb{P}_n\{f(Z)\} = \int f(z)d\mathbb{P}_n(z)$.

If we assume π_t and ω_t were correctly parametrically modeled, then one could use the following simple inverse-probability-weighted (IPW) estimator

$$\hat{\psi}_{inc.ipw}(t;\delta) = \mathbb{P}_n\left\{\prod_{t=1}^{T}\left(\frac{\delta A_t + 1 - A_t}{\delta\hat{\pi}_t(H_t) + 1 - \hat{\pi}_t(H_t)} \cdot \frac{\mathbb{1}(R_{t+1} = 1)}{\hat{\omega}_t(H_t,A_t)}\right)Y\right\}.$$

Note that this IPW estimator is a special case of $\hat{\psi}_{inc.pi}$ where $\hat{m}_t$ is set to zero for all t.

However, to develop general Z-estimators with desired convergence rates with nonparametric models requires strong empirical process conditions (e.g., Donsker-type or low entropy conditions) that restrict the flexibility of the nuisance estimators. This is due to using the data twice (once for estimating the nuisance functions, again for estimating the average of the uncentered influence function), which can cause overfitting. Hence, to avoid this downside and to make our estimator more practically useful, we use sample splitting, following [155, 18, 67, 111]. As will be seen shortly, sample splitting allows us to achieve fast parametric$\sqrt{n}$ rates even when all the nuisance functions $\boldsymbol{\eta}$ are flexibly estimated at much slower rates than$\sqrt{n}$.

To this end we randomly split the observations $(Z_1,...,Z_n)$ into K disjoint groups, using a random variable S_i, $i = 1,...,n$, drawn independently of the data, where each $S_i \in \{1,...,K\}$ denotes the group membership for unit i. Then our proposed estimator is given by

$$\widehat{\psi}_t(\delta) = \mathbb{P}_n\left\{\varphi(Z;\hat{\boldsymbol{\eta}}_{-S},\delta,t)\right\} \equiv \frac{1}{K}\sum_{k=1}^{K}\mathbb{P}_n^{(k)}\{\varphi(Z;\hat{\boldsymbol{\eta}}_{-k},\delta,t)\} \tag{2.6}$$

where we let $\mathbb{P}_n^{(k)}$ denote sample averages only over the set of units $\{i : S_i = k\}$ in group k, and let $\hat{\boldsymbol{\eta}}_{-k}$ denote the nuisance estimator constructed excluding group k. We detail exactly how to compute the proposed estimator $\widehat{\psi}_t(\delta)$ in Algorithm 4 in section A.1 of the appendix.

Computing the estimator is easily amenable to parallelizable due to the sample splitting. It is worth noting that our method effectively utilizes all the observed samples available at each time t without discarding any observations in advance.

2.5.2 Convergence Theory

Now we provide a theorem that details the main large-sample property of our proposed estimator. In the theorem we verify that $\widehat{\psi}_t(\delta)$ is$\sqrt{n}$-consistent and asymptotically normal even when all the nuisance functions are estimated at much slower than $n^{-1/2}$ rates.

In what follows we denote the $L_2(\mathbb{P})$ norm of function f by $\|f\| = \left(\int f(z)^2 d\mathbb{P}(z)\right)^{1/2}$, to distinguish it from the ordinary L_2 norm $\|\cdot\|_2$ for a fixed vector. Moreover note that although we write m_t for the pseudo-regression functions defined in Theorem 2.4.1 for the sake of brevity, in principle they should be indexed by both time t and the given increment parameter δ as $m_{t,\delta}$. The next theorem shows uniform convergence of $\hat{\psi}_t(\delta)$, which lays the foundation for inference.

Theorem 2.5.1. *Define the variance function as* $\sigma^2(\delta,t) = \mathbb{E}\left[(\varphi(Z;\boldsymbol{\eta},\delta,t) - \psi_t(\delta))^2\right]$ *and let* $\hat{\sigma}^2(\delta,t) = \mathbb{P}_n\left[(\varphi(Z;\hat{\boldsymbol{\eta}}_{-S},\delta,t) - \hat{\psi}_t(\delta))^2\right]$ *denote its estimator. Assume:*

1) *The set* $\mathscr{D} = [\delta_l, \delta_u]$ *is bounded with* $0 < \delta_l \leq \delta_u < \infty$.

2) $\mathbb{P}[\mid m_t(H_t,A_t,R_{t+1}=1)\mid \leq C] = \mathbb{P}[\mid \hat{m}_t(H_t,A_t,R_{t+1}=1)\mid \leq C] = 1$ *for some constant* $C < \infty$ *and* $\forall t$.

3) $\sup_{\delta\in\mathscr{D}}\left|\frac{\hat{\sigma}^2(\delta,t)}{\sigma^2(\delta,t)}-1\right| = o_{\mathbb{P}}(1)$, *and* $\|\sup_{\delta\in\mathscr{D}}|\varphi(Z;\boldsymbol{\eta},\delta,t)-\varphi(Z;\hat{\boldsymbol{\eta}}_{-S},\delta,t)|\| = o_{\mathbb{P}}(1)$.

4) $\left(\underset{\delta\in\mathscr{D}}{sup}\|m_{\delta,t}-\widehat{m}_{\delta,t}\|+\|\pi_t-\widehat{\pi}_t\|\right)\left(\|\widehat{\pi}_s-\pi_s\|+\|\widehat{\omega}_s-\omega_s\|\right) = o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)$ *for* $\forall s \le t$.

Then we have

$$\frac{\hat{\psi}_t(\delta)-\psi_t(\delta)}{\hat{\sigma}(t,\delta)/\sqrt{n}} \rightsquigarrow \mathbb{G}(\delta,t)$$

in $l^\infty(\mathscr{D})$*, where* $\mathbb{G}$ *is a mean-zero Gaussian process with covariance* $\mathbb{E}[\mathbb{G}(\delta_1,t_1)\mathbb{G}(\delta_2,t_2)] = \mathbb{E}\left[\widetilde{\varphi}(Z;\boldsymbol{\eta},\delta_1,t_1)\widetilde{\varphi}(Z;\boldsymbol{\eta},\delta_2,t_2)\right]$ *and* $\widetilde{\varphi}(Z;\boldsymbol{\eta},\delta,t) = \frac{\varphi(Z;\boldsymbol{\eta},\delta,t)-\psi_t(\delta)}{\sigma(\delta,t)}$.

A proof of the above theorem is given in the appendix A.4.7. We also analyze the second order remainder terms of the efficient influence function given in Lemma A.4.2, and keep the intervention distribution completely general (see section A.4.8, A.4.9 in the appendix). Therefore, the results can be applied to study other stochastic interventions under the presence of right-censoring (which is beyond the scope of this paper).

Assumptions 1), 2) and 3) in Theorem 2.5.1 are all very weak. Assumptions 1) and 2) are mild boundedness conditions, where assumption 2) could be further relaxed at the expense of a less simple proof, for example with bounds on L_p norms. Assumption 3) is also a basic and mild consistency assumption, with no requirement on rates of convergence. The main substantive assumption is Assumption 4), which requires that product of nuisance function estimation errors must vanish at fast enough rates. Note that unlike the result from [67], we have additional nuisance function ω in the condition. One sufficient condition for Assumption 4 to hold is that all the nuisance functions are consistently estimated at a rate of $n^{-1/4}$ or faster.

Lowering the bar from$\sqrt{n}$ to $n^{-1/4}$ indeed allows us to employ a richer set of modern machine learning methods by reducing the burden of nonparametric modeling. Such rates are attainable under diverse structural constraints, e.g., [149, 106, 61, 44]. In this paper we are agnostic about how such rates might be attained. In practice, we may want to consider using different estimation techniques for each of $\boldsymbol{\pi},\boldsymbol{m},\boldsymbol{\omega}$ based on our prior knowledge or use ensemble learners.

Based on the result in Theorem 2.5.1, given the value of δ and t we can construct pointwise $1-\alpha$ confidence intervals for $\psi_t(\delta)$ as

$$\widehat{\psi}_t(\delta) \pm z_{1-\alpha/2}\frac{\hat{\sigma}^2(\delta,t)}{\sqrt{n}}$$

where $\hat{\sigma}^2(\delta,t)$ is the variance estimator defined in Theorem 2.5.1. As in [67] we can use the multiplier bootstrap for uniform inference, by replacing the $z_{1-\alpha/2}$ critical value with one c_α

satisfying

$$\mathbb{P}\left(\sup_{\delta\in\mathscr{D},1\leq t\leq T}\left|\frac{\widehat{\psi}_t(\delta)-\psi_t(\delta)}{\widehat{\sigma}(\delta,t)/\sqrt{n}}\right|\leq c_\alpha\right)=1-\alpha+o(1).$$

This is due to the fact that we only add a finite number T timepoints into the function class of φ at maximum (see A.4.8 in the appendix for more detailed discussion). We refer to [67] for details on how to construct c_α via a bootstrap procedure.

2.6 Infinite Time Horizon Analysis

The great majority of causal inference literature considers a finite time horizon where the number of timepoints is small and fixed, or even just equal to one, a priori ruling out much significant (if any) longitudinal structure. However, in practice more and more studies accumulate data across very many timepoints, due to ever increasing advances in data collection technology. In fact, in many applications the number of timepoints T can even be comparable to or larger than sample size n, rendering most of the classical methods based on finite time horizons futile. For example, [81] describe how new mobile and wearable sensing technologies have revolutionized randomized trial and other health-care studies by providing data at very high sampling rates (e.g., 10-500 times per second). [77, 104] use $T = 210$ timepoints in their study in which they present the micro-randomized trial for just-in-time adaptive interventions via mobile applications. As we collect such more granular and fine-grained data, some recent studies explore efficient off-policy estimation techniques on infinite-time horizon (e.g. Liu et al. [90] in reinforcement learning). Interestingly, there has been no formal analysis for general longitudinal studies.

Therefore here we analyze the behavior of the IPW version of our proposed incremental effect estimator (relative to a standard IPW estimator of a classical deterministic effect), in a more realistic regime where the number of timepoints can scale with sample size. To the best of our knowledge, this is one of the first such infinite-horizon analyses in causal inference, outside of some recent examples involving dynamic treatment regimes [83, 32]. Specifically, we study the relative efficiency bound in the number of timepoints T and show how deterministic effects are afflicted by a large variance inflation relative to incremental effects in the infinite time horizon. Even when two estimators target different effects, often the efficiency helps guide us through the problem of selecting the estimand [e.g., 3, 22], particularly when we do not have strong preference for one over the other.

We proceed with comparing the deterministic effect of receiving treatment at every timepoint and the incremental effect for $\delta > 1$ (the result for effects of receiving control at every timepoint is similar and presented in the Section A.4.5 of appendix). The incremental intervention effect is asymptotically equivalent to the deterministic effect for the always-treated as $\delta \to \infty$ (the larger δ is, the more closer the two effects are).

For the sake of simplicity, we consider the case where propensity scores are known and do not vary with covariates (i.e., $\pi_t(H_t) = p$ for all t) and there is no dropout (i.e. $d\mathbb{P}\{R_{t+1} = 1\} = 1$ a.e. $[\mathbb{P}]$ for all $t = 1,...,T$). In other words, we consider a simple setup where the propensity scores are all equal to p and the pseudo-regression functions m_t's are zero in the full nonparametric efficiency bounds.

In this setup we have unbiased estimators of the always-treated effect $\psi_{at} = \mathbb{E}(Y^{\bar{\mathbf{1}}})$ and the incremental effect $\psi_{inc} = \mathbb{E}(Y^{\overline{Q}(\delta)})$ given by

$$\widehat{\psi}_{at} = \prod_{t=1}^{T} \left(\frac{A_t}{p} \right) Y$$

and

$$\widehat{\psi}_{inc} = \prod_{t=1}^{T} \left(\frac{\delta A_t + 1 - A_t}{\delta p + 1 - p} \right) Y$$

respectively, where $Y = Y_T$. We now explore the asymptotic relative efficiency of these estimators in the case where $T \to \infty$. In the next theorem, we show that one can achieve the near-exponential efficiency gain by targeting ψ_{inc} instead of ψ_{at}.

Theorem 2.6.1. *Consider the estimators and assumptions defined above. Suppose* $|Y| \leq b_u$ *for some constant* $b_u > 0$ *and* $\mathbb{E}\left[\left(Y^{\bar{\mathbf{1}}}\right)^2\right] > 0$. *Then for any* $T \geq 1$,

$$C_T \left[\left\{ \frac{\delta^2 p^2 + p(1-p)}{(\delta p + 1 - p)^2} \right\}^T - p^T \right] \leq \frac{Var(\widehat{\psi}_{at})}{Var(\widehat{\psi}_{inc})} \leq C_T \zeta(T;p) \left\{ \frac{\delta^2 p^2 + p(1-p)}{(\delta p + 1 - p)^2} \right\}^T$$

where $C_T = \frac{b_u^2}{\mathbb{E}\left[\left(Y^{\bar{\mathbf{1}}}\right)^2\right]}$ *and* $\zeta(T;p) = \left(1 + \frac{c\left(\mathbb{E}\left[Y^{\bar{\mathbf{1}}}\right]\right)^2}{(1/p)^T \mathbb{E}\left[\left(Y^{\bar{\mathbf{1}}}\right)^2\right]}\right)$ *for any fixed value of c such that* $\frac{1}{1-p^T\left(\mathbb{E}\left[Y^{\bar{\mathbf{1}}}\right]\right)^2 / \mathbb{E}\left[\left(Y^2\right)^{\bar{\mathbf{1}}}\right]} \leq c.$

A proof of the above theorem can be found in the Section A.4.5 of the appendix and is based on similar logic used in deriving the g-formula [108]. In the proof, we give more general results for any deterministic effects $\mathbb{E}(Y^{\bar{a}_T})$, $\forall \bar{a}_T \in \overline{\mathscr{A}}_T$. Note that we only require two very mild assumptions: the boundedness assumption on Y and $\mathbb{E}[(Y^{\bar{\mathbf{1}}})^2] > 0$ which is equivalent to say that $Y^{\bar{\mathbf{1}}}$ is a non-degenerate random variable.

Theorem 2.6.1 allows us to precisely quantify the asymptotic relative efficiency gain. Importantly, since $\frac{\delta^2 p^2 + p(1-p)}{(\delta p + 1 - p)^2} < 1$ for $\delta > 1$ and $\zeta(T;p) \to 1$ monotonically at an exponential rate in T, the efficiency gain is also almost exponential in T. The result for the never-treated effects is similar as well, and given in Section A.4.5 of the appendix. It is clear to see $\lim_{\delta\to\infty}\frac{\delta^2p^2+p(1-p)}{(\delta p+1-p)^2}=1^-$. Hence, the result of Theorem 2.6.1 can be framed as a trade-off between efficiency gain and bias in targeting the same effects.

Theorem 2.6.1 implies that $\widehat{\psi}_{inc}$ will be always more efficient than $\widehat{\psi}_{at}$ if we intend to incorporate substantial number of timepoints. In what follows we refine this statement so that one can characterize the minimum threshold of the number of timepoints to make the claim true, under the same condition used in Theorem 2.6.1.

Corollary 2.6.1. *There exists a finite number T_{min} such that*

$$Var(\widehat{\psi}_{inc}) < Var(\widehat{\psi}_{at})$$

for every $T > T_{min}$, where T_{min} is never greater than

$$\min\left\{T : \left[\frac{\delta^2 p+1-p}{(\delta p+1-p)^2}\right]^T - \frac{c_1}{p^T} + 2 < 0\right\} \quad \textit{where} \quad c_1 = \frac{\mathbb{E}\left[\left(Y^{\overline{1}}\right)^2\right]}{b_u^2}.$$

A proof can be found in Section A.4.6 in the appendix. The proof of the above corollary relies upon the fact that $var(\widehat{\psi}_{inc})$ can be represented as the variance of the weighted sum of all the distinct deterministic intervention effects $\overline{a}_T \in \overline{\mathscr{A}}_T$ (Lemma A.4.7). The constant c_1 is simply the normalized second order moment and can be translated into the average magnitude of $Y^{\overline{1}}$. In other words, the larger $|Y^{\overline{1}}|$ is, the smaller T_{min} is.

Remark 1. *It may be possible to tighten the upper bound for T_{min}, but in practice the value of T_{min} is typically already small. To illustrate, consider the setup where $Y \in [0,1]$ and $\delta = 2.5, p = 0.5$, and two extreme cases: $c_1 = 0.95$ ($Y^{\overline{1}}$ is dispersed mostly around $\{0,1\}$) and $c_1 = 0.05$ ($Y^{\overline{1}}$ is concentrated around 0). Then the corresponding T_{min} values are 2 and 6 respectively. If we use $\delta = 5, p = 0.5$, the numbers will become 3 and 9 respectively.*

Our proof of Theorem 2.6.1 and Corollary 2.6.1 can be generalized to the case where the nuisance functions need to be estimated, but we feel the simple case captures the main ideas, and the general case would only add complexity. Numerical simulations given in Section A.2 of the appendix support our result. Our result in this section provides a crucial insight into the longitudinal studies with many timepoints, indicating massive efficiency gains are possible by studying incremental rather than classical deterministic effects.

2.7 Experiments

2.7.1 Simulation Study

In this section we explore finite-sample performance of the proposed estimator $\hat{\psi}_t(\delta)$ via synthetic simulation for an observational study. We consider the following data generation model

$$X_t = (X_{1,t}, X_{2,t}) \sim N(0, \mathbf{I}),$$

$$\pi_t(H_t) = expit\Big(\mathbf{1}^\top X_t + 2\sum_{s=t-2}^{t-1}(A_s - 1/2)\Big),$$

$$\omega_t(H_t, A_t) = expit\Big(C_0 + \sum_{s=1}^{t} A_s\Big)\,, \quad C_0 \sim \mathscr{U}[u_l, 5],$$

$$\big(Y\big|\overline{X}_t, \overline{A}_t\big) \sim N\big(\mu(\overline{X}_t, \overline{A}_t), 1\big)$$

for all $t = 1, ..., t$ where we set $\mu(\overline{X}_t, \overline{A}_t) = 10 + A_t + A_{t-1} + |((\mathbf{1}^\top X_t + \mathbf{1}^\top X_{t-1})\;|$ and $t = 50$. $\mathscr{U}[u_l, 5]$ is a uniform random variable with interval $[u_l, 5]$. Basically in this setup we assume that the more likely to have been treated, the less likely to drop out from the study.

We use three baseline methods: the naive Z-estimator ($\hat{\psi}_{inc.pi}$) and the IPW type estimator ($\hat{\psi}_{inc.ipw}$), both of which are defined in Section 2.5.1, and the incremental-effect estimator ($\hat{\psi}_{inc.nc}$) proposed by Kennedy [67], which does not take right-censoring into account. Since finite-sample properties of $\hat{\psi}_{inc.nc}$ were already extensively explored in Kennedy [67], here we focus more on dropout effect on performance.

To estimate nuisance parameters, we form an ensemble of some widely-used nonparametric models. Specifically, we use cross-validation-based superleaner ensemble algorithm [136] via the `SuperLearner` package in R to combine support vector machine, random forest, k-nearest neighbor regression. For the proposed method, we use sample splitting with $K = 2$ splits as described in Algorithm 4.

We repeat simulation S times in which we draw n samples each simulation. We use D values of δ equally spaced on the log-scale within $[0.1, 5]$. Then performance of each estimator is assessed via normalized root-mean-squared error (RMSE) defined by

$$\widehat{RMSE} = \frac{1}{D}\sum_{d=1}^{D}\left[\frac{1}{S}\sum_{s=1}^{S}\left\{\frac{\hat{\psi}_s(t;\delta_d) - \psi(t;\delta_d)}{\overline{\psi}(t;\delta_d)}\right\}^2\right]$$

where $\hat{\psi}_s(t;\delta_d)$ and $\psi(t;\delta_d)$ are an estimated value of estimator for s-th simulation with value δ_d and a true value of the target parameter with δ_d respectively, and $\overline{\psi}(t;\delta_d)$ means a sample average of $\psi(t;\delta_d)$ across different values of δ_d. We present the results in Table 2.1.

Setup	$\widehat{RMSE}$				Average Dropouts (%)
	$\hat{\psi}_{inc.pi}$	$\hat{\psi}_{inc.ipw}$	$\hat{\psi}_{inc.nc}$	$\hat{\psi}_{proposed}$	
$S = 500, n = 1000, D = 25, u_l = 1$	0.85	0.82	0.91	**0.55**	36.0
$S = 500, n = 5000, D = 25, u_l = 1$	0.69	0.60	0.73	**0.36**	35.1
$S = 500, n = 1000, D = 25, u_l = 5$	0.72	0.81	**0.63**	0.64	4.7
$S = 500, n = 5000, D = 25, u_l = 5$	0.58	0.65	0.40	**0.38**	4.9

Table 2.1 Normalized RMSE across different baselines and simulation settings.

As shown in Table 2.1, the proposed estimator performs in general better than the other baseline methods especially when lots of data are dropped out, as anticipated from the theory. $\hat{\psi}_{inc.pi}$ and $\hat{\psi}_{inc.ipw}$ in general show fairly large RMSE, since they are not expected to converge at $\sqrt{n}$ rates. Under the substantial dropout rates, $\hat{\psi}_{inc.nc}$ shows even worse performance than these estimators due to the smaller number of effective samples [1]. On the other hand, $\hat{\psi}_{inc.nc}$ shows comparable performance to the proposed estimator when there is only a small portion of censored data.

2.7.2 Application

Here we illustrate the proposed methods in analyzing the Effects of Aspirin on Gestation and Reproduction (EAGeR) data, which evaluates the effect of daily low-dose aspirin on pregnancy outcomes and complications. The EAGeR trial was the first randomized trial to evaluate the effect of pre-conception low-dose aspirin on pregnancy outcomes ([119, 97]). However, to date this evidence has been limited to intention-to-treat analyses.

The design and protocol used for the EAGeR study have been previously documented [120]. Overall, 1,228 women were recruited into the study (615 aspirin, 613 placebo) and 11% of participants chose to drop out of the study before completion. Roughly 43,000 person weeks of information were available from daily diaries, as well as study questionnaires, and clinical and telephone evaluations collected at regular intervals over follow-up. The dataset is characterized by a substantial degree of non-compliance (more than 50% at the end of the study), and thereby is susceptible to positivity violation.

We used our incremental propensity score approach to evaluate the effect of aspirin on live birth and pregnancy loss in the EAGeR trial, accounting for time-varying exposure and dropout. The EAGeR dataset has been compiled as described in (2.1). Here, the study terminates at week 89 ($T = 89$). We use 24 baseline covariates (e.g., age, race, income,

[1] For $\hat{\psi}_{inc.nc}$, we discard samples that have been dropped out.

education, etc.) and 5 time-dependent covariates (compliance, conception, vaginal bleeding, nausea and GI discomfort). A_t is a binary treatment variable coded as 1 if a woman took aspirin at time t and 0 otherwise. $R_t = 1$ indicates that the woman is observed in the study at time t. Lastly, Y_t is an indicator of having a pregnancy outcome of interest at time t. We are particularly interested in two types of pregnancy outcomes: live birth and pregnancy loss (fetal loss). We perform separate analysis for each of the two cases.

For comparative purposes, we estimate the simple complete-case effect

$$\widehat{\psi}_{CC} = \mathbb{P}_n(Y|\overline{A}_T = 1, R_T = 1) - \mathbb{P}_n(Y|\overline{A}_T = 0, R_T = 1).$$

which relies on both non-compliance and drop-out being completely randomized. The value of $\widehat{\psi}_{CC}$ is 0.052 (5.2%) for live birth and 0.012 (1.2%) for pregnancy loss, both of which are close to the intention-to-treat estimates reported in [120, 119].

Here, we give a brief discussion why standard approaches might fail for our analysis of the EAGER dataset. We found a strong evidence of positivity violation in the EAGER dataset due to non-compliance; as shown in Figure A.3 in Section A.3 of the appendix, the average propensity score quickly drops to zero as t grows. So it would be hard to imagine having *all* of the study participants take aspirin at each time. Standard approaches such as widely-used marginal structural models (MSMs) [112] typically require treatment positivity, and thus are likely to fail for our analysis. In fact, when we model the effect curve by $\mathbb{E}[Y^{\overline{a}_T}] = m(\overline{a}_T;\beta) = \beta_0 + \sum_{t=1}^{T} \beta_{1t} a_t$ so that the coefficient for exposure can vary with time, then an inverse-weighted MSM estimator indeed fails and no coefficient estimates can be found even for moderate value of T, e.g. $T =\sim 10$. (see Figure A.3-(b) in the appendix for a closer look on why). This positivity violation precludes most of the standard approaches for time-varying treatments including MSMs. Not to mention that the MSM approach relies on parametric models.

Therefore, in order to avoid positivity violation, we alter our target contrast from the standard ATE to the mean outcome we would have observed in a population if “observed" versus none (*not* all versus none as in the ATE) were treated. Then we apply some of other *nonparametric* approaches available in the literature; we use the g-computation (plug-in) estimator [108] and the sequential doubly robust (SDR) estimator proposed by Luedtke et al. [92]. In short, the result based on the g-computation estimator (Figure A.4 in the appendix) shows that the counterfactual mean outcomes for never-takers are worse-off than the observed, whereas the result based on the SDR estimator (Figure A.5 in the appendix) suggests that such effects look no longer statistically significant so we cannot make any firm conclusion. Above all else, the alternative effect we have estimated here entails the fundamental limitation on the target effect since in many cases the always-taker group is typically of utmost interest for researchers. See Section A.3 of the appendix for more details.

Now, we estimate the incremental effect curve $\psi_T(\delta)$, which represents the probability of having live birth or pregnancy loss at the end of the study ($T = 89$) if the odds of taking aspirin were multiplied by factor δ. Specifically, this effect compares the outcome probabilities that would be observed if the odds of taking aspirin for all women was increased by a factor of δ at all timepoints, across different values of δ. Again, we use the cross-validated superleaner algorithm [136] to combine support vector machine, random forest, k-nearest neighbor regression, and multivariate adaptive regression splines, to estimate a tuple of nuisance functions (m_t, ω_t, π_t) at every t. We use sample splitting as in Algorithm 4 with $K = 2$ splits, and use 10,000 bootstrap replications to compute pointwise and uniform confidence intervals. Results are shown in Figure 2.1.

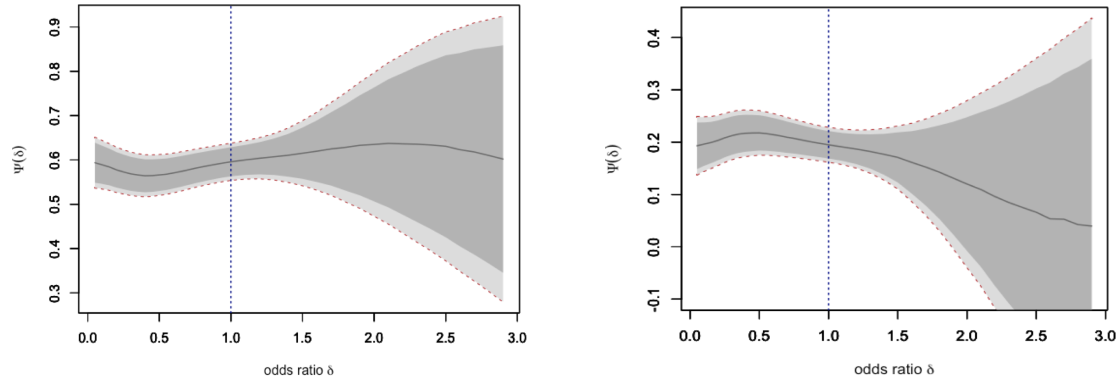


Fig. 2.1 Estimated incremental effect curves which represent the probability of having a live birth (Left) and a pregnancy loss (Right). In each figure, lighter grey area with red dotted line represents a 95% uniform band and darker grey area represents a 95% pointwise band.

We find the estimated curve is almost flat for live birth, and has a negative gradient with respect to δ (odds ratio) in general for pregnancy loss. Thus, unlike the previous findings, our result seems indicative of a positive effect of low-dose aspirin on reducing the risk of pregnancy loss; if odds of taking aspirin were increased proportionally for all individuals, the mean risk of pregnancy loss would drop from $\mathbb{P}_n(Y) = 19.3\%$ observationally to 13.1%, if the odds doubled. However, one needs to take the wider band at large δ into consideration. This analysis provides considerably more nuance than the alternative contrast used in the g-computation and SDR estimators or a standard MSM approach, and requires none of the parametric and positivity assumptions.

In conclusion, our analysis suggests new evidence that the low-dose aspirin therapy can be associated with decrease in the risk of pregnancy loss, but its accuracy is still afflicted with some uncertainties.

2.8 Discussion

Incremental interventions are a novel class of stochastic dynamic intervention where positivity assumptions can be completely avoided. However, they had not been extended to repeated outcomes, and without further assumptions do not give identifiability under dropout - both very common in practice. In this paper we solved this problem by showing how incremental intervention effects are identified and can be estimated when drop-out occurs (conditionally) at random. Even in the case of many dropouts, our proposed method efficiently uses all the data without sacrificing robustness. We give an identifying expression for incremental effects under monotone dropout, without requiring any positivity assumptions. We establish general efficiency theory and construct the efficient influence function, and present nonparametric estimators which converge at fast rates and yield uniform inferential guarantees, even when all the nuisance functions are estimated with flexible machine learning tools at slower rates. Furthermore, we studied the relative efficiency of incremental effects to conventional deterministic dynamic intervention effects in a novel infinite time horizon setting in which the number of timepoints can possibly grow with sample size, and showed that incremental effects are more efficient than deterministic effects and yield near-exponential efficiency gains in the infinite-time regime.

There are a number of avenues for future work. The first is application to other substantive problems in medicine and the social sciences. For example, in a forthcoming paper we analyze the effect of aspirin on pregnancy outcomes with more extensive data. It will also be important to consider other types of non-monotone missingness where the standard time-varying missing-at-random assumption A2-M may not be appropriate ([124, 126]). We expect our approach can be extended to other important problems in causal inference; for example, one could develop incremental effects for continuous treatments and instruments [74, 73], or for mediation in the same spirit as [26], but generalized to the longitudinal case with dropout. Developing incremental-based sensitivity analyses for the longitudinal missing-at-random assumption would also be important.

Chapter 3

Causal effects based on distributional distances

3.1 Introduction

We begin by considering a simple randomized experiment with a binary treatment $A \in \{0,1\}$ and an outcome $Y \in \mathbb{R}$ where $(A,Y) \sim \mathbb{P}$ for an unknown distribution $\mathbb{P}$, which is arguably one of the most classical and widely used setups in causal inference problems (e.g., A/B testing). Here, one often pursues the average treatment effect (ATE) of A on Y, defined as

$$\mathbb{E}[Y^1 - Y^0] \tag{3.1}$$

where Y^a denotes the *counterfactual* or *potential* outcome that would have been observed under $A = a$ for $a \in \{0,1\}$ [118].

In this paper, we provide a novel insight on causal inference by considering causal effects defined by means of a distributional distance between counterfactual outcome distributions. For example, in the above randomized experiment, letting Q^a denote the distribution of Y^a, we target the distributional distance between Q^1 and Q^0 defined by

$$D(Q^1, Q^0) \tag{3.2}$$

where D is a distance defined on distribution inputs. In practice, the simple randomized experiment described above is often not enough for causal effects of our interest. Therefore in our work, we also consider randomized studies where we have multiple data sources (e.g., A/B testing across different websites or countries) or general observational studies.

Our problem differs from traditional causal inference by relying on a more nuanced measure of treatment effect. Note that the traditional ATE defined in (3.1) can be zero even when the treatment has a significant impact. For example, if we suppose that $Y^0 = 0$ but $\mathbb{P}(Y^1 = 1) = \mathbb{P}(Y^1 = -1) = 1/2$, then the ATE is exactly zero. Should policy makers conclude in this case that the treatment really has no impact? This might be misleading, since the treatment indeed yields extreme harms and benefits to half the population. Therefore, more nuanced measure of treatment effects than the ATE are needed. On the other hand, unlike the ATE one may easily notice that the causal effect that we consider in (3.2) is substantially positive. Directly comparing counterfactual outcome distributions, such as the one in (3.2), can distinguish such subtle cases and in general provide more nuanced and valuable information about treatment effects beyond the ATE. So we can always use it jointly with the ATE in a complementary sense.

Relation to previous work. Here we give a brief review of some related literature, and refer to cited references for more details. There have been several attempts to incorporate distribution data into learning tasks in the modern machine learning. For example, distribution regression has been discussed in a regression framework for functional data [e.g., 105, 36, 125]. [60] studied smooth distance functionals using the theory of influence functions and sample splitting, and [59] gave minimax lower bounds for observational L_1 distances. There has also been substantial work in econometrics considering counterfactual quantile estimation, for example by [37] and [116]. However, these topics were not studied in causal inference framework.

We extend this previous work by proposing and studying non-smooth L_1 distributional distances between counterfactuals, not only in simple randomized experiments but also in more complex multi-source and observational studies. Studying the counterfactual versions of distributional distance functionals itself brings a number of non-trivial theoretical subtleties (see, for example, [60, 59, 125]). The same goes for the non-smooth L_1 distance compared to the quantile and cumulative distribution function (cdf)-based effects previously studied by [37, 116]. Nonetheless, the L_1 distance provides a number of advantages. First, it is a simple one-number summary of distributional differences, unlike the quantile and cdf effects which are potentially complex curves. Thus it can be a simple tool to use as a first step in assessing whether there is effect modification beyond a mean shift (e.g., when the average effect is zero). Second, even if one is interested in quantiles/cdfs, the L_1 distance can be used to test hypotheses that these quantities really differ. Finally, the L_1 distance is interpretable as it means the average absolute difference in densities, and is invariant under monotone transformations of Y (which is not true for many other distances including L_2 distance) [24].

Importantly, in this paper we also detail how to effectively estimate each counterfactual outcome distribution, which may provide an in-depth way of analyzing "what" drives the treatment effects. This problem is basically equivalent to density estimation with outcomes missing at random, which has received very little attention compared to the standard density estimation literature in statistics (with a few exceptions of [117, 113]). We develop a novel doubly-robust estimator for nonparametric counterfactual density estimation in observational studies, which, to the best of our knowledge, has not yet appeared in the literature and address the improvements as compared to the plug-in estimator. Finally, we provide a bootstrap approach to obtaining confidence intervals by characterizing the asymptotic convergence of our proposed estimators, which is useful for the inferential and testing procedures.

3.2 Preliminaries

3.2.1 Setup and Identification

Throughout, we consider binary treatments $A \in \mathscr{A} := \{0,1\}$ and real-valued outcome $Y \in \mathscr{Y} \subset \mathbb{R}^d$. Although $d = 1$ is the most common case in practice, we allow $d > 1$. Unless otherwise mentioned, we let $\mathbb{P}$ be an observed data distribution on a compact subset. In particular, we let Y have a density p with respect to d-dimensional Lebesgue measure λ_d. For a treatment assignment $A = a$ we define a counterfactual distribution Q^a as the distribution of Y^a. For the distributional distance, we take D to be the L_1 distance between densities, as in $D(P_1,P_2) = \|p_1 - p_2\|_1 = \int |p_1(u) - p_2(u)|du$ for two distributions P_1,P_2 and their corresponding densities p_1, p_2 with respect to λ_d. Further, the counterfactual density q^a is defined as a Radon-Nikodym derivatives of Q^a with respect to λ_d.

In what follows, we describe three different settings for which we develop our estimators.

Single-source randomized study: $\boldsymbol{Z = (A,Y)}$. In the simple randomized study we observe i.i.d samples $(Z_1,...,Z_n)$ where $Z = (A,Y) \sim \mathbb{P}$. For our causal parameter $D(Q^1,Q^0)$ in (3.2) to be identified, we require the following consistency and randomization assumptions.

- (C1) *Consistency*: $Y = Y^a$ if $A = a$
- (C2) *Randomization*: $A \perp\!\!\!\perp Y^a$

These assumptions are typically hold by design in randomized experiments. Randomization requires that treatment is independent of potential outcomes. Consistency implicitly conveys a no-interference condition: one subject's outcomes cannot be affected by others' treatments. Under assumptions (C1) and (C2), it is straightforward to see that $Q^a = \mathbb{P}(Y|A = a)$.

Consequently, we have the following identifying expression

$$D(Q^1,Q^0) = \int |q^1(y) - q^0(y)|dy = \int \big|p(y|A=1) - p(y|A=0)\big|dy. \tag{3.3}$$

Multi-source randomized study: $\boldsymbol{Z} = ((\boldsymbol{A},\boldsymbol{Y})_{\mathbb{P}},\mathbb{P})$. Now suppose that we are interested in causal effects over multiple sources of $\mathbb{P}$ still with the same data structure (A,Y). So the distributional properties of the data may vary across different $\mathbb{P}$'s. To this end, we let $\mathbb{D}$ denote the set of all distributions on $(\mathscr{A},\mathscr{Y})$ which have a density with respect to the Lebesgue measure. Then let $\mathscr{P}$ be a probability measure on a measurable space $(\mathbb{D},\sigma(\mathsf{D}))$ where $\sigma(\mathsf{D})$ is a σ-field generated by a measurable function $\mathsf{D}:\mathbb{D}\to\overline{\mathbb{R}}^{+0}$ which is defined on L_1 distance. Now suppose we have N distinct $\mathbb{P}_i$'s which are i.i.d. samples from the superpopulation distribution $\mathscr{P}$ on $\mathbb{D}$, that is,

$$\mathbb{P}_1,\mathbb{P}_2,...,\mathbb{P}_N \overset{\text{i.i.d.}}{\sim} \mathscr{P}$$

where $(A,Y)_{\mathbb{P}_i} \sim \mathbb{P}_i, i = 1,...,N$. Namely, each $(A,Y)_{\mathbb{P}_i}$ is a single-source experiment under $\mathbb{P}_i$ with n_i samples. Hence our target parameter in this case is given by

$$\mathbb{E}_{\mathscr{P}}[D(Q^1,Q^0)]. \tag{3.4}$$

This setting may be appropriate for when we conduct randomized experiments over different environments which can be assumed to be independent. For identification, we need slightly different assumptions from the simple, single-source randomized study.

- (C1) *Consistency*: $Y = Y^a$ if $A = a$
- (C2') *Conditional randomization*: $A \perp\!\!\!\perp Y^a$ for each $(A,Y) \sim \mathbb{P}_i$.

By the law of iterated expectation we have $\mathbb{E}_{\mathscr{P}}[D(Q^1,Q^0)] = \mathbb{E}_{\mathscr{P}}[\mathbb{E}\{D(Q^1,Q^0) \mid \mathbb{P}\}]$. Hence under assumptions (C1) and (C2'), conditioned on the sampled distribution $\mathbb{P}$, $D(Q^1,Q^0)$ is identified in the same way as in (3.3), and consequently the target effect (3.4) is identified.

Observational study: $\boldsymbol{Z} = (\boldsymbol{X},\boldsymbol{A},\boldsymbol{Y})$. In observational studies, the treatment happened naturally according to some unknown process, and was not under experimenter's control. Thus, randomization assumption no longer holds by design. Instead in general we try to collect as many relevant covariates as possible, in an attempt to ensure that treatment is at least conditionally randomized. We consider data structure $Z = (X,A,Y)$, with X on some compact support $\mathscr{X} \subset \mathbb{R}^k$. Our target parameter is still $D(Q^1,Q^0)$ in (3.2) but in observational study, we require a different set of assumptions for identification as follows.

- (C1) *Consistency*: $Y = Y^a$ if $A = a$
- (C3) *No unmeasured confounding*: $A \perp\!\!\!\perp Y^a \mid X$
- (C4) *Positivity*: $\mathbb{P}(A = a|X) > 0$ a.s $\mathbb{P}$

No unmeasured confounding (or exhangeability) will hold if the collected covariates can explain treatment assignment, to the extent that after conditioning on them treatment is not further related to potential outcomes. Positivity requires everyone to have some chance of being treated at all treatment levels. Under these assumptions we have $q^a(y) = \int_{\mathcal{X}} p(y|X = x, A = a)d\mathbb{P}(x)$, and thus our target parameter is identified as

$$\begin{aligned} D(Q^1,Q^0) &= \int |q^1(y) - q^0(y)|dy \\ &= \int \left| \int_{\mathcal{X}} \Big\{ p(y|X = x, A = 1) - p(y|X = x, A = 0) \Big\} d\mathbb{P}(x) \right| dy. \end{aligned} \tag{3.5}$$

Kernel-smoothed counterfactual density. Finding an efficient estimator for the counterfactual density q^a is challenging and still an open problem. In this paper, we instead target the kernel-smoothed version of our counterfactual density defined as

$$q^a_h(y) := \mathbb{E}\left\{ \frac{1}{h^d} K\left(\frac{\|Y^a - y\|_2}{h} \right) \right\}, \tag{3.6}$$

with a valid kernel K and its bandwidth h. The smoothing bias vanishes as h goes to zero [1]. With this kernel-smoothed counterfactual density q^a_h, our target parameter for each scenario is still identified under the exactly same set of assumptions. Specifically, for single- and multi-source randomized experiments under assumptions (C1),(C2) or (C1),(C2') we have an identifying expression for q^a_h as

$$q^a_h(y) = \mathbb{E}\left\{ \frac{1}{h^d} K\left(\frac{\|Y - y\|_2}{h} \right) \mid A = a \right\},$$

and for an observational study under assumptions (C1), (C3), (C4) we have

$$q^a_h(y) = \mathbb{E}\left\{ \mathbb{E}\left[\frac{1}{h^d} K\left(\frac{\|Y - y\|_2}{h} \right) \Big| X, A = a \right] \right\},$$

[1] For example, $\sup_{q^a \in \Sigma(\beta,L)} |q^a_h - q^a| = O(h^\beta)$ where $\Sigma(\beta,L)$ is a Hölder class of functions with constants $\beta > 0, L$ [e.g., 131].

and consequently we will obtain identifying expressions for $D(Q^1_{h_1}, Q^0_{h_0})$ as previously by simply replacing the identifying expressions for q^a in (3.3) - (3.5) with the ones above for q^a_h, where $Q^a_{h_a}$ is the distribution having the density $q^a_{h_a}$.

Throughout, we base our analysis on a fixed bandwidth case, following for example Chen et al. [17], Rinaldo and Wasserman [107], Chazal et al. [16]. Using a fixed bandwidth provides several advantages. First, we do not need strong smoothness assumptions about the form of the density. In fact, the kernel-smoothed density can exist even if Y^a itself does not have a density in the usual sense Rinaldo and Wasserman [107]. Second, fixed bandwidths may more closely mirror practical data analysis, since we typically face a single dataset with a particular sample size, rather than a sequence of datasets of increasing size. Finally, with a fixed bandwidth we avoid the need for any impractical undersmoothing to remove bias Wasserman [146], and can achieve faster rates of convergence towards the smoothed parameter. We aim to consider varying bandwidth analyses in future work.

3.2.2 Bootstrap and Stochastic Convergence of Empirical Process

For building valid confidence intervals for further quantifying randomness of our estimates, we use theories on the bootstrap method and an empirical process. Introduced in [30], Bootstrapping is a method for estimating the variance of an estimator and thus for finding confidence intervals. When the target parameter is nonparametric, such as the causal effect defined in (3.2), the stochastic convergence of an empirical process is required to guarantee an asymptotic validity of the bootstrap procedure. The rest of this subsection is devoted to provide a brief review for techniques in stochastic convergence of empirical process that are essential to construct confidence intervals in Section 3.4, and readers who are not interested in the details may skip the rest of the section. We refer to [139, 78] and reference therein for further details.

Suppose an i.i.d sample $(Z_1,...,Z_n) \sim \mathbb{P}$ on $\mathscr{Z}$, and let $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ be the empirical measure. Let $(Z_1^*,\ldots,Z_n^*)$ be the bootstrapped samples, i.e. a set of samples with replacement from the original sample $(Z_1,\ldots,Z_n)$, and let $\mathbb{P}_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i^*}$ be the bootstrap empirical measure. Bootstrapping is used to infer information of unknown measure $\mathbb{P}_n - \mathbb{P}$ by known and computable measure $\mathbb{P}_n^* - \mathbb{P}_n$.

One theoretical guarantee for bootstrap is that $\sqrt{n}(\mathbb{P}_n - \mathbb{P})$ and $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$ converges to same Brownian Bridge. Let $\mathscr{F} \subset \mathbb{R}^{\mathscr{Z}}$ be a class of measurable functions. We let $\ell_\infty(\mathscr{F})$ be the collection of all bounded functions $\phi : \mathscr{F} \to \mathbb{R}$ equipped with the sup norm (or uniform norm) $\|\cdot\|_\infty$. A random measure μ is understood in $\ell_\infty(\mathscr{F})$ as $\mu(f) = \int f d\mu$. For random measures $\{\mu_n\}_{n\in\mathbb{N}}$ and μ, we say $\mu_n \to \mu$ weakly in $\ell_\infty(\mathscr{F})$ if and only if

$\mathbb{E}[\phi(\mu_n)] \to \mathbb{E}[\phi(\mu)]$ for every bounded continuous map $\phi : \ell_\infty(\mathscr{F}) \to \mathbb{R}$. Now we have the following theorem for convergence of the bootstrap.

Theorem 3.2.1. *(Gine and Zinn [40, Theorem 2.4], Kosorok [78, Theorem 2.6])* $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \to \mathbb{G}$ *weakly in* $\ell_\infty(\mathscr{F})$ *if and only if* $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n) \to \mathbb{G}$ *a.s. weakly in* $\ell_\infty(\mathscr{F})$ *for a limit process* $\mathbb{G}$. *If either convergence happens, the limit process* $\mathbb{G}$ *is a centered Gaussian process with* $Cov[\mathbb{G}(f), \mathbb{G}(g)] = \int fg d\mathbb{P} - \int f d\mathbb{P} \int g d\mathbb{P}$ *for* $f, g \in \mathscr{F}$.

Therefore, once $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \to \mathbb{G}$ weakly in $\ell_\infty(\mathscr{F})$ is shown, Theorem (3.2.1) implies that $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n) \to \mathbb{G}$ weakly in $\ell_\infty(\mathscr{F})$ a.s. as well, and the unknown measure $\sqrt{n}(\mathbb{P}_n - \mathbb{P})$ can be asymptotically approximated by know and computable measure $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$. One way to show $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \to \mathbb{G}$ weakly in $\ell_\infty(\mathscr{F})$ (i.e. $\mathscr{F}$ is $\mathbb{P}$-Donsker) is to use the bracketing entropy argument as detailed in, for example, Van Der Vaart and Wellner [142, Chapter 2.5].

3.3 Proposed Estimator and Error Analysis

3.3.1 Single-source randomized study

To estimate the counterfactual density function q^a - we propose a conditional kernel density estimator defined as

$$\widehat{q^a_{h_a}}(y) = \frac{\mathbb{1}(n_a > 0)}{n_a} \sum_{j=1}^{n} \frac{1}{h_a^d} K\Big(\frac{||y - Y_j||_2}{h_a}\Big) \mathbb{1}(A_j = a), \tag{3.7}$$

where K is a valid kernel function with a fixed bandwidth $h_a > 0$ and $n_a = \sum_i \mathbb{1}(A_i = a)$. In practice, one may set $h_1 = h_0 = h$. Given this estimator, we have the following identity.

Proposition 3.3.1. *Under the causal assumptions (C1), (C2),* $\widehat{q^a_h}$ *defined in* (3.7) *satisfies that*

$$\mathbb{E}[\widehat{q^a_h} \mid A_i = a, \forall i] = q^a_h.$$

The proof of the above proposition is given in Section B.2.1 in the appendix. Now we propose the following plug-in estimator for the single-source randomized study

$$D(\widehat{Q^1_{h_1}}, \widehat{Q^0_{h_0}}), \tag{3.8}$$

where $\widehat{Q^a_{h_a}}$ denotes a distribution induced from $\widehat{q^a_{h_a}}$.

To evaluate performance of the proposed estimator, we aim to upper bound the L_1 risk (mean absolute deviation) of our plug-in estimator defined by

$$\mathbb{E}\left[\left|D(\widehat{Q^1_{h_1}},\widehat{Q^0_{h_0}})-D(Q^1_{h_1},Q^0_{h_0})\right|\right]. \tag{3.9}$$

To proceed, first we need to bound $\mathbb{E}[D(Q^a_h,\widehat{Q^a_h})]=\mathbb{E}[\|q^a_h-\widehat{q^a_h}\|_1]$, the L_1 risk of our kernel density estimator. In what follows, we make two basic assumptions on the counterfactual distribution Q^a and on the kernel function K to construct the kernel density estimator in (3.7).

- (A1) *Bounded density and support of the counterfactual distribution*: Probability distribution Q^a has a density $q^a=\frac{dQ^a}{d\lambda_d}$ with respect to the Lebesgue measure λ_d on $\mathbb{R}^d$ where $q^a\leq q_{\max}<\infty$, and is supported on a compact set $\mathscr{Y}\subset\mathbb{R}^d$.

- (A2) *Finite L_2 norm and bounded support of the kernel function*: The kernel function $K:\mathbb{R}^d\to\mathbb{R}$ has finite L_2 norm $\|K\|_2:=\sqrt{\int K(u)^2du}<\infty$ and has a bounded support, i.e. there exists $R_K<\infty$ with $\mathrm{supp}(K)\subset\mathbb{B}_{R_K}(0)$, where $\mathbb{B}_{R_K}(0)=\{u\in\mathbb{R}^d:\|u\|_2\leq R_K\}$.

Both the assumptions (A1) and (A2) are all weak and commonly found in nonparametric statistics. Now the following lemma gives upper bound of $\mathbb{E}[D(Q^a_h,\widehat{Q^a_h})]$.

Lemma 3.3.1. *Let $\widehat{Q^a_h}$ denote estimated distribution for true distribution Q^a_h under treatment $A=a$ with kernel bandwidth h. Then for $Z=(A,Y)\sim\mathbb{P}$, under the assumptions (A1) and (A2), we have*

$$\mathbb{E}[D(Q^a_h,\widehat{Q^a_h})]=O\left(\frac{1}{\sqrt{n\pi_a h^d}}\right) \tag{3.10}$$

where $\pi_a=\mathbb{P}(A=a)$.

The proof of Lemma 3.3.1 is given in Section B.2.2 of the appendix. Consequently, we have the following theorem regarding the upper bound of the L_1 risk (3.9).

Theorem 3.3.1 (L_1 risk of the estimator $D(\widehat{Q^1_{h_1}},\widehat{Q^0_{h_0}})$)**.** *Under assumptions (A1) and (A2),*

$$\mathbb{E}\left[\left|D(\widehat{Q^1_{h_1}},\widehat{Q^0_{h_0}})-D(Q^1_{h_1},Q^0_{h_0})\right|\right]=O\left(\frac{1}{\sqrt{n\pi_1h_1^d}}+\frac{1}{\sqrt{n\pi_0h_0^d}}\right). \tag{3.11}$$

The proof will be given in Section B.2.3 of the appendix. The above theorem shows that having the kernel bandwidth fixed our error vanishes at $\sqrt{n}$ rates.

3.3.2 Multi-source randomized study

For the multi-source randomized experiment where we have $Z = \left((A,Y)_{\mathbb{P}_i}, \mathbb{P}_i\right)$, $\mathbb{P}_i \sim \mathscr{P}$ for $i = 1,...,N$, to estimate the target parameter $\mathbb{E}_{\mathscr{P}}[D(Q^1_{h_1}, Q^0_{h_0})]$ we propose the sample mean of plug-in estimators

$$\frac{1}{N}\sum_{i=1}^{N} D\left(\widehat{(Q^1_{h_1})_i}, \widehat{(Q^0_{h_0})_i}\right) \tag{3.12}$$

, where each $\widehat{(Q^a_{h_a})_i}$ is an estimated counterfactual distribution for assignment $A = a$ and subpopulation $\mathbb{P}_i$ via the kernel density estimator in (3.7) with a prespecified bandwidth h_a. Thus now we are interested in upper bounding the L_1 risk

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} D\left(\widehat{(Q^1_{h_1})_i}, \widehat{(Q^0_{h_0})_i}\right) - \mathbb{E}_{\mathscr{P}}[D(Q^1_{h_1}, Q^0_{h_0})]\right|\right]. \tag{3.13}$$

The following theorem provides the error bound of (3.13).

Theorem 3.3.2 (L_1 *risk of the estimator* $\frac{1}{N}\sum_{i=1}^{N} D\left(\widehat{(Q^1_{h_1})_i}, \widehat{(Q^0_{h_0})_i}\right)$). *Under assumptions (A1) and (A2),*

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} D\left(\widehat{(Q^1_{h_1})_i}, \widehat{(Q^0_{h_0})_i}\right) - \mathbb{E}_{\mathscr{P}}[D(Q^1_{h_1}, Q^0_{h_0})]\right|\right] \\
&= O\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{n_i\pi_{1,i}h^d_{1,i}}} + \frac{1}{\sqrt{n_i\pi_{0,i}h^d_{0,i}}}\right) + \frac{\sigma_{\mathscr{P}}}{\sqrt{N}}\right)
\end{aligned} \tag{3.14}$$

where n_i, $h_{a,i}$ is the total number samples and the bandwidth used for kernel density estimation for a subpopulation $\mathbb{P}_i$ respectively, and $\sigma_{\mathscr{P}} = \sqrt{Var_{\mathscr{P}}\left[D(Q^1_{h_1}, Q^0_{h_0})\right]}$. In particular, when $n_i = n$, $h_{a,i} = h_a$, $\pi_{1,i} = \pi_1$, $\pi_{0,i} = \pi_0$ for all i, then

$$\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N} D\left(\widehat{(Q^1_{h_1})_i}, \widehat{(Q^0_{h_0})_i}\right) - \mathbb{E}_{\mathscr{P}}[D(Q^1_{h_1}, Q^0_{h_0})]\right|\right] = O\left(\frac{1}{\sqrt{nh^d_1\pi_1}} + \frac{1}{\sqrt{nh^d_0\pi_0}} + \frac{\sigma_{\mathscr{P}}}{\sqrt{N}}\right).$$

We give the proof of Theorem 3.3.2 in Section B.2.4 of the appendix. Notice that the error bound in (3.14) consists of two parts. The first part is simply the average estimation error over N different single-source randomized experiments. The second part is related to the heterogeneity of treatment effects across subpopulations and will be negligible if $D_{\mathbb{P}_i}(Q^1_{h_1}, Q^0_{h_0})$ does not vary too much across $\mathbb{P}_i$'s.

3.3.3 Observational study

An observational study requires more careful argument to develop the estimator. Since the identifying expression (3.5) contains conditional densities not only depending on $Y \in \mathbb{R}^d$ but also potentially high-dimensional $X \in \mathbb{R}^k$, a plug-in estimator might yield impractically slow convergence rates. Recall that under the causal assumptions (C1), (C3), (C4), from (3.6) we have

$$\mathbb{E}\left\{\mathbb{E}\left[T_{h,y}(Y)\middle|X,A=a\right]\right\} = q_h^a(y)$$

for a given h, where we define $T_{h,y}(Y) = \frac{1}{h^d}K\left(\frac{\|Y-y\|_2}{h}\right)$. In what follows we propose a novel doubly robust style estimator for $q_h^a(y)$ by

$$\widehat{\psi}_h^a(y) = \mathbb{P}_n\left\{\frac{\mathbb{1}(A=a)}{\widehat{\pi}_a(X)}\Big(T_{h,y}(Y) - \widehat{\mu}_A(X)\Big) + \widehat{\mu}_a(X)\right\}, \tag{3.15}$$

where $\pi_a(X) = \mathbb{P}(A=a|X)$, $\mu_A(X) = \mathbb{E}[T_{h,y}(Y)|A,X]$, $\mu_a(X) = \mathbb{E}[T_{h,y}(Y)|A=a,X]$ and $\widehat{\pi}_a(X)$, $\widehat{\mu}_A(X)$, $\widehat{\mu}_a(X)$ denote their estimates respectively. The estimator in (3.15) resembles the doubly robust (or semiparametric) estimator for the ATE, where we have replaced Y in the original estimator with $T_{h,y}(Y)$. The doubly robust estimator is known to be efficient, model-free estimators, in the sense that they achieve $\sqrt{n}$-consistency and asymptotic normality even when π and μ are estimated flexibly with nonparametric models, without committing a priori to particular estimators or function classes. For more details on doubly robust estimators and related topics, we refer the interested readers to [12, 130, 137, 65].

Next we propose our estimator for observational studies by

$$D(\widehat{Q_h^1}, \widehat{Q_h^0}) \tag{3.16}$$

where $\widehat{Q_h^0}$ a distribution induced by $\widehat{\psi}_h^a$ in (3.15). As will be verified shortly in Lemma 3.3.2, construction based on $\widehat{\psi}_h^a$ endows the same kind of double robustness property which can be found in the ordinary doubly robust estimator for the ATE to our estimator. For the sake of simplicity here we use a single bandwidth h, acknowledging that we can also proceed with the different bandwidths as previously. Hereafter, for a given function f, we use the notation $\|f\|_q = (\int |f(z)|^q d\mathbb{P}(z))^{\frac{1}{q}}$ to denote the $L_q(\mathbb{P})$-norm of f. Before formally stating the theorem, we enumerate additional assumptions we need as below.

- (B1) *Convergence rate of nuisance parameters*: Let $\overline{\pi}_a$ and $\overline{\mu}_a$ denote fixed functions to which $\widehat{\pi}_a$ and $\widehat{\mu}_a$ converge in the sense that $\|\widehat{\pi}_a - \overline{\pi}_a\|_2 = o_{\mathbb{P}}(r(n))$ and $\|\widehat{\mu}_a - \overline{\mu}_a\|_2 = o_{\mathbb{P}}(s(n))$. We require $r(n)s(n) = O(n^{-\frac{1}{2}})$ but we only require *either* $\overline{\pi}_a = \pi_a$ *or* $\overline{\mu}_a = \mu_a$ where π_a and μ_a are true parameters.

- (B2) *Uniform boundedness*: $\|1/\widehat{\pi}_a\|_\infty$, and $\|\widehat{\mu}_a\|_2$ are uniformly bounded.
- (B3) *Sample splitting*: The estimators for nuisance functions $(\widehat{\pi}_a, \widehat{\mu}_a)$ are computed in a separate independent sample.

Note that all the extra assumptions (B1) - (B3) are quite weak as well. Assumption (B1), the main substantive assumption in this subsection, says that at least one of the estimators $\widehat{\pi}_a$ or $\widehat{\mu}_a$ must be consistent for the true π_a or μ_a in terms of the L_2 norm at the rate of $o_{\mathbb{P}}(s(n)), o_{\mathbb{P}}(r(n))$ respectively. Since only one of the nuisance estimators is required to be consistent (not necessarily both), our estimator shows double robustness (see Lemma B.2.3 in the appendix). Compared to naive plug-in estimators where $\sqrt{n}$ rates are never attainable in nonparametric models, the requirement on double rates of $s(n)r(n) = n^{-\frac{1}{2}}$ brings significant improvement since under reasonable structural assumptions on regression functions (e.g., sparsity) $\sqrt{n}$ rates are attainable through many nonparametric methods. One sufficient condition for (B1) would be $s(n) = n^{-\frac{1}{4}}$ and $r(n) = n^{-\frac{1}{4}}$. Assumption (B2) involves a minimal regularity condition on the reciprocal of estimator $\widehat{\pi}_a$ and its limit $\overline{\pi}_a$. Assumption (B3) enables us to accommodate the added complexity from estimating both nuisance functions and ψ_h^a without relying on complicated empirical process conditions (e.g., Donsker condition) [18, 66].

Now we give our result on error analysis. The next lemma describes how accurately we can approximate the true counterfactual distribution via the proposed density estimator (3.15).

Lemma 3.3.2. *Let $\overline{\pi}_a, \overline{\mu}_a$ be fixed functions to which $\widehat{\pi}_a$ and $\widehat{\mu}_a$ asymptotically converge. Then under assumptions (A1), (A2), (B2), and (B3), together with the causal assumptions (C1), (C3), (C4), we have*

$$\mathbb{E}\left[D(\widehat{Q_h^a}, Q_h^a)\right] = O\left(\frac{1}{\sqrt{n}}\right) + \|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2. \tag{3.17}$$

Lemma 3.3.2 is elaborated in Lemma B.2.6 in the appendix in more details. Importantly, to the best of our knowledge this result has not yet appeared in nonparametric counterfactual density estimation. One may notice that the product of two L_2 norms in (3.17) becomes negligible under Assumption (B1). Now we provide our main theorem for this subsection.

Theorem 3.3.3 (L_1 *risk of the estimator* $\widehat{D(Q_h^1, Q_h^0)}$ *for observational study*)**.** *Under assumptions (A1), (A2), (B1), (B2), and (B3), together with the causal assumptions (C1), (C3), (C4), we have*

$$\mathbb{E}\left[\left|\widehat{D(Q_h^1, Q_h^0)} - D(Q_h^1, Q_h^0)\right|\right] = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) \tag{3.18}$$

where we use random sample splitting so we estimate ψ_h^a and $(\widehat{\pi}_a, \widehat{\mu}_a)$ on separate sample sets.

The proof of Theorem 3.3.3 requires more involved argument than previous randomized experiments and is given in B.2.5 of the appendix. The result of Theorem 3.3.3 implies that the proposed estimator can be estimated at the fast $\sqrt{n}$ rates as well, even when all the nuisance parameters μ, π are estimated flexibly at much slower rates than $\sqrt{n}$.

Remark 2. *1 . (Bandwidth selection) For each of the proposed estimators, the fixed kernel smoothing bandwidth h must be specified in advance through a separate procedure. This turns out to be a very challenging problem; since we do not have "ground truth", standard approaches such as cross-validation cannot be applied. Here, we are basically being agnostic about the optimal bandwidth choice rule as it is beyond what we focus on in this paper. Instead, we proceed on an ad hoc basis as follows.*

1. *We generate an artificial dataset that resembles the observed data distribution but has all the counterfactual outcomes. For example, for the given $\mathsf{D}^{obs} = \{A_i, X_i, Y_i\}_{i=1}^n$, we generate an artificial one $\mathsf{D}^{af} = \{A_i, X_i, Y_i^0, Y_i^1\}_{i=1}^n$. This can be done via something akin to matching [1], for instance.*
2. *Then we estimate $D(\widehat{Q_h^1}, \widehat{Q_h^0})$, $D(Q_h^1, Q_h^0)$ using D^{obs}, D^{af} respectively.*
3. *We find h that minimizes the mean squared error.*

The procedure described above should be done on the separate dataset, whence we compute the proposed estimators. Again, note that the method is pretty much ad hoc without formal validity. The optimal bandwidth choice problem can be tackled via theoretical analysis, which also would be closely related to another interesting future work. For example, we conjecture our proposed estimators in Section 3.3 may be minimax optimal when the bandwidth is tuned in a particular way.

3.4 Asymptotic Convergence and Confidence Interval

3.4.1 Asymptotic Convergence

In the previous section we described effective ways to estimate the counterfactual density and our target causal effect in various scenarios and analyzed their error rates. In this section we characterize an asymptotic behavior of our proposed counterfactual density estimators, and delineate how to construct a confidence interval via bootstrapping. To this end we require slightly stronger version of the previous assumptions (A2), (B2) as follows.

- (A2') *Finite L_∞ norm, Lipschitz, and bounded support of the kernel function*: The kernel function $K : \mathbb{R}^d \to \mathbb{R}$ has finite L_∞ norm $\|K\|_\infty := \sup_u |K(u)| < \infty$ and has a bounded support, i.e. there exists $R_K < \infty$ with $\mathrm{supp}(K) \subset \mathbb{B}_{R_K}(0)$, where $\mathbb{B}_{R_K}(0) = \{u \in \mathbb{R}^d : \|u\|_2 \leq R_K\}$. Also, K is Lipschitz, i.e. there exists $L_K < \infty$ with $|K(u_1) - K(u_2)| \leq L_K \|u_1 - u_2\|$.
- (B2') *Uniform boundedness*: $\|1/\widehat{\pi}_a\|_\infty$, $1/\overline{\pi}_a$ and $\|\widehat{\mu}_a\|_2$ are uniformly bounded.

Note that in (A2') the bounded norm and bounded support in above assumptions are still considered mild. Also, the smoothness condition on the kernel function is commonly found in nonparametric literature [e.g., 131].

In the next theorem we characterize an asymptotic property of the proposed counterfactual density estimator for the single-source randomized study (3.7). Hereafter, we use $\rightsquigarrow$ for denoting convergence in distribution [2].

Theorem 3.4.1. *Under assumptions (A1), (A2'), for a treatment assignment $A = a$ we have*

$$\sqrt{n} D(\hat{Q}_h^a, Q_h^a) \rightsquigarrow \frac{1}{\pi_a} \int |\mathbb{G}(y) - q_h^a(y)\mathbb{G}(a)| \, dy,$$

where $\mathbb{G}$ is a centered Gaussian process [3] with $Cov[\mathbb{G}(y_1), \mathbb{G}(y_2)] = \pi_a \mathbb{E}[T_{h,y_1}(Y) T_{h,y_2}(Y)] - \pi_a^2 q_h(y_1) q_h(y_2)$, $Cov[\mathbb{G}(y), \mathbb{G}(a)] = \pi_a(1-\pi_a) q_h(y)$, and $Var[\mathbb{G}(a)] = \pi_a(1-\pi_a)$, where we write $T_{h,y}(Y) = \frac{1}{h^d} K\left(\frac{\|Y-y\|_2}{h}\right)$ and $q_h(y) = \mathbb{E}[T_{h,y}(Y)]$.

The proof of Theorem 3.4.1 is given in Section B.2.6 of the appendix. Characterization of the asymptotic behavior for observational study appears a little bit different, as stated in the next theorem.

Theorem 3.4.2. *Under the assumptions (A1), (A2'), (B1), (B2'), (B3), it follows*

$$\sqrt{n} D(\hat{Q}_{h_a}^a, Q_{h_a}^a) \rightsquigarrow \int |\mathbb{G}(y)| dy,$$

where $\mathbb{G}$ is a centered Gaussian process with $Cov[\mathbb{G}(y_1), \mathbb{G}(y_2)] = \mathbb{E}\left[f_{y_1}^a f_{y_2}^a\right] - \mathbb{E}\left[f_{y_1}^a\right] \mathbb{E}\left[f_{y_2}^a\right]$, $f_y^a := \frac{\mathbf{1}(A=a)}{\overline{\pi}_a(X)} \left(T_{h,y}(Y) - \overline{\mu}_A(X)\right) + \overline{\mu}_a(X)$. $\overline{\pi}$, $\overline{\mu}$ are defined in Lemma 3.3.2.

The proof of Theorem 3.4.2 is given in Section B.2.7 of the appendix. Theorem 3.4.1 and 3.4.2 lay the foundation to construct confidence intervals for the proposed estimators as detailed in the next section.

[2] In our context, weak convergence is equivalent to "convergence in distribution" or "convergence in law". We sometimes interchangeably use those terms in our paper in order to conserve the original statement in the theorems that we cite from other literature.

[3] Here, the index set is understood as a multiset $(\mathscr{Y} \cup \mathscr{A}, m)$ where the multiplicity $m = 2$ only for elements in $\mathscr{Y} \cap \mathscr{A}$ so that we can use the indices $y \in \mathscr{Y}$ and $a \in \mathscr{A}$ together.

3.4.2 Confidence Interval via Bootstrapping

In this section we present a bootstrap approach to constructing confidence interval for each of the proposed estimators. For $\alpha \in (0,1)$, a $1-\alpha$ confidence interval $\hat{C}_\alpha$ for our target parameter θ is an interval satisfying

$$\liminf_{n\to\infty} \mathbb{P}(\theta \in \hat{C}_\alpha) \geq 1-\alpha,$$

where θ is the estimator for single- and multi-source randomized study, and observational study respectively, as presented in (3.8), (3.12), and (3.16).

We construct the confidence interval $\hat{C}_\alpha$ centered at the causal estimator $\hat{\theta}$ and of width $2c_n$, i.e., $\hat{C}_\alpha = [\hat{\theta} - c_n, \hat{\theta} + c_n]$, where $\hat{\theta} = D(\widehat{Q^1_{h_1}}, \widehat{Q^0_{h_0}})$ for the single-source randomized study or observational study ($h_0 = h_1$), and $\hat{\theta} = \frac{1}{N}\sum_{i=1}^N D((\widehat{Q^1_{h_1}})_i, (\widehat{Q^0_{h_0}})_i)$ for the multi-source randomized study.

Then $\hat{C}_\alpha$ is a valid $1-\alpha$ asymptotic confidence set if and only if

$$\liminf_{n\to\infty} \mathbb{P}(|\hat{\theta} - \theta| \leq c_n) \geq 1-\alpha. \tag{3.19}$$

We use bootstrapping to compute the confidence interval. Algorithms detailing how to compute $\hat{C}_\alpha$ for each of the proposed estimators are given in Algorithm 1, 2, and 3 in the following page. Proposed bootstrapping algorithms provide a straightforward way to derive estimates of the radius of the interval c_n and are simple to implement in practice.

As briefly described in Section 3.2.2, the validity of the bootstrap confidence interval is based on the stochastic convergence of the empirical process. Suppose we have an original i.i.d sample set $(Z_1, ..., Z_n) \sim \mathbb{P}$ and a bootstrapped set $(Z_1^*, \ldots, Z_n^*)$, and their empirical measures $\mathbb{P}_n$, $\mathbb{P}_n^*$ respectively. The main theory that underpins our bootstrapping algorithm is that the empirical process and its bootstrapped version converge to the same limiting distribution (Theorem 3.2.1). For example, one sufficient condition we find in Algorithm 1 for (3.19) to hold is $\liminf_{n\to\infty} \mathbb{P}\left(\sqrt{n}D(\widehat{Q^a_{h_a}}, Q^a_{h_a}) \leq \hat{z}^a_{\alpha/2}\right) \geq 1 - \frac{\alpha}{2}$. Thus for the case of single-source randomized study, it suffices to prove that $\sqrt{n}D(\widehat{Q^a_h}, Q^a_h)$ and $\sqrt{n}D(\widehat{Q^a_h}^*, \widehat{Q^a_h})$ converge to the same limiting distribution. Convergence of $\sqrt{n}D(\widehat{Q^a_h}, Q^a_h)$ can be obtained from Theorem 3.4.1, and applying Theorem 3.2.1 implies that indeed $\sqrt{n}D(\widehat{Q^a_h}^*, \widehat{Q^a_h})$ converges to the same limiting distribution. We can show validity of other bootstrapping algorithms basically in the same manner as well. These results are summarized in the following theorem.

Theorem 3.4.3. *Under assumptions (A1), (A2') for single- and multi-source randomized study, and under assumptions (A1), (A2'), (B1), (B2'), (B3) for observational study, corresponding confidence intervals $\hat{C}_\alpha$ constructed via Algorithm 1, 2, 3 are valid confidence* *intervals, i.e.*

$$\liminf_{n\to\infty} \mathbb{P}\left(\theta \in \hat{C}_\alpha\right) \geq 1-\alpha$$

for given level of α*.*

We give a proof of the above Theorem in section B.2.8, B.2.9, and B.2.10 of the appendix.

Algorithm 1 Bootstrapping algorithm for single-source randomized study.

```
1. We generate $B$ bootstrap samples $\{\tilde{Z}_1^1, \ldots, \tilde{Z}_n^1\}, \ldots, \{\tilde{Z}_1^B, \ldots, \tilde{Z}_1^B\}$, by sampling with replacement from the original sample.
2. On each bootstrap sample, compute $T_i^a = \sqrt{n} D(\widehat{Q_{h_a}^a}^i, \widehat{Q_{h_a}^a})$, where $\widehat{Q_{h_a}^a}^i$ is the estimated distribution of kernel density estimator $\widehat{Q_{h_a}^a}$ computed on $i$th bootstrap samples $\{\tilde{Z}_1^i, \ldots, \tilde{Z}_n^i\}$.
3. Compute $\frac{\alpha}{2}$-quantile $\hat{z}_{\alpha/2}^a = \inf\left\{z : \frac{1}{B}\sum_{i=1}^B I(T_i^a > z) \leq \frac{\alpha}{2}\right\}$.
4. Define $\hat{C}_\alpha = \left[D(\widehat{Q_{h_a}^1}, \widehat{Q_{h_a}^0}) - \frac{\hat{z}_{\alpha/2}^0}{\sqrt{n}} - \frac{\hat{z}_{\alpha/2}^1}{\sqrt{n}}, D(\widehat{Q_{h_a}^1}, \widehat{Q_{h_a}^0}) + \frac{\hat{z}_{\alpha/2}^0}{\sqrt{n}} + \frac{\hat{z}_{\alpha/2}^1}{\sqrt{n}}\right]$.
```

Algorithm 2 Bootstrapping algorithm for multi-source randomized study.

1. For each $i = 1,\dots,N$, we generate ith bootstrap samples $\{Z^*_{\mathbb{P}_i,1},\dots,Z^*_{\mathbb{P}_i,n^i}\}$ by sampling with replacement from the ith original sample $\{Z_{\mathbb{P}_i,1},\dots,Z_{\mathbb{P}_i,n_i}\}$.
2. On each bootstrap sample $\{Z^*_{\mathbb{P}_i,1},\dots,Z^*_{\mathbb{P}_i,n^i}\}$, compute $D_i^a = \sqrt{n}D((\widehat{Q^a_{h_a}})^*_i,(\widehat{Q^a_{h_a}})_i)$, where $(\widehat{Q^a_{h_a}})^*_i$ is the estimated distribution of kernel density estimator $\widehat{Q^a_{h_a}}$ computed on ith bootstrap samples $\{Z^*_{\mathbb{P}_i,1},\dots,Z^*_{\mathbb{P}_i,n^i}\}$.
3. Compute $\bar{D}^a = \frac{1}{N}\sum_{i=1}^N D_i^a$.
4. We generate B bootstrap distributions $\{\mathbb{P}_1^{(1)},\dots,\mathbb{P}_N^{(1)}\},\dots,\{\mathbb{P}_1^{(B)},\dots,\mathbb{P}_N^{(B)}\}$, by sampling with replacement from the original distribution $\{\mathbb{P}_1,\dots,\mathbb{P}_N\}$.
5. On each bootstrap sample, compute

$$T_j = \left|\frac{1}{\sqrt{N}}\sum_{i=1}^N D((\widehat{Q^1_{h_1}})_i^{(j)},(\widehat{Q^0_{h_0}})_i^{(j)}) - \frac{1}{\sqrt{N}}\sum_{i=1}^N D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)\right|,$$

   , where $(\widehat{Q^a_{h_a}})_i^{(j)}$ is the estimated distribution of kernel density estimator $\widehat{Q^a_{h_a}}$ computed on the sample $\{Z_{\mathbb{P}_i^{(j)},1},\dots,Z_{\mathbb{P}_i^{(j)},n}\}$ from the bootstrapped distribution $\mathbb{P}_i^{(j)}$.
6. Compute α-quantile $\hat{z}_\alpha = \inf\left\{z : \frac{1}{B}\sum_{j=1}^B I(T_j > z) \le \alpha\right\}$.
7. Define

$$\begin{aligned}\hat{C}_\alpha = \Bigg[&\frac{1}{N}\sum_{i=1}^N D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i) - \frac{\bar{D}^1}{\sqrt{n}} - \frac{\bar{D}^0}{\sqrt{n}} - \frac{\hat{z}_\alpha}{\sqrt{N}},\\ &\frac{1}{N}\sum_{i=1}^N D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i) + \frac{\bar{D}^1}{\sqrt{n}} + \frac{\bar{D}^0}{\sqrt{n}} + \frac{\hat{z}_\alpha}{\sqrt{N}}\Bigg].\end{aligned}$$

Algorithm 3 Bootstrapping algorithm for observational study.

1. We generate B bootstrap samples $\{\tilde{Z}_1^1,\dots,\tilde{Z}_n^1\},\dots,\{\tilde{Z}_1^B,\dots,\tilde{Z}_1^B\}$, by sampling with replacement from the original sample.
2. On each bootstrap sample, compute $T_i^a = \sqrt{n}D(\widehat{Q^a_{h_a}}^i,\widehat{Q^a_{h_a}})$, where $\widehat{Q^a_{h_a}}^i$ is the estimated distribution of kernel density estimator $\widehat{Q^a_{h_a}}$ computed on ith bootstrap samples $\{\tilde{Z}_1^i,\dots,\tilde{Z}_n^i\}$.
3. Compute $\frac{\alpha}{2}$-quantile $\hat{z}^a_{\alpha/2} = \inf\{z : \frac{1}{B}\sum_{i=1}^B I(T_i^a > z) \le \frac{\alpha}{2}\}$.
4. Define $\hat{C}_\alpha = \left[D(\widehat{Q^1_{h_a}},\widehat{Q^0_{h_a}}) - \frac{\hat{z}^0_{\alpha/2}}{\sqrt{n}} - \frac{\hat{z}^1_{\alpha/2}}{\sqrt{n}}, D(\widehat{Q^1_{h_a}},\widehat{Q^0_{h_a}}) + \frac{\hat{z}^0_{\alpha/2}}{\sqrt{n}} + \frac{\hat{z}^1_{\alpha/2}}{\sqrt{n}}\right]$.

3.5 Numerical Illustration

Here we present a series of simulation studies using both synthetic and real-world data to illustrate our method. We consider three different setups; we generate two synthetic datasets for a single-source experiment and one for a multi-source experiment, and use the real-world data for an observational study. For each setup, we illustrate how the proposed estimator can uncover clues on the distributional shift induced by a given treatment, which otherwise would not have been revealed by traditional methods.

3.5.1 Single-source experiment

For a single-source experiment, we generate two pairs of counterfactual distributions having the exact same mean as illustrated in Figure 3.1. The first pair consists of two beta distributions and the second consists of a univariate Gaussian and a mixture of two Gaussian distributions. To generate data (A,Y) from each pair of distributions, we randomly sample 100 points from each of q^0, q^1 respectively. In both cases, we set $\mathbb{P}(A=1)=1/2$.

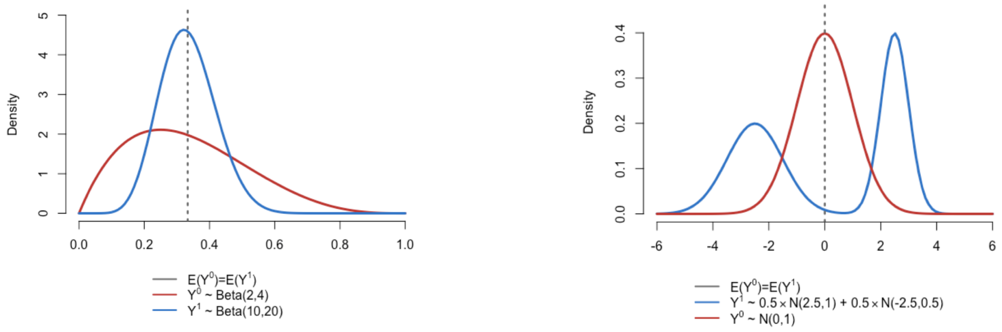


Fig. 3.1 Two pairs of counterfactual distributions that have the same mean. In each pair, the distribution for the treated is largely different from the control, for example in terms of the variance, skewness, and (the number of) mode.

Then we estimate causal effects defined in (3.2) using the proposed estimator. For baseline methods, we use the difference-in-means ($\hat{\psi}_{\text{diff}}$) and Horvitz-Thompson estimators ($\hat{\psi}_{\text{HT}}$), two of most widely used methods in randomized experiments, whose target parameter is the ATE defined in (3.1). Given dataset (A,Y) and known $\pi=\mathbb{P}(A=1)$, the two baseline estimators are defined as below.

$$\hat{\psi}_{\text{diff}} = \mathbb{P}_n[Y|A=1] - \mathbb{P}_n[Y|A=0] = \frac{\mathbb{P}_n[AY]}{\mathbb{P}_n[A]} - \frac{\mathbb{P}_n[(1-A)Y]}{\mathbb{P}_n[1-A]}$$

$$\hat{\psi}_{\text{HT}} = \mathbb{P}_n[\frac{AY}{\pi} - (\frac{1-A}{1-\pi})Y]$$

Even though the mean of the two counterfactual distributions in each pair in Figure 3.1 is the same, the given treatment brings a substantial change to shape of the distribution. In the second example of the unimodal and bimodal distributions, this becomes much more obvious; we also have a considerable degree of the effect heterogeneity in this example. By construction, the baseline estimators whose target parameter is (3.1) estimate zero effects for both cases.

We present the value of the two baseline estimators and our proposed estimator defined in (3.8) together with the 95% confidence interval in Table 3.1, using the synthetic data distributions described in Figure 3.1. For bootstrapping we use $B = 100$. When computing the proposed estimator the numerical integration is done via Monte Carlo with uniform sampling, and we use bandwidth $h_0 = h_1 = 0.005$ for kernel density estimation. As expected, all the baseline estimators report that treatment effects are insignificant, whereas the proposed estimator gives a significant clue on a substantial shift in counterfactual distribution.

	Two beta distributions		**Uni- vs. Bi-modal**	
Estimator	Point Estimation	95% CI	Point Estimation	95% CI
Difference-in-means	0.002	$(-0.015, 0.020)$	0.013	$(-0.189, 0.215)$
Horvitz-Thompson	0.014	$(-0.030, 0.059)$	0.012	$(-0.188, 0.212)$
Difference-in-distribution	**0.735**	$\mathbf{(0.718, 0.752)}$	**0.311**	$\mathbf{(0.250, 0.371)}$

Table 3.1 Estimated causal effects across different estimators for the two experimental setups described in Figure 3.1.

3.5.2 Multi-source experiment

For the multi-source experiment, we setup the super-distribution $\mathscr{P}$ as below

$$\begin{aligned} &\mathbb{P}_i \sim (1-A)\mathcal{N}(0, u_1^2) + A\left\{w\mathcal{N}((1-w)u_2, u_3^2) + (1-w)\mathcal{N}(-wu_2, u_4^2)\right\}, \quad \forall i \\ &u_1 \sim \mathscr{U}(0.5, 1.5)) \qquad u_2 \sim \mathscr{U}(1, 5) \qquad u_3 \sim \mathscr{U}(0.5, 1.5) \\ &u_4 \sim \mathscr{U}(0.5, 1.5) \qquad w \sim \mathscr{U}(0.25, 0.75) \qquad A \sim Bernoulli(0.5) \end{aligned}$$

, where we set $N = 50, n = 100$. Under this setting, for each $\mathbb{P}_i \sim \mathscr{P}$ we have

$$\mathbb{P}_i(Y^A) \Rightarrow \begin{cases} \text{unimodal}, & \text{if } A = 0 \\ \text{bimodal}, & \text{if } A = 1. \end{cases}$$

Note that for each i a pair $\mathbb{P}_i(Y^1), \mathbb{P}_i(Y^0)$ looks like the second example in Figure 3.1 and has the same mean by design. Having all the other conditions be the same with the previous single-source experiment setup, we estimate two baselines $\hat{\psi_{\text{diff}}}$, $\hat{\psi_{\text{HT}}}$, and our proposed estimator (3.12) as before, and present the results with 95% confidence intervals in Table 3.2. As shown in Table 3.2, the proposed estimator suggests there has been a significant shift in counterfactual distribution caused by the given treatment.

Estimator	Point Estimation	95% CI
Difference-in-means	0.039	$(-0.361, 0.435)$
Horvitz-Thompson	0.037	$(-0.368, 0.432)$
Difference-in-distributions	**0.194**	$\mathbf{(0.105, 0.284)}$

Table 3.2 Estimated causal effects across different estimators for the multi-source experiment

3.5.3 Real-world Example: Effect of Free Lunch on Achievement Gap

In this subsection, we illustrate the use of the proposed causal effects with an analysis of the effect of free lunch on achievement gap. Disparity in academic achievement across races is a severe social problem in the US. For example, the achievement gap between white and black students has narrowed very little over the last 50 years, despite supposed progress in race relations and increased emphasis on closing such discrepancies [47].

On the other hand, many public schools in the US provide free lunch for qualifying students, with the aim of equalizing performance based on the clear relationship between students' learning and overall nutritional status [151]. Surprisingly, the debate over free lunch programs involved little discussion about its impact on academic achievement, for example as to whether providing the free meal plans at school could improve the educational achievement gaps between different races. Here we attempt to investigate the causal effect of offering more free lunches upon the improvement on the achievement gap between different ethnicities.

We use datasets from the Stanford Education Data Archive (SEDA) [4] in which we collect the test score gaps between ethnicities, percent free lunch in average and other socioeconomic, and demographic characteristics of geographical school districts during 2009-2013 on a district basis. We consider a school district treated if it is providing above-average school level free lunch to ethnic minorities. Our outcome is Math and ELA test score gaps between White and two ethnic minorities, Black and Hispanic. Detailed information about dataset can be found in Table B.1 and B.2 in section B.1 of the appendix.

[4] https://cepa.stanford.edu/seda/data-archive

We first estimate the causal effect of free lunch on test gaps each year by employing three baseline methods that are widely used for observational studies in causal inference literature. Given the data structure (X, A, Y) we provide a description of each baseline estimator as below.

- *The plug-in regression estimator* $\hat{\psi}_{\text{pi}}$

  $$\hat{\phi_{\text{pi}}} = \mathbb{P}_n[\hat{\mu}_1(X) - \hat{\mu}_0(X)]$$

  where μ is regression function $\mathbb{E}[Y \mid A = a, X]$ to be estimated.

- *The inverse probability weighting estimator* $\hat{\psi_{\text{IPW}}}$

  $$\hat{\phi}_{\text{IPW}} = \mathbb{P}_n \left[\frac{AY}{\hat{\pi}(X)} - (\frac{1-A}{1-\hat{\pi}(X)})Y \right]$$

  where $\pi(X) = \mathbb{P}(A = 1|X)$.

- *The doubly-robust (semi-parametric) estimator* $\hat{\psi}_{\text{DR}}$

  $$\hat{\phi}_{\text{DR}} = \hat{\phi}_{\text{pi}} + \mathbb{P}_n \left\{ \left(\frac{A}{\hat{\pi}(X)} - \frac{1-A}{1-\hat{\pi}(X)} \right) [Y - \hat{\mu}_A(X)] \right\}$$

  .

Acknowledging that it would cause slow convergence rates for $\hat{\psi}_{\text{pi}}$ and $\hat{\psi_{\text{IPW}}}$, here we employ a nonparametric model to estimate both π, μ; we use Random Forests via `ranger` package in R. More details about these estimators (e.g. asymptotic properties) can be found, for example, in [94].

For each year, we estimate $\hat{\psi}_{\text{pi}}$, $\hat{\psi_{\text{IPW}}}$, $\hat{\psi}_{\text{DR}}$, and our proposed estimator (3.16), and present the results with 95% confidence intervals. For the sake of brevity, only the results for year 2009 are presented in Table 3.3 [5]. Most of the baseline methods appear to be not significant in that their 95% confidence interval contains zero. On the other hand, the proposed estimator suggests a substantial shift in counterfactual distribution induced by the free lunch program for both the White vs Black and White vs Hispanic. One may further look into and decide whether such distributional change is meaningful from the the policy-making perspective through extra tests if necessary.

[5] Since the results for other years are more or less similar, we move them to Table B.3, B.4, B.5 in section B.1 of the appendix.

	White-Black		White-Hispanic	
Estimator	Math	ELA	Math	ELA
Plug-in regression	-0.030 $(-0.086, 0.026)$	0.025 $(-0.085, 0.036)$	-0.029 $(-0.057, 0.028)$	-0.024 $(-0.084, 0.036)$
IPW	-0.020 $(-0.059, 0.019)$	-0.012 $(-0.056, 0.032)$	-0.022 $(-0.059, 0.015)$	-0.002 $(-0.029, 0.028)$
Doubly Robust	-0.056 $(-0.070, -0.042)$	-0.035 $(-0.048, -0.024)$	-0.049 $(-0.062, -0.037)$	-0.013 $(-0.026, 0.001)$
Difference-in-distributions	**0.752** (0.724,0.780)	**0.638** (0.596,0.680)	**0.702** (0.650,0.754)	**0.529** (0.480,0.579)

Table 3.3 Estimated causal effect of free lunch on test gaps in 2009 (with 95% CI)

3.6 Discussion

As illustrated in the introduction, there are often times when mere comparison of average effects reveals potentially less valuable information about how treatment works on outcomes. In this study, we pursue a more nuanced way to explore causal effects beyond the ATE by considering estimating causal effects based on the discrepancy between unobserved counterfactual distributions.

We provide a novel way to estimate each of the counterfactual outcome distributions for efficient estimation of our target functional $D_1(Q^0, Q^1)$ with the non-smooth L_1 distance by considering single- and multi-source randomized studies, as well as observational studies. We analyzed error bounds and asymptotic properties of the proposed estimators. To the best of our knowledge, our doubly robust style estimator for an observational study is the first result on efficient nonparametric counterfactual density estimation. We further propose methods to construct confidence intervals for the unknown mean distribution distance by analyzing the asymptotic convergence of our counterfactual density estimators.

Our proposed method can be always used jointly with the ATE, as a first step in assessing whether there is effect modification beyond a mean shift; for instance, when the ATE is nearly zero but our estimator is large, we should be cautious before making a decision merely based on the former. On top of that, one may build upon our proposed framework to meet their own analytical goals. For example, we conjecture that when used together with the ATE our method may provide an alternative indirect approach to test the degree of heterogeneity in treatment effects, in the sense that a large value of $D_1(Q_1, Q_0)$ with the nearly zero ATE implies considerable variation in subgroup effect in terms of magnitude or direction and thereby can be used as an evidence of heterogeneous treatment effects under certain circumstances (see the second example used in Section 3.5.1) [6]. Furthermore, even when we are interested in other types of functional such as quantiles and cdfs between different counterfactual distributions, we can use our results to construct and test hypotheses with respect to these quantities. It is also worth noting that our method can be extended to the weighted distributional distances D_w defined by $D_w = \int w(u)|p(u) - q(u)|du$. In the ordinary L_1 distance we have $w = 1$ everywhere, but one can appropriately tailor the weight function w if necessary; one example can be a sigmoid function, which would be useful when we care more about positive effects.

Our work leads to many opportunities for important future work from theoretical perspective as well. We plan several extensions of our work from L_1 distance to more general functionals. Considering varying bandwidth in our counterfactual density estimator would be another important future extension. Moreover, as mentioned in Remark 1, we conjecture that our proposed estimators may be minimax optimal when the bandwidth is tuned in a particular way, but we leave that to future work.

[6] This must be accompanied with other statistical metrics that would help us to properly define the degree of heterogeneity in treatment effects.

Chapter 4

Causal Clustering

4.1 Introduction

Statistical causal inference is about estimating what would happen to some response when a “cause” of interest is changed or intervened upon, possibly contrary to an observed fact. This is essential for answering many important questions in health, public policy, economics, and across science: e.g., how would survival change under medical treatment A vs. B, or what are the economic effects of policy X vs. Y? To mathematically frame such causal problems, we use *counterfactual* or *potential* outcomes [118]. We consider a setup where we observe n iid samples of $Z = (X, A, Y) \sim \mathbb{P}$, where $X \in \mathbb{R}^d$ are covariates, $A \in \{0, 1, ..., p-1\}$ denotes treatment, and $Y \in \mathbb{R}$ is an outcome of interest. The potential or counterfactual outcome we would have observed for a unit had they received treatment $A = a$ is denoted Y^a for $a \in \mathscr{A}$. To compare population-average outcomes between two treatment levels (e.g., $A = 1$ versus $A = 0$), we can formulate the population-level *average treatment effect* (ATE) as

$$\mathbb{E}(Y^1 - Y^0). \tag{4.1}$$

The ATE is one of the most popular target effects in causal inference, and can be identified and estimated under a proper set of assumptions in both randomized and observational studies [e.g., 52, 56]. There have been much work concerning efficient estimation of the ATE, and its analogs in more complex data structures such as censored longitudinal data [137]. Recently there has been also a huge interest in incorporating the benefits of machine learning into estimating such causal parameters [e.g., 138, 19, 109].

4.1.1 Heterogeneity in Treatment Effects

A potential shortcoming of the ATE in (4.1) is that it can mask heterogeneity in causal effects, e.g., across subgroups of different units. Identifying such treatment effect heterogeneity and corresponding subgroups is of great importance in policy evaluation, drug development, and health care service, and has generated growing recent interest. For example, patients with different subtypes of cancer often react differently to the same treatment; however, our understanding of cancer subtypes at the molecular level is limited [50], and there is little consensus about which treatments are most effective for which patients [79]. Typically, the functional form of the relationship between treatment effects and the attributes of units is not known a priori and thus such effect heterogeneity has to be explored via data driven methods. There has been a lot of recent work in this area [5, 145, 153, 29, 55, 39, 42, 43, 147, 122], but there are many open problems and it has not been studied as extensively as other branches of causal inference [70].

Most approaches for studying effect heterogeneity target the conditional average treatment effect (CATE), defined as follows.

Definition 4.1.1 (Conditional average treatment effect (CATE))**.**

$$\tau(X) = \mathbb{E}[Y^1 - Y^0 \mid X]. \tag{4.2}$$

The CATE captures how treatment effects vary with covariates. Various methods have been proposed to obtain estimates of and inferences for the CATE, with a special emphasis in recent years on incorporating flexible machine learning tools. For example, van der Laan and Luedtke [134] provided a framework of efficient CATE estimation based on the loss-based super-learning approach. Athey and Imbens [5] developed a popular tree-based method. Imai et al. [55] and Wager and Athey [145] adapted random forest and support vector machine classifiers. Shalit et al. [122] presented error bounds using domain adaptation. Künzel et al. [82] proposed meta-algorithms for CATE estimation, with a particular focus on unbalanced designs. Nie and Wager [99] gave a novel adaptation of RKHS regression methods and studied conditions for oracle efficiency. Kennedy [70] gave generic model-free error bounds and pursued fastest possible convergence rates.

4.1.2 Motivation

In contrast to previous work, which pursues methods with a definite supervised learning flavor, we instead consider assessing effect heterogeneity via an unsupervised learning perspective. Namely, rather than estimating the CATE specifically, we aim to infer the properties and structure of effect heterogeneity by finding underlying subgroups and clusters. Our work is therefore more descriptive and discovery-based, which we feel fills a gap in the literature. This is exactly analogous to the clustering versus regression distinction in standard statistical learning [25, Theorem 2.2]; to the best of our knowledge, clustering methods have yet to be exploited in causal inference, let alone in the heterogeneous effects problem.

Thus in this paper we propose adapting unsupervised learning methods for understanding treatment effects: we develop *Causal Clustering*, a new approach for analyzing effect heterogeneity that leveraging tools from clustering analysis. Specifically, we pursue an efficient way to uncover subgroup structure in conditional treatment effects by harnessing widely-used clustering methods. Relative to standard CATE estimators, our framework provides complementary tools for ascertaining subgroups with similar treatment effects, exploiting flexible unsupervised machine learning methods. Importantly, causal clustering can be particularly useful in outcome-wide studies with multiple treatment levels [143, 144], where rather than probing a high-dimensional CATE surface to assess structure one can instead find lower-dimensional clusters with similar treatment effects.

4.1.3 Clustering with Unknown Outcomes

Our problem largely differs from the standard clustering setup since, as we indicate in Section 4.2 in more detail, the variable to be clustered consists of unknown regression functions that need to be estimated. Clustering with this kind of unknown "pseudo-outcome" has not been studied as extensively as standard clustering that is performed on deterministic, fully observed data. Some recent work has considered cluster analysis with partially observed data, as in for example Serafini et al. [121] who explored missing data problems in clustering and Haviland et al. [49] who studied group-based trajectory modeling with non-random dropout. They use parametric approach to model partially unobserved outcomes which are in a vector form in fixed dimensions, unlike fully unobserved regression functions in our paper. Su et al. [123] has considered clustering with measurement errors, by modeling marginal outcome distribution through deconvoluting density estimation, but still on outcomes in a vector form. Kumar and Patel [80] considered clustering on unknown model parameters, without theoretical arguments, relying on parametric assumption. Importantly, as far as we are aware, none of the existing methods in clustering literature have explored general nonparametric approaches to clustering on unknown pseudo-outcomes.

4.1.4 Paper Organization

The remainder of the paper is structured as follows. In Section 2, we present the idea of causal clustering and associated assumptions and notation. In Section 3, we show that k-means, density-based, and hierarchical clustering algorithms can be successfully adopted into our framework with simple plug-in estimators, albeit with a cost in error rates coming through a first-order nuisance error. In Section 4, we develop a more efficient bias-correced estimator for k-means causal clustering using nonparametric efficiency theory, which attains fast convergence rates to the true cluster centers under weak nonparametric conditions on nuisance estimators. There, we also give conditions for asymptotic normality of the cluster centers. In Section 5, we argue why our framework can be easily generalized to outcome-wide studies. Section 6 provides simulation studies as well as case studies with real data. Section 7 concludes with a discussion.

4.2 Setup & Notation

As in the previous section, we consider an observational study consisting of n iid samples of $Z = (X, A, Y) \sim \mathbb{P}$, where we let $\mathscr{X} \in \mathbb{R}^d$, $\mathscr{A} = \{1, ..., p\}$, and $\mathscr{Y} \in \mathbb{R}$ denote the support of our pre-treatment covariate (X), treatment (A), and outcome (Y) variables respectively. Note that we allow multi-level treatments, i.e. p distinct levels of treatment with $p \geq 2$ where the index starts from 1. For conditional effects like the CATE in (4.2) to be identified, we require the following standard causal assumptions.

Assumption A1. *(consistency)* $Y = Y^a$ *if* $A = a$.

Assumption A2. *(no unmeasured confounding)* $A \perp\!\!\!\perp Y^a \mid X$.

Assumption A3. *(positivity)* $\mathbb{P}(A = a \mid X)$ *is bounded away from 0 a.s.* $[\mathbb{P}]$.

Assumptions (A1)-(A3) are the standard assumptions commonly adopted in the causal inference literature [52]. Assumption (A1) means that observed outcomes must equal corresponding potential outcomes under the observed treatment sequence; it could be violated for example in the presence of interference. Assumption (A2) is sometimes called *(conditional) randomization* or *exchangeability* and holds by design in a randomized experiment. However, in an observational study it requires sufficiently many relevant confounders to be collected. Assumption (A3) implies that everyone must have some positive probability of receiving each treatment level. This is needed since otherwise some counterfactuals would never be observed even in an infinite superpopulation.

Under Assumptions (A1)-(A3), it is well-known that the counterfactual regression function under $a \in \mathscr{A}$ is identified as

$$\mu_a(X) \equiv \mathbb{E}[Y^a \mid X] = \mathbb{E}[Y \mid X, A = a]$$

Therefore, for $\forall a, a' \in \mathscr{A}$ a pairwise CATE of the treatment a relative to a' is given by

$$\begin{aligned} \tau_{aa'}(X) &\equiv \mathbb{E}[Y \mid X, A = a] - \mathbb{E}[Y \mid X, A = a'] \\ &= \mu_a(X) - \mu_{a'}(X) \end{aligned} \tag{4.3}$$

Next we define the *conditional counterfactual mean vector* that maps the covariates into the p-dimensional mean-outcome space.

Definition 4.2.1 (Conditional counterfactual mean vector)**.** *We define* $\boldsymbol{\mu}$ *by*

$$\boldsymbol{\mu}(X) = \left[\mathbb{E}[Y^1 \mid X], \ldots, \mathbb{E}[Y^p \mid X]\right]^\top. \tag{4.4}$$

In other words, $\boldsymbol{\mu}$ is a surjection of covariate space $\mathscr{X}$ onto $\mathbb{R}^p$. Under Assumptions (A1)-(A3), $\boldsymbol{\mu}$ could be estimated by estimating each regression function μ_a with observed data.

Each point projected through the conditional counterfactual mean vector carries implicit information about the CATE. If all coordinates of a point $\boldsymbol{\mu}(X)$ were the same, this would mean no treatments had any effects on the conditional mean scale. Furthermore for two units i, j, we have

$$\boldsymbol{\mu}(X_i) \approx \boldsymbol{\mu}(X_j) \Rightarrow \tau_{aa'}(X_i) \approx \tau_{aa'}(X_j) \quad \text{for all } \ a, a' \in \mathscr{A}.$$

Namely, adjacent units in the conditional counterfactual mean vector space would show similar reactions toward a given set of treatments in terms of the CATE. This gives some motivation for uncovering subgroup structure via cluster analysis on the image of the conditional counterfactual mean vector.

In Figure 4.1, we illustrate the idea of causal clustering through the case of binary treatments ($p = 2$). We generate 900 samples in the conditional counterfactual mean vector space using a mixture of six Gaussian distributions with different means and covariance functions, where the overall ATE is set to be exactly zero. By construction, there are roughly six clusters where units within each cluster are more homogeneous in terms of the CATE. When it comes to analyzing the heterogeneity of treatment effect, often people rely on histogram of the CATE as in Figure 4.1-(c). However in this illustration drawing histogram

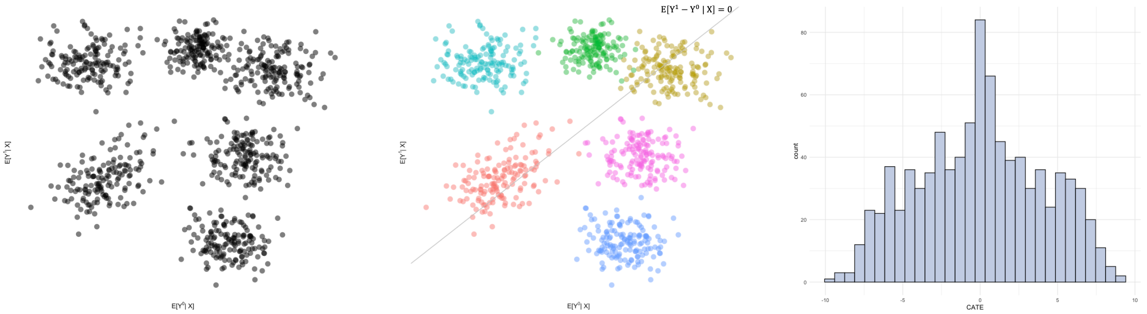


Fig. 4.1 Illustration for causal clustering for the case of binary treatments with a mixture of six Gaussian distributions, where $\boldsymbol{\mu} \in \mathbb{R}^2$ and $\mathbb{E}[Y^1 - Y^0] = 0$. The symmetrical, bell-shaped histogram of CATE in (c) is not informative on the underlying patterns of the mixture distributions shown in (b), which could be exploited via clustering analysis.

still does not give the whole story. On the other hand, cluster analysis with $\boldsymbol{\mu}$'s can effectively discover a subgroup structure that would be obscure when just looking at the histogram of CATE.

Although the aforementioned clustering idea is simple and intuitive, standard results from the clustering literature cannot be applied off-the-shelf, since the variable to be clustered is $\boldsymbol{\mu}$, which consists of unknown outcome regression functions that need to be estimated. In contrast, standard clustering is performed on observed data, not unknown pseudo-outcomes like $\boldsymbol{\mu}$. Consequently, it is unclear which if any theoretical guarantees of the original clustering algorithms hold for causal clustering, and under what conditions.

For our analysis, we simply write $\boldsymbol{\mu} \equiv \boldsymbol{\mu}(X) = [\mu_1(X), ..., \mu_p(X)]^\top$ and $\widehat{\boldsymbol{\mu}} \equiv \widehat{\boldsymbol{\mu}}(X) = [\widehat{\mu}_1(X), ..., \widehat{\mu}_p(X)]^\top$ when the dependency on X is clear from the context, and consider sets of points $\mathsf{U}^n = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_n\}$ and $\mathsf{O}^n = \{\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2, ..., \widehat{\boldsymbol{\mu}}_n\}$ induced from our data $\mathsf{D}^n = \{Z_1, Z_2, ..., Z_n\}$ where each $Z_i \overset{\text{i.i.d}}{\sim} \mathbb{P}$.

Hereafter, we let $\|x\|_p$ denote L_p norm for any fixed vector x. When we are given a fixed operator f, we let $\mathbb{P}_n$ and $\mathbb{P}$ denote the empirical measure over D^n and the conditional expectation over $\mathbb{P}$, respectively, as in $\mathbb{P}_n(f) = \mathbb{P}_n\{(f(Z)\} = \frac{1}{n}\sum_{i=1}^n f(Z_i)$ and $\mathbb{P}(f) = \int f(z)d\mathbb{P}(z)$. We also use $\|f\|_{\mathbb{P},p}$ to denote the $L_p(\mathbb{P})$ norm defined by $\|f\|_{\mathbb{P},p} = [\mathbb{P}(f^p)]^{1/p} = [\int f(z)^p d\mathbb{P}(z)]^{1/p}$. In particular, we use $\|\cdot\|$ as a shorthand notation for $L_2(\mathbb{P})$ norm as $L_2(\mathbb{P})$ is used most frequently in this paper. Moreover throughout the development, for $x \in \mathbb{R}^d$ and $r > 0$, we let $\mathbb{B}(x,r)$ denote the open ball centered at x with radius r with respect to L_2 norm, i.e. $\mathbb{B}(x,r) = \{y \in \mathbb{R}^d : \|x-y\|_2 < r\}$ and use the notation $\overline{\mathbb{B}(x,r)}$ for the closed ball.

Lastly, we impose the following mild boundedness assumption to ensure that our cluster analysis is performed on the compact space.

Assumption A4. *$\|Y\|_\infty < \infty$ and $\|\boldsymbol{\mu}\|_2, \|\widehat{\boldsymbol{\mu}}\|_2 \leq B$ for some finite constant B.*

Remark 3. *One may flexibly tailor the conditional counterfactual mean vector in* (4.2.1) *to fit a specific purpose. We give two examples here. For the sake of simplicity, let us assume $\mathscr{A} = \{0,1\}$. 1) Suppose that we only care about the magnitude of the CATE. Then we can simply redefine $\boldsymbol{\mu} = \mu_1 - \mu_0$* [1] *and perform clustering analysis with this new $\boldsymbol{\mu}$. 2) Next suppose, for example, that we are interested in how a treatment shifts the median of an outcome variable as in the context of the quantile treatment effects [e.g. 21, 103, 154]. In this case, we can redefine our conditional counterfactual mean vector by $\boldsymbol{\mu} = (Q_0(q), Q_1(q))$ for some prespecified $q \in (0,1)$ (for median, $q = 1/2$), where $Q_a(q)$ is the quantile function of our potential outcome Y^a, i.e. $Q_a(q) = \inf\{y \in \mathbb{R} : q \leq F_{Y^a}(y)\}$ where $F_{Y^a} = \mathbb{P}(Y^a \leq y \mid X)$.*

Remark 4. *A growing number of recent studies seek to adopt outcome-wide approaches where our outcome variable is essentially multivariate [143, 88, 144]. Suppose that we assess causal effects over m different outcomes. We let $Y^a_{(l)}$ denote a potential outcome for the l-th outcome under treatment $a \in \mathscr{A}$ and $\mu_a^{(l)} \equiv \mathbb{E}[Y^a_{(l)} \mid X]$. Our framework is easily extendable to outcome-wide studies by simply letting, for example, $\boldsymbol{\mu} = (\mu_0^{(1)}, ..., \mu_p^{(1)}, \mu_0^{(2)}, ..., \mu_p^{(2)}, ..., \mu_0^{(m)}, ..., \mu_p^{(m)})$.*

4.3 Analysis on Three Causal Clustering Algorithms

In this section, we analyze three causal clustering algorithms. Specifically, we provide error analysis of plug-in estimators for k-means, density-based, and hierarchical clustering algorithms, and show that they can be successfully adopted into our framework at the cost of first-order nuisance error rates.

4.3.1 k-means Clustering

Originally from signal processing, *k*-means (a.k.a vector quantization) is the one of the oldest, and the most popular approaches to clustering. It works by finding *k* representative points which defines a Voronoi tessellation. There has been a substantial amount of research on *k*-means clustering (see, for review, [57] or the monograph of [41]). It is one of the few clustering algorithms whose theoretical properties are relatively well-understood, since the analysis is relatable to principal components analysis [28, 152].

[1] One may instead use a proper dissimilarity measure to get the same result.

We call a set of those k representative points a codebook $C = \{c_1,...,c_k\}$ where $c_j \in \mathbb{R}^p, j = 1,...,k$. Let $\mathit{\Pi}_C[x]$ be the projection of $x \in \mathbb{R}^p$ onto C:

$$\mathit{\Pi}_C[x] = \underset{c \in C}{\operatorname{argmin}} \|c - x\|_2^2.$$

Then define the *population clustering risk* $R(C)$ and *empirical clustering risk* $R_n(C)$ by

$$R(C) = \mathbb{E}\|\boldsymbol{\mu} - \mathit{\Pi}_C[\boldsymbol{\mu}]\|_2^2, \qquad R_n(C) = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{\mu}_i - \mathit{\Pi}_C[\boldsymbol{\mu}_i]\|_2^2$$

One may construct an ordinary k-means clustering scheme by computing the optimal codebook $\widehat{C}^*$ that minimizes $R_n(C)$ as an estimate of the optimal population codebook C^*. That is,

$$\widehat{C}^* = \underset{C \in \mathscr{C}_k}{\operatorname{argmin}} R_n(C), \qquad C^* = \underset{C \in \mathscr{C}_k}{\operatorname{argmin}} R(C)$$

where $\mathscr{C}_k$ denotes all codebooks of length k in the image of $\boldsymbol{\mu}$. The common way to find such $\widehat{C}^*$ is known as Lloyd's algorithm [91, 62] but there are other recent developments as well [84]. A solution of such algorithms normally depends on the starting values. Some popular methods for choosing good starting values are discussed in Arthur and Vassilvitskii [4], Tseng and Wong [127].

The problem of evaluating how good $\widehat{C}^*$ is, compared to the truly optimal C^*, has been extensively studied particularly in perspective of the excess risk analysis. Pollard [101] proved that k-means is risk consistent in the sense that $R(\widehat{C}^*) - R(C^*) \xrightarrow{a.s.} 0$. Borrowing techniques from statistical learning theory, the standard result by Linder et al. [89] states that when an input vector is almost surely bounded we achieve $\mathbb{E}\left[R(\widehat{C}) - R(C^*)\right] = O\left(\sqrt{\frac{\log n}{n}}\right)$. The lower bound is found by Bartlett et al. [8] as $O(1/\sqrt{n})$ which is later achieved by [10]. However, it has been shown that faster rates of $O(\log n/n)$, $O(1/n)$ can be achieved under certain conditions as well [e.g., see Section 1 of 87].

All the previous studies including the mentioned above assume fixed, deterministic training samples. However in our setting we cannot compute $\widehat{C}^*$ as in ordinary k-means since we do not observe U^n. Instead for k-means causal clustering, we propose the following plug-in estimator to compute the optimal codebook $\widehat{C}$ from O^n by

$$\begin{aligned} \widehat{C} &= \underset{C \in \mathscr{C}_k}{\operatorname{argmin}} \widehat{R}_n(C), \\ \text{where} \quad \widehat{R}_n(C) &= \frac{1}{n}\sum_{i=1}^{n} \|\widehat{\boldsymbol{\mu}}_i - \mathit{\Pi}_C[\widehat{\boldsymbol{\mu}}_i]\|_2^2. \end{aligned} \tag{4.5}$$

We aim to verify that under which conditions $\widehat{C}$ is still risk consistent and compute the convergence rate. By borrowing similar techniques used in [89], in what follows we provide an error bound of L_1-risk for our k-means causal clustering.

Theorem 4.3.1. *Suppose we are given O^n where $\widehat{\boldsymbol{\mu}}$ is estimated in the separate sample set $\mathsf{D}_0^n = \{Z_{n+1}, ..., Z_{2n}\}$. Then under assumptions (A1)-(A4), there exists an integer n_0 such that for every $n > n_0$*

$$\mathbb{E}\left|R(\widehat{C}) - R(C^*)\right| \leq 32B^2\sqrt{\frac{k(p+1)\log n}{n}} + 4\sqrt{2}B\sum_a \|\widehat{\mu}_a - \mu_a\|_1 .$$

A proof of above theorem is given in Section C.2.1 of the appendix. Note that in Theorem 4.3.1 we use sample splitting to avoid imposing any extra conditions on the function class of μ_a. The first term of the error bound in Theorem 4.3.1 is the same order as the rates given in [89]. Therefore, Theorem 4.3.1 implies that the extra price that we pay as regards the excess risk is the estimation error of outcome regression functions.

The fact that $\widehat{C}$ is risk consistent does not always imply that $\widehat{C}$ is actually close to the true codebook C^*. The classical result of Pollard et al. [102] addressed this issue by finding conditions to assure asymptotic normality of $\widehat{C^*}$ to C^* by assuming a unique optimal codebook. On the other hand, in our case even in the context where there is a unique optimal codebook, quite different configurations of centers C may give rise to very similar values of the excess risk $R(C) - R(C^*)$ due to discrepancy between U^n and O^n. We will save our discussion of this topic until Section 4.4.

4.3.2 Hierarchical Clustering

Hierarchical clustering methods build a set of nested clusters at different resolutions, and the resulting hierarchy is usually depicted by a binary tree or dendrogram. Hence, they require no prior specification of the number of clusters and they permit the data to be understood simultaneously at many levels of granularity based on the predefined similarity measure. We will be considering algorithms whose only access to their data is via a pairwise similarity function $d : \mathbb{R}^p \times \mathbb{R}^p \to [-1, 1]$. There are two types of hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). Here we only consider agglomerative approach which is more common in practice [148].

Many agglomerative clustering algorithms extend d so that we can compute the distance, or *linkage*, $D(A, B)$ between sets of points A and B to form the cluster hierarchy [e.g., 58, 23, 31]. The most common ways of extending the distance are to use *single, average*, or *complete* linkages. The following lemma provides an error bound of computing the set distance in O^n.

Lemma 4.3.1. *Let D denote the single linkage between sets of points. Then for any two sets A, B in U^N and the corresponding estimates $\widehat{A}, \widehat{B}$ in O^n, we have*

$$\left| D(A,B) - D(\widehat{A},\widehat{B}) \right| \leq \sqrt{2} \sum_{a \in \mathscr{A}} \|\widehat{\mu}_a - \mu_a\|_\infty .$$

The same result holds for average and complete linkages.

The proof of the lemma is given in Section C.2.2 of the appendix. Unlike k-means clustering it is not straightforward to analyze the performance of hierarchical clustering with respect to the true target hierarchy which is an infinite set of clusters across different resolutions. More importantly, in the presence of noise the standard linkage-based algorithms might fail.

Balcan et al. [7] proposed a new robust agglomerative hierarchical clustering algorithm that can handle above issues. Their algorithm produces clustering that contains a pruning which is close to the target clustering at a prespecified error rate in the presence of noise, and can be implemented even under the inductive setting where we use only small subset of entire sample. Suppose we have N samples in total. We consider a subset S of size n, $n \ll N$, and a clustering problem (S, l) in the conditional counterfactual mean vector space where each point $\boldsymbol{\mu} \in S$ has a true cluster label $l(\boldsymbol{\mu}) \in \{C_1, ..., C_k\}$. Further we let $C(\boldsymbol{\mu})$ denote a cluster corresponding to the label $l(\boldsymbol{\mu})$, and $n_{C(\boldsymbol{\mu})}$ denote the size of the cluster $C(\boldsymbol{\mu})$. To proceed we define the following *good-neighborhood property* to quantify the level of noisiness in our population distribution.

Definition 4.3.1 ((α, ν)-good neighborhood property for distribution)**.** *For $\boldsymbol{\mu}' \in \mathsf{U}^N$, let $\mathbb{C}(\boldsymbol{\mu}') = \{\boldsymbol{\mu} : C(\boldsymbol{\mu}) = C(\boldsymbol{\mu}')\}$, i.e. a set whose label is equal to $C(\boldsymbol{\mu}')$, and $r_{\boldsymbol{\mu}'} = \inf_r \{r : \mathbb{P}[\boldsymbol{\mu} \in \mathbb{B}(\boldsymbol{\mu}', r)] = \mathbb{P}[\mathbb{C}(\boldsymbol{\mu}')]\}$. The distribution $\mathbb{P}_{\alpha,\nu}$ satisfies (α, ν)-good neighborhood property if $\mathbb{P}_{\alpha,\nu} = (1-\nu)\mathbb{P}_\alpha + \nu \mathbb{P}_{noise}$ where $\mathbb{P}_\alpha$ is a probability distribution which satisfies*

$$\mathbb{P}\{\boldsymbol{\mu} \in \mathbb{B}(\boldsymbol{\mu}', r_{\boldsymbol{\mu}'}) \setminus \mathbb{C}(\boldsymbol{\mu}')\} \leq \alpha$$

for any $\boldsymbol{\mu}' \in \mathsf{U}^N$, and $\mathbb{P}_{noise}$ is any valid distribution.

The good-neighborhood property in Definition 4.3.1 is distributional extension of the original good neighborhood property proposed by Balcan et al. [6, 7]. Next, we assume the following mild boundedness condition on our population density.

Assumption A5. *$\mathbb{P}_{\alpha,\nu}$ in Definition 4.3.1 has a bounded Lebesgue density.*

In the next theorem, we specify the cost of performing causal clustering via the robust agglomerative hierarchical clustering.

Theorem 4.3.2. *Suppose that* U^N *consists of* N *i.i.d samples from* $\mathbb{P}_{\alpha,\nu}$ *that satisfies the* (α,ν)*-good neighborhood property in Definition 4.3.1. Also assume that* $\widehat{\boldsymbol{\mu}}$ *is estimated in the separate sample set* $\mathsf{D}_0^n = \{Z_{N+1},...,Z_{N+n}\}$*. Consider a random subset* $\mathsf{U}^n \subset \mathsf{U}^N$ *and corresponding* O^n *in which clustering to be performed. Now let* $\gamma = \sum_{a\in\mathscr{A}} \|\widehat{\mu}_a - \mu_a\|_\infty$*, and for any* $\delta_N, \delta_n \in (0,1)$*, define*

$$\alpha' = \alpha + O\left(\sqrt{\frac{1}{N}\log\frac{1}{\delta_N}}\right), \quad \nu' = \nu + O\left(\sqrt{\frac{1}{N}\log\frac{1}{\delta_N}}\right), \quad \beta = O\left(\gamma + \frac{1}{n}\log(\frac{1}{\delta_n})\right).$$

Then as long as the smallest target cluster has size greater than $12(\nu'+\alpha'+\beta)N$*, the robust hierarchical clustering [7, Algorithm 2] in* O^n *with* $n = \Theta\left(\frac{1}{\min(\alpha'+\beta,\nu')}\ln\frac{1}{\delta\min(\alpha'+\beta,\nu')}\right)$ *produces a hierarchy with a pruning that have error at most* $\nu'+\delta$ *with respect to the true target clustering with probability at least* $1-\delta-\delta_N-\delta_n$*.*

The proof is given in Section C.2.2 in the appendix. The above theorem assumes the inductive setting where we use only subset of size n from the entire sample set of size N. The main implication of Theorem 4.3.2 is that roughly speaking, the natural misclassification error α has increased by $O(\sum_{a\in\mathscr{A}} \|\widehat{\mu}_a - \mu_a\|_\infty) + o(1)$ due to the cost of causal clustering [2].

4.3.3 Density-Based Clustering

The idea of density-based clustering was first introduced as an efficient clustering algorithm for large-scale, noisy datasets [33, 53]. It works by detecting areas where points are concentrated and where they are relatively sparse or empty. The density-based methods provide advantages over other clustering methods through their noise handling capabilities and ability to determine non-spherical shaped clusters. Here, we focus on the level-set approach (Hartigan [48]; **?**] and the references therein.).

To avoid confusion with the notation on probabilistic arguments, we slightly abuse the notation in this subsection; we set $|\mathscr{A}| = d$ so now $\boldsymbol{\mu} \in \mathbb{R}^d$. Further we let P be the probability distribution of $\boldsymbol{\mu}$ to distinguish it from $\mathbb{P}$, and p be the corresponding Lebesgue density. We also let K denote a valid kernel function, i.e. a nonnegative function with $\int K(u)du = 1$. We construct the oracle kernel density estimator $\widetilde{p}_h$ with bandwidth $h > 0$ as

$$\widetilde{p}_h(\boldsymbol{\mu}') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K\left(\frac{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}'\|}{h}\right),$$

[2] This can be clear by comparing the result of Theorem 4.3.2 with Theorem 11 of [7]

for $\forall \boldsymbol{\mu}' \in \mathbb{R}^d$. Then we define an average oracle kernel density estimator by $p_h \equiv \mathbb{E}(\widetilde{p}_h)$ and the corresponding upper level set by $L_{h,t} = \{\boldsymbol{\mu} : p_h(\boldsymbol{\mu}) > t\}$. Suppose that for each t, $L_{h,t}$ can be decomposed into finitely many disjoint sets: $L_{h,t} = C_1 \cup \cdots \cup C_{l_t}$. Then $\mathscr{C}_t = \{C_1, ..., C_{l_t}\}$ is the *level set clusters* of our interest at level t.

With regard to the analysis of topological properties of the distribution P, the upper level set of p_h serves a very similar role to the upper level set of the true density p, while offering several advantages [35, 76]. For example, p_h is always well-defined even when p is not, p_h provides simplified topological information, and the convergence rate of the kernel density estimator to p_h is faster than to p. For such reasons, we typically target the level set $L_{h,t}$ induced from p_h instead of the one induced from p.

When each $\boldsymbol{\mu}_i$ is observed, the level sets can be estimated by computing $\widetilde{L}_{h,t} = \{\boldsymbol{\mu} : \widetilde{p}_h(\boldsymbol{\mu}) > t\}$. Specifically, for each t we let $\widetilde{\mathscr{W}_t} = \{\boldsymbol{\mu} : \widetilde{p}_h(\boldsymbol{\mu}) > t\}$, and construct a graph G_t where each $\boldsymbol{\mu}_i \in \widetilde{\mathscr{W}_t}$ is a vertex and there is an edge between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ if and only if $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \leq h$. Then the clusters at level t are estimated by taking the connected components of the graph G_t which is called a *Rips* graph. *Persistent homology* measures how the topology of R_t varies by the value of t. See Kent et al. [75], Bobrowski et al. [13] for details in algorithm and its theoretical properties.

However in our case, the oracle kernel density estimator $\widetilde{p}_h$ is not computable since we do not observe each $\boldsymbol{\mu}_i$. Thus we construct a plug-in version of the kernel density estimator for $\widehat{\boldsymbol{\mu}}$ as

$$\widehat{p}_h(\boldsymbol{\mu}') = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{c_K h^d} K\left(\frac{\|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}'\|}{h}\right)$$

with a normalizing constant c_K, and target the corresponding level set $\widehat{L}_{h,t} = \{\boldsymbol{\mu} : \widehat{p}_h(\boldsymbol{\mu}) > t\}$.

In order to handle the additional complication in estimating $\widehat{L}_{h,t}$, we impose the same bounded-density assumption (A5) on the distribution P and introduce the following mild regularity conditions on the kernel K.

Assumption A6. *The kernel function K has a support on $\overline{\mathbb{B}(0,1)}$. Moreover, it is Lipschitz continuous with constant M_K, i.e. for all $x, y \in \mathbb{R}^d$, $|K(x) - K(y)| \leq M_K \|x - y\|_2$.*

In the next theorem, we claim that provided that the target level set $L_{h,t}$ is stable enough, i.e. it does not change too much when t perturbs, our level set estimator $\widehat{L}_{h,t}$ is close to the target level set $L_{h,t}$ in the Hausdorff distance H.

Theorem 4.3.3. *Suppose that $L_{h,t}$ is stable and let $H(\cdot,\cdot)$ be the Hausdorff distance between two sets. We further assume that $\widehat{\boldsymbol{\mu}}$ is estimated in the separate sample set $\mathsf{D}_0^n =$*

$\{Z_{n+1},...,Z_{2n}\}$. *Let the bandwidth vary with n such that* $\{h_n\}_{n\in\mathbb{N}} \subset (0,h_0)$ *and*

$$\limsup_n \frac{(\log(1/h_n))_+}{nh_n^d} < \infty.$$

Then, under the assumptions (A1)-(A6),

$$H(\widehat{L}_t, L_{h,t}) = O_P\left(\sqrt{\frac{(\log(1/h_n))_+}{nh_n^d}} + \frac{1}{h_n^{d+1}} \min\left\{\sum_a \|\widehat{\mu}_a - \mu_a\|_1, h_n\right\}\right)$$

See Appendix C.1.1 for the definition of the stability of the level set and the Hausdorff distance. The proof of Theorem 4.3.3 is given in the appendix C.2.3. The above theorem guarantees the estimated level sets are not drastically different from $L_{h,t}$. Hence we again verify that causal clustering also can be done via level-set density-based clustering at the additional cost of estimating the nuisance regression functions for the outcome process.

The three clustering algorithms analyzed in this section are developed based on different theories, and each has its own merits and risks. Therefore, which method to use should depend on data.

4.4 Efficient k-means Causal Clustering

All the estimators proposed in Section 4.3 are essentially nonparametric plug-in type, and we showed that their convergence rates is in general dominated by the estimation rate of the outcome regression function μ_a. Although they are easy to implement, the dependence on the estimation rate of μ_a would be problematic to attain $\sqrt{n}$ rates in nonparametric models, unless we assume unrealistic structural assumptions such as high-order smoothness. Furthermore, although the risk function characterizes an important feature of the clustering scheme in k-means clustering, characterization of convergence properties of the cluster centers themselves could convey more valuable information. This section is devoted to developing a more efficient estimator for k-means causal clustering based on the nonparametric efficiency theory and influence functions.

4.4.1 Setup

Consider the population clustering risk $R(C)$ given a codebook $C \in \mathscr{C}_k$ as in Section 4.3.1. In the sequel, the set of minimizers of the clustering risk will be denoted by $\mathscr{M}^*$, i.e.

$\mathscr{M}^* = \{C^* \in \mathscr{C}_k : R(C^*) = \min_{C \in \mathscr{C}_k} R(C)\}$. Then we consider the *kernel-smoothed* risk function

$$R_h(C) \equiv \mathbb{E}\|\boldsymbol{\mu} - \widetilde{\Pi}_C(\boldsymbol{\mu};h)\|_2^2 \tag{4.6}$$

, where the non-smooth projection function Π_C has been smoothed with kernel $\boldsymbol{K}$ and bandwidth $h > 0$ by

$$\begin{aligned} \widetilde{\Pi}_C(\boldsymbol{\mu};h) &= \sum_r \omega_r(\boldsymbol{\mu};C,h)c_r, \\ \text{where } \omega_r(\boldsymbol{\mu};C,h) &\equiv \frac{\boldsymbol{K}\left(\frac{\|\boldsymbol{\mu}-c_r\|_2}{h}\right)}{\sum_l \boldsymbol{K}\left(\frac{\|\boldsymbol{\mu}-c_l\|_2}{h}\right)}. \end{aligned} \tag{4.7}$$

For the brevity of proofs we use Gaussian (radial basis function) kernel: $\boldsymbol{K}(\boldsymbol{\mu}, c_r) = \exp(-\frac{\|\boldsymbol{\mu}-c_r\|_2}{h})$. However, we remark that other types of kernel also can be employed as long as they are bounded and sufficiently smooth [3]. Also to simplify the notation, we drop the dependency on W and C,h in the weight ω_r when it is clear from context.

First, for a given codebook C we aim to develop a doubly robust, efficient influence function based estimator for $R_h(C)$ so that it can eventually be an efficient estimator for the original risk $R(C)$ with proper choice of h. Next, we propose a minimizer of the estimator for $R_h(C)$ as our estimator for the optimal cluster codebook C^* and show that it is risk consistent at fast rates. Finally, we will argue that under proper conditions our proposed estimator is consistent and asymptotically normal to the true optimal codebook.

In our development, we will show that utilizing information on treatment process gives better efficiency. Hereafter, we define $\pi_a(X) \equiv \mathbb{P}[A = a \mid X]$, a conditional probability of receiving the treatment $a \in \mathscr{A}$. When $p = 2$, we let $\pi \equiv \pi_1$ be the propensity score.

4.4.2 Proposed estimator

To find conditions that the smoothing approximation from R_h to R is negligible is relatively straightforward (see Lemma C.3.3 in Section C.3.1 of the appendix). Therefore we will put more weight on finding an efficient estimator for the smoothed function R_h throughout the development. To show this emphasis, we use $\psi(Z;C,h,\eta) \equiv R_h(C)$ given a fixed codebook C, where η denotes a set of all nuisance parameters $(\pi_1,...,\pi_p,\mu_1,...,\mu_p)$.

In order to develop the efficient estimator for $\psi(Z;C,h,\eta)$ we use the efficient influence function approach. The *efficient influence function* is important to construct optimal estimators since its variance equals the efficiency bound (in asymptotic minimax sense). Using the efficient influence function also endows our estimators with favorable properties such as

[3] To be formal, all partial derivatives up to order kp must exist and be bounded.

double robustness or general second-order bias, which leads to relaxation of nonparametric conditions on the nuisance parameter estimation. There is at most one efficient influence function in nonparametric models. We refer the interested reader to Section 1.2 and references therein for more detailed information about the influence function and nonparametric effiency theory.

We let $\phi(Z;C,h,\eta)$ be the efficient influence function of $\psi(Z;C,h,\eta)$. We hide the dependency on Z, h and η when it is clear in context, and use shorthand notations $\phi_C \equiv \phi(Z;C,h,\eta)$, $\psi_C \equiv \psi(Z;C,h,\eta)$. Further, we let $\varphi_C \equiv (Z;C,h,\eta)$ denote the uncentered efficient influence function of ψ_C: i.e., $\varphi_C = \phi_C + \psi_C$. The next theorem gives the efficient influence function for our target parameter ψ_C under a nonparametric model.

Theorem 4.4.1 (Efficient influence function)**.** *Under a nonparametric model, the uncentered efficient influence function φ_C for ψ_C is as given by*

$$\varphi_C(Z) = \sum_{a\in\mathscr{A}} \left\{ 2\left[\sum_r f_r^a(\boldsymbol{\mu})\right] \sum_{a'\in\mathscr{A}} \left\{ \sum_r \left[\frac{\partial f_r^a}{\partial \mu_{a'}} \frac{\mathbb{1}(A=a')}{\pi_{a'}}(Y-\mu_{a'}) \right] \right\} + \left[\sum_r f_r^a(\boldsymbol{\mu})\right]^2 \right\}$$

where for $a,a' \in \mathscr{A}$

$$f_r^a(\boldsymbol{\mu};C,h) = \omega_r(\mu_a - c_{ra}),$$

$$\frac{\partial \omega_r}{\partial \mu_{a'}} = -\frac{\omega_r}{h}\left\{ \frac{\mu_{a'} - c_{ra'}}{\|\boldsymbol{\mu} - c_r\|_2} - \sum_j \frac{\mu_{a'} - c_{ja'}}{\|\boldsymbol{\mu} - c_j\|_2}\omega_j \right\}$$

, $c_r = [c_{r1},...,c_{rp}]^\top$. *The weight term ω_r is the same as in* (4.7).

The proof is given in Section C.3.1 of the appendix. We compute the remainder of the first order von Mises expansion of the efficient influence function in Lemma C.3.2 of the same section, which will be one of the key ingredients to develop our theory.

In order to flexibly incorporate modern machine learning tools without requiring complex empirical process conditions, we use sample splitting [19, 18]. We randomly split the observations $(Z_1,...,Z_n)$ into S disjoint groups, and let $\mathbb{P}_n^s$ denote the empirical measure only over the set of units in group s, $s \in \{1,...,S\}$, and let $\hat{\eta}_{-s}$ denote a set of the nuisance estimators constructed excluding the group s. Then for a given C, the efficient influence function based estimator for ψ_C is given by

$$\widehat{\psi}_C = \frac{1}{S}\sum_{s=1}^{S} \mathbb{P}_n^s\{\varphi(Z;C,h,\hat{\eta}_{-s})\}. \tag{4.8}$$

Finally, our proposed estimator for the optimal codebook C^* is given by

$$\widehat{C} = \underset{C \in \mathscr{C}_k}{\operatorname{argmin}} \widehat{\psi}_C. \tag{4.9}$$

In next section, we will argue that the proposed estimator $\widehat{C}$ in (4.9) indeed has favorable theoretical properties in regard to both excess risk and cluster codebook.

4.4.3 Theoretical Properties

In the following two subsections, we analyze theoretical properties of our estimator in two aspects: convergence in the excess risk and asymptotic normality of $\widehat{C}$.

Excess Risk Analysis

This subsection is devoted to finding conditions where the excess risk $R(\widehat{C}) - R(C^*)$ vanishes fast at $\sqrt{n}$ rates.

Given a codebook $C = \{c_1, ..., c_k\}$, we define the *Voronoi cell* associated with c_i as the closed set defined by

$$V_i(C) = \left\{ \boldsymbol{\mu} \mid \|\boldsymbol{\mu} - c_i\|_2 \leq \|\boldsymbol{\mu} - c_j\|_2, \forall j \neq i \right\},$$

and its boundary by

$$\partial V_i(C) = \left\{ \boldsymbol{\mu} \mid \|\boldsymbol{\mu} - c_i\|_2 = \|\boldsymbol{\mu} - c_j\|_2, \forall j \neq i \right\}.$$

Thus the entire boundary induced from a given quantization with C can be written by

$$\partial C = \bigcup_i \partial V_i(C).$$

Next, we define a neighborhood of ∂C in which the distance from $\boldsymbol{\mu}$ to nearest cluster centers only differs up to t. Namely, for $t > 0$ we define the *t-neighborhood* $N_C(t)$ by

$$N_C(t) = \bigcup_i \left\{ \boldsymbol{\mu} \in V_i(C) \;\middle|\; \min_{j \neq i} \left\{ \left| \|\boldsymbol{\mu} - c_j\|_2 - \|\boldsymbol{\mu} - c_i\|_2 \right| \right\} \leq t \right\}.$$

For example, in 2-dimensional Euclidean space for each pair c_i, c_j, $\left| \|\boldsymbol{\mu} - c_j\|_2 - \|\boldsymbol{\mu} - c_i\|_2 \right| \leq t$ forms a region surrounded by two hyperbolas which are symmetric around the line $\{\boldsymbol{\mu} \mid \partial V_i(C) = \partial V_j(C)\}$. Now we introduce the following (κ, α)*-margin condition*.

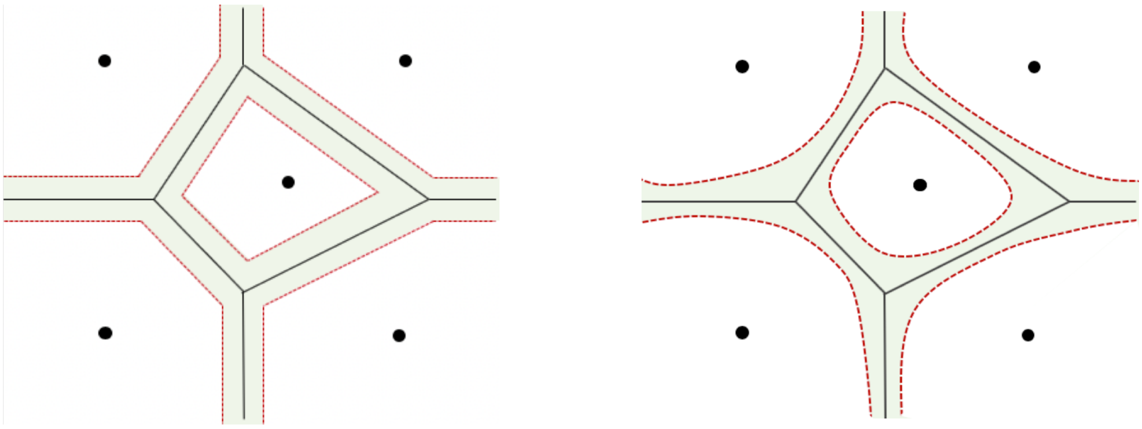


Fig. 4.2 Illustration for the margin condition in [85] (Left) and the margin condition in Definition 4.4.1 (Right), where we restrict the probability mass in the shaded area, inside red-dashed lines. Two areas are equal up to a constant.

Definition 4.4.1 ((κ,α)-Margin condition)**.** *Let us define* $p(t) := \sup_{C\in\mathscr{M}^*} \mathbb{P}(\boldsymbol{\mu} \in N_C(t))$. *A distribution* $\mathbb{P}$ *satisfies a* (κ,α)*-margin condition with radius* $\kappa > 0$ *and rate* $\alpha > 0$ *if and only if for all* $0 \leq t \leq \kappa$,

$$p(t) \lesssim t^{\alpha}.$$

The above margin condition requires a local control of the probability mass around ∂C for $C \in \mathscr{M}^*$, hence implies that every classification associated with an optimal codebook forms a natural classification in some sense. Here, smaller α implies weaker condition. Due to the boundedness assumption (A4), our margin condition is essentially equivalent to the margin condition used by Levrard [85, 86] who studied a nonasymptotic bounds for k-means clustering in the sense that the volumes of the t-neighborhood are equal up to a constant (see Figure 4.2) [4]. This type of margin condition is also adopted in causal inference problems involving estimation of non-smooth target parameters [e.g., 72, 135, 93].

Next theorem gives the conditions under which $\widehat{\psi}_C$ reasonably well approximates $R(C)$ when $C \in \mathscr{M}^*$.

Lemma 4.4.1. *Along with the causal and boundedness assumptions* (A1) $\sim$ (A4) *assume the following:*

(a) The (κ,α)*-margin condition*

(b) The estimators $\widehat{\mu}_a, \widehat{\pi}_a$ *are consistent in the sense that* $\|\widehat{\mu}_a - \mu_a\| = o_{\mathbb{P}}(1)$, $\|\widehat{\pi}_a - \pi_a\| = o_{\mathbb{P}}(1)$

[4] In the study of Levrard [85, 86], α is set to 1

(c) There exists $\gamma \in (0,1)$ *such that* $nh^{\alpha\gamma} = O(1)$

(d) $$\left(kh^{\frac{\alpha}{2}-1}+1\right)\sum_{a'\in\mathscr{A}}\|\pi_{a'}-\widehat{\pi}_{a'}\|_{\mathbb{P},4}\,\|\mu_{a'}-\widehat{\mu}_{a'}\|_{\mathbb{P},4}+\left(k^2h^{\frac{\alpha}{2}-2}+1\right)\sum_{a',a''\in\mathscr{A}}\|\mu_{a'}-\widehat{\mu}_{a'}\|_{\mathbb{P},4}\,\|\mu_{a''}-\widehat{\mu}_{a''}\|_{\mathbb{P},4}$$
$$+kh^{\alpha}+\frac{1}{nh^2}=o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)$$

Then for any optimal codebook $C \in \mathscr{M}^*$*, we have*

$$\sqrt{n}\,(\widehat{\psi}_C - R(C)) \rightsquigarrow N\left(0, var\left(\sum_{a\in\mathscr{A}} \bar{\phi}^{a}_{C^*}\right)\right)$$

where $\bar{\phi}^{a}_{C^*}$ *is determined in the proof.*

Hence given any optimal codebook $C \in \mathscr{M}^*$, $\widehat{\psi}_C$ is a $\sqrt{n}$-consistent, asymptotic normal estimator for the original risk $R(C)$. A proof of the above theorem is given in Section C.3.2 of the appendix.

In Assumption (a), the radius κ is only required to be fixed and positive, but $\alpha > 4$ must be satisfied due to assumptions (c) and (d). The left-hand side of Assumption (d) is basically the upper bound of the second-order remainder of the von Mises expansion of ϕ_C plus the last term of $1/nh^2$ which characterizes the lower bound of our bandwidth h.

Theorem 4.4.1 provides guarantees to achieve $\sqrt{n}$ rates in terms of the L_4 estimation rate for the nuisance parameters. Even though L_4 error rates are somewhat less common than L_2 rates (i.e., square loss), for many nonparametric classes of interest including smooth, Hölder, and Sobolev classes, minimax L_p error rates have been characterized for general $p > 0$ [e.g., 38, Corollary 1] and can be applied directly here. Moreover, if we assume the moment comparison condition [e.g., 38, Section 6] on our function class, L_4 error rates are always upper bounded by L_2 error rates. In this case, we can appeal to the result on ordinary L_2 error rates. See, for example, [38] and references therein for more detailed discussion.

The result in Lemma 4.4.1 is valid only for $C \in \mathscr{M}^*$. In order to analyze the excess risk, we show consistency of $\widehat{C}$ in the following lemma.

Lemma 4.4.2. *Along with the assumptions (a) - (d) in Lemma 4.4.1, assume that*

(e) C^* *is unique up to relabeling of its coordinates,*

Then $\widehat{C}$ *converges in probability to* C^**.*

A proof can be found in Section C.3.3 of the appendix. The uniqueness condition (e) is also used in the previous work of [101, 102] in order to show consistency of the empirical risk minimizer $\widehat{C}^*$. Based on Lemma 4.4.2, 4.4.1, we compute an asymptotic bound for the excess risk as stated in the next theorem.

Theorem 4.4.2. *Suppose (a) - (e). Then we have*

$$R(\widehat{C}) - R(C^*) = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right).$$

See Section C.3.4 of the appendix for the proof. It is worth noting that since $R(\cdot)$ is a continuous, bounded function whose domain $\mathscr{C}_k$ is compact, the uniqueness condition (e) guarantees that C^* is a *well-separated* point of maximum of R; in other words $\inf_{C \notin \mathbb{B}_\delta(\mathscr{M}^*)} R(C) > R(C^*)$ for any $C^* \in \mathscr{M}^*$ and every $\delta > 0$ where $\mathbb{B}_\delta(\mathscr{M}^*) = \{C : d_{\text{codebook}}(C, C^*) < \delta, \forall C^* \in \mathscr{M}^*\}$ for any valid metric d_{codebook} for codebooks.

Here we give a brief discussion of bandwidth selection for our estimator. For finite α, our result holds as long as $h = \Omega(n^{-1/4})$ and $h = O(n^{-1/\alpha\gamma})$ for some $\gamma \in (0,1)$ and $\alpha > 4$. α should be as small as possible as since larger α implies that we require the stronger margin condition. Consequently, one could take $h \sim n^{-1/4} \log n$. Unfortunately when we assume the optimistic strong margin condition, it is not clear how to characterize the optimal h. This issue will be discussed in some more detail later in Chapter 4.6.

Asymptotic Normality of Codebook

One may find our approach more beneficial if we can apply the central limit theorem argument to $\widehat{C}$. In this subsection, we find conditions to assure $\sqrt{n}$-consistency as well as asymptotic normality of $\widehat{C}$. To this end we consider (κ, ∞)-margin condition where we assume zero probability mass inside the κ-neighborhood $N_C(\kappa)$ for a given C and some $\kappa > 0$. Although κ can be arbitrarily small, the above condition is much stronger than the original margin condition as it implies that every classification associated with optimal codebooks must form a non-overlapping, hard-margin natural classifier.

In what follows, we show an asymptotic normality of our estimated codebook $\widehat{C}$ using this stronger version of margin condition.

Theorem 4.4.3. *Under the assumptions (a) - (e) in Lemma 4.4.1 and Lemma 4.4.2, where we replace Assumption (a) by (κ, ∞)-margin condition, we have*

$$\sqrt{n}(\widehat{C} - C^*) \rightsquigarrow N\left(0, \Sigma'_{C^*,\eta}\right)$$

where the $kp \times kp$ covariance matrix $\Sigma'_{C^*,\eta}$ is specified in (C.25) in Section C.3.5 of the appendix.

The stronger version of margin condition is the price we pay to make the central limit theorem applicable to our $\widehat{C}$. A proof of this theorem is given in Section C.3.5 of the appendix.

4.5 Experiments

4.5.1 Simulation Study

Here we explore finite-sample properties for the k-means causal clustering approaches that we developed in Section 4.3.1 and 4.4. Simulation is particularly designed to demonstrate validity of our theoretical results in Section 4.4.3.

We consider the following data generating process with sample size n. First, we fix k, p, each of which is randomly drawn from a set $\{2, ..., 10\}$. For each pair (k, p), we randomly pick k points in a bounded hypercube $[0,1]^p$ in a way that every pairwise mutual Euclidean distance is always greater than 0.2. A set of these k points is our true (optimal) codebook $C^* = \{c_1^*, ..., c_k^*\}$ [5]. Then we assign roughly equal number of units to each cluster center c_j^*, $j = 1, 2, .., k$. Specifically for each unit $i = 1, ..., n$, we draw a label $I \in \{1, ..., k\}$ from a multinomial distribution: $\text{multi}(p_1, ..., p_k)$ with $p_1 = \cdots = p_k = 1/k$. Given this label information, we set $\boldsymbol{\mu} = c_I + \varepsilon^{\text{truc}}$ where $\varepsilon^{\text{truc}}$ follows a truncated normal distribution of $N(0, 1/2)$ with the threshold of $\min\limits_{c_i, c_j \in C} d(c_i, c_j)/2 - 0.01$. This guarantees that the nearest center for units with label j is always c_j. Next, we model our observed data generating process by $A \sim \text{multi}(\pi_1, ..., \pi_p)$ and $Y = \mu_A + Z$, where all the π's are roughly equal and $Z \sim N(0,1)$. Finally, we assume that $\widehat{\mu}_a = \mu_a + \xi$ and $\widehat{\pi}_a = \pi_a + \zeta$, where $\xi \sim N(0, n^{-(r_\mu + 0.01)})$ and $\zeta \sim N(0, n^{-(r_\pi + 0.01)})$ respectively.

Note that under the above simulation setup, we have $\|\widehat{\mu}_a - \mu_a\|_{\mathbb{P},4} = o(n^{-r_\mu})$ and $\|\widehat{\pi}_a - \pi_a\|_{\mathbb{P},4} = o(n^{-r_\pi})$. This is simple yet enough to verify our theoretical results. For example, when $\widehat{\mu}_a$ converges to its true values at slower rates (e.g., when $r_\mu = 1/4$), the nonparametric plug-in estimator proposed in Section 4.3.1 should not perform better than the efficient k-means causal clustering in Section 4.4.

We randomly pick 10 different pairs of (k, p) and vary the sample size n from 250 to 10k for each (k, p) pair. For each (k, p, n) tuple, we generate data according to the above specified process, and then compute $\widehat{C}_{\text{pi}}$, the plug-in estimator in (4.5), and $\widehat{C}_{\text{eff}}$, the efficient estimator in (4.9), and their risk $R(\widehat{C}_{\text{pi}})$ and $R(\widehat{C}_{\text{eff}})$, respectively. We repeat the simulation $J = 100$

[5] Rigorously speaking, the true codebook in this setting are not exactly the optimal codebook C^* defined to be a minimizer of the risk function as in Section 4.3.1. However, for sufficiently large n ($>\sim$ 1000) they become almost identical.

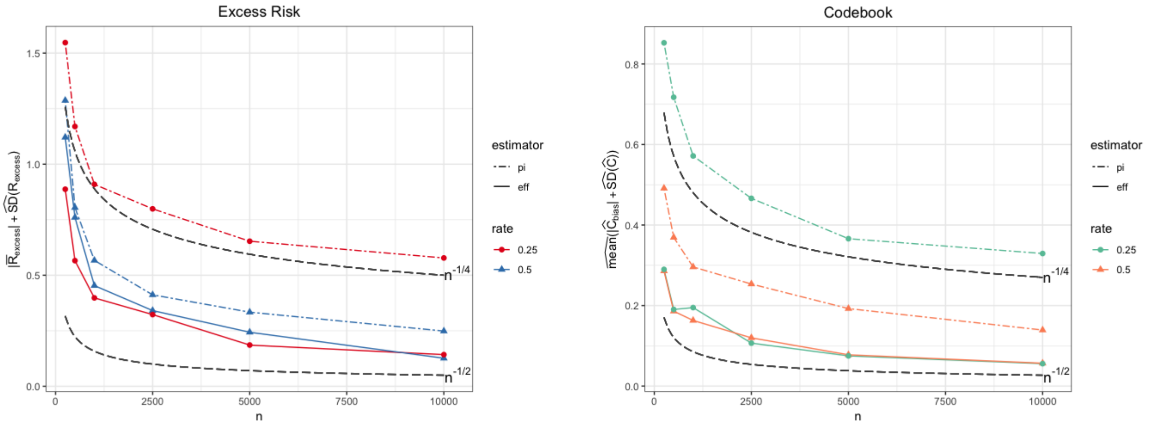


Fig. 4.3 Finite sample performance of the plug-in estimator (pi) and the efficient influence function based estimator (eff) with respect to the excess risk (left) and codebook (right), across different sample sizes (n=250 $\sim$ 10k) and nuisance estimation rates (1/4, 1/2). Each point is obtained with 100 simulations. Two reference curves in black dotted line are scaled by an appropriate constant.

times for each (k,p,n). The entire simulation are done twice across different nuisance estimation rates: i.e., $(r_\mu, r_\pi) = (1/2, 1/2), (1/4, 1/4)$.

First, we consider the excess risk $R_{\text{excess}} = R(\widehat{C}) - R(C^*)$. Then the performance of estimators in excess risk is assessed via

$$\left|\overline{R}_{\text{excess}}\right| + \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(R_{\text{excess},j} - \overline{R}_{\text{excess}}\right)^2} \equiv \left|\overline{R}_{\text{excess}}\right| + \widehat{SD}\left(R_{\text{excess}}\right)$$

where $\overline{R}_{\text{excess}} = \frac{1}{J}\sum_{j=1}^{J} R_{\text{excess},j}$, an average over J simulations. Next, we also assess accuracy of $\widehat{C}$ for estimating the true codebook C^* similarly via

$$\frac{1}{kp}\sum_{r,a}\left\{\left|\bar{c}_{ra} - c^*_{ra}\right| + \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(c_{ra,j} - \bar{c}_{ra}\right)^2}\right\} \equiv \widehat{\text{mean}}\left(\left|\widehat{C}_{\text{bias}}\right| + \widehat{SD}(\widehat{C})\right)$$

where $\bar{c}_{ra} = \frac{1}{J}\sum_{j=1}^{J} c_{ra,j}$. Here $\widehat{\text{mean}}$ and $\widehat{SD}$ represent sample mean and sample standard deviation operators respectively. We use $h = n^{-1/4}\log n$ for the kernel bandwidth and random starting values for the minimization step in (4.9). Finally for each n, values for these performance measures are averaged over different (k,p) pairs. Results are given in Figure 4.3.

For both fast ($r_\mu = r_\pi = 1/2$) and slow ($r_\mu = r_\pi = 1/4$) rates at which the nuisance functions are estimated, the performance of $R(\widehat{C}_{\text{eff}})$ and $\widehat{C}_{\text{eff}}$ with respect to their true value

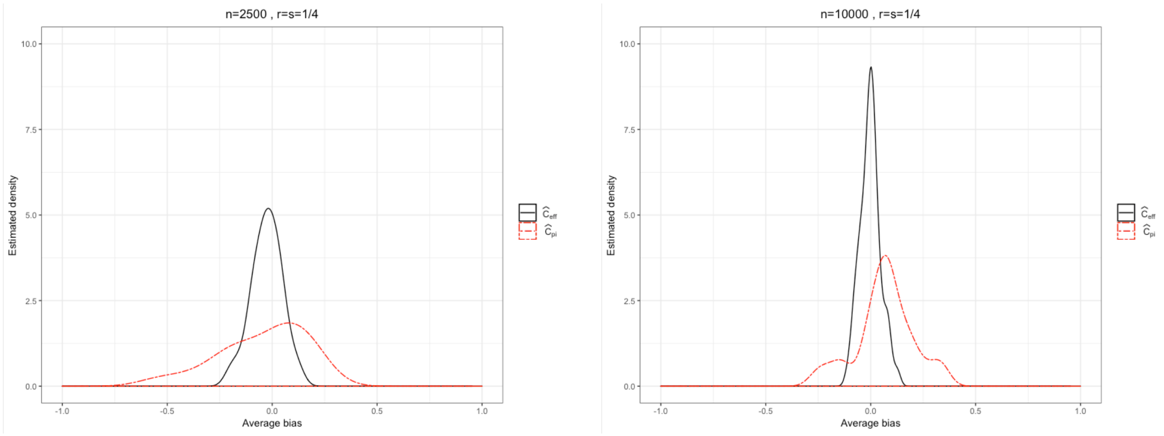


Fig. 4.4 Estimated density of average bias across all coordinates in $\widehat{C} - C^*$ at different sample sizes (n=2.5k, 10k), when simulation is repeated 100 times each. Nuisance functions are estimated at $n^{1/4}$ rates.

is improved as n grows, nearly at $n^{1/2}$ rates. This is expected by Theorem 4.4.2 and Theorem 4.4.3. On the other hand, when the nuisance functions are estimated at the slow rates the plug-in based estimators $R(\widehat{C}_{\text{pi}})$ and $\widehat{C}_{\text{pi}}$ show much worse performance, roughly at $n^{1/4}$ rates, since they are no longer expected to converge at $n^{1/2}$ rates as described in Theorem 4.3.1.

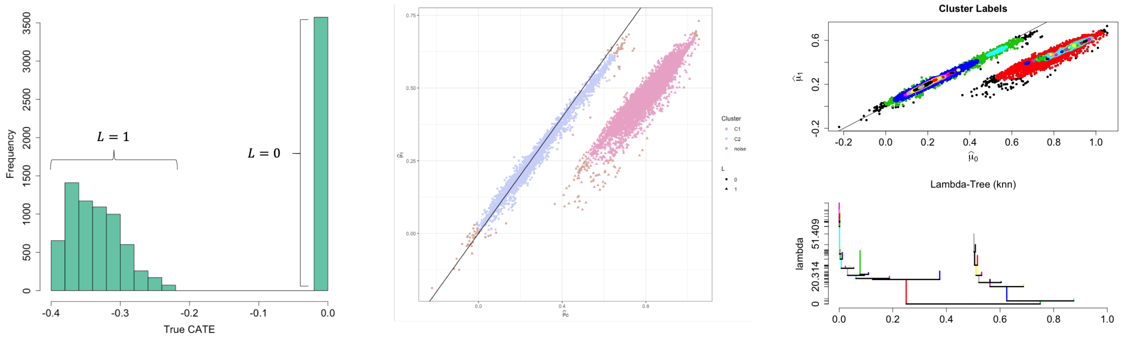
To further verify benefits of the efficient k-means causal clusters in Section 4.4, we also estimate density of average bias $\frac{1}{kp}\sum_{r,a}(\widehat{c}_{ra} - c^*_{ra})$ across 100 simulations for $\widehat{C}_{\text{eff}}$ and $\widehat{C}_{\text{pi}}$ respectively, at two different sample sizes $n = 2.5$k, 10k. Here all the nuisance functions are estimated at the slow $n^{1/4}$ rates. As shown in Figure 4.4, $\widehat{C}_{\text{eff}}$ is substantially more concentrated around the true values.

4.5.2 Illustration

In this section, we illustrate our method through two case studies. We use the semi-synthetic data on voting study [99] and the real-world data on substance abuse treatment [95].

Voting study. Nie and Wager [99] considered a dataset on the voting study originally used by Arceneaux et al. [2], where they generated synthetic treatment effect to make the task of estimating heterogeneous treatment effects non-trivial. We use the same setup of Nie and Wager [99, Chapter 2], where we have binary treatments, binary outcomes, and 11 pretreatment covariates (including state, age, gender, etc.), and the true CATE $\tau(\cdot)$ in (4.1.1) is known [6]. While Nie and Wager [99] specifically focused on accurate estimation of $\tau(\cdot)$, here we aim to illustrate how our causal clustering can be useful to discover an interesting

[6] Roughly 36% of samples are set to have zero CATE values.



(a) (b) (b)

Fig. 4.5 (a) Histogram of the true CATE in the test set. We define a true label L as an indicator variable whose value is 1 for negative CATE. (b) The result of density-based causal clustering. Units in the two clusters C1 and C2 are assigned to $\widehat{L} = 1$ and $\widehat{L} = 0$, respectively. (c) Points in the clusters C1 and C2 are concentrated around the right upper area (large μ_0, μ_1) and the lower left area (small μ_0, μ_1), respectively.

subgroup structure. We randomly chose a training set of size 130,000 and a test set of size 10,000 from the entire sample, and estimate the conditional counterfactual mean vector $\boldsymbol{\mu}$ in the training set and perform the causal clustering in the test set.

We fit models for $\boldsymbol{\mu}$ via Random Forests (RF), Generalized Boosted Models (GBM), and Lasso using `ranger`, `gbm`, `glmnet` R packages respectively, and chose the GBM based on the cross-validated (CV) error. Points for $\widehat{\boldsymbol{\mu}}$ show non-spherical shapes so we proceed with the level-set density clustering discussed in Section 4.3.3 [7], via the `TDA` R package. We only consider two clusters corresponding to the two largest branches at the bottom of the tree (see Figure 4.5-(c)). Roughly 4% of the points are classified as noise.

In Figure 4.5-(b), we see two clusters that are clearly separable from each other, one with nearly zero CATE (Cluster C2) and the other with substantially negative CATE (Cluster C1), which seems consistent with the shape of the histogram of true CATE shown in Figure 4.5-(a). To verify that we did not get our findings just by chance, we define the true label $L := \mathbb{1}\{\tau(X) < 0\}$ and its estimate via the causal clustering $\widehat{L} := \mathbb{1}\{\widehat{\boldsymbol{\mu}} \in \text{C1}\}$. We repeat simulation 100 times, each with different synthetic effect assignment, and compute the error $\mathbb{P}_n\{\mathbb{1}(L \neq \widehat{L})\}$ in the test set across different simulations. All the errors are exactly zero, and thus we confirm our finding is not a coincidence.

[7] Typically, this plug-in method suffers from inefficiency compared to the efficient k-means method proposed in Section 4.4, but here it can be justified by a large number of samples.

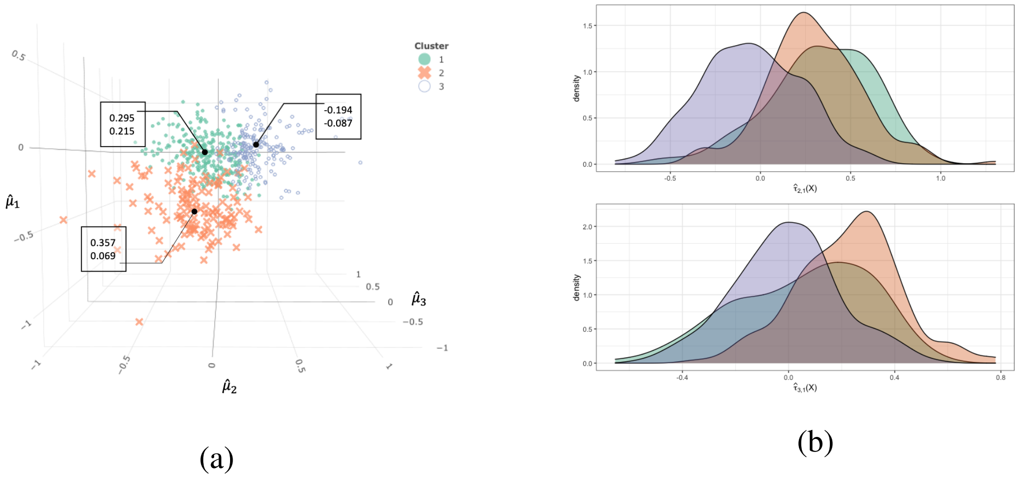


Fig. 4.6 (a) The three clusters in $\widehat{\boldsymbol{\mu}}$. The average CATEs of receiving the MET&CBT-5 (upper) and the SCY (lower) over traditional programs (community) are presented together inside the box for each cluster. (b) The density plots for the pairwise CATEs $\widehat{\tau}_{2,1}(X)$ (upper) and $\widehat{\tau}_{3,1}(X)$ (lower) across three clusters.

Another interesting fact which can be discovered by causal clustering here is the difference in distribution of (μ_0, μ_1) between the two clusters. In general, units in the cluster C1 have larger μ_0, μ_1 than the cluster C2. This is more clearly illustrated in Figure 4.5-(c).

Substance abuse treatment. McCaffrey et al. [95] studied the relative effects of three treatment programs (community/MET&CBT-5/SCY) for adolescent substance abuse. Instead of using the full data originally collected by the Substance Abuse Mental Health Services Administration's Center for Substance Abuse Treatment (SAMHSA CSAT), we use a random subset of the data which is readily available via the `twang` R package. The subset of data to be analyzed contains 600 samples, 200 youths in each treatment, and 5 covariates (age, ethnicity, criminal history, etc.). Our outcome is the substance frequency score, where higher scores indicate increased frequency of substance use. See McCaffrey et al. [95], Burgette et al. [15] for a more detailed description of the dataset.

Cross-sectional scatter plots reveal several spherical chunks in the estimated conditional counterfactual mean vector space ($\widehat{\boldsymbol{\mu}}$). To implement the efficient k-means clustering algorithm, we set $k = 3$ for the number of clusters, which is determined by the Elbow Method, $S = 2$ for sample splitting, and use $h \sim n^{-1/4} \log n$ as before. Here we fit the RF model for all the nuisance components $\widehat{\mu}$'s and $\widehat{\pi}$'s as it delivers the lowest CV error. The results are presented in Figure 4.6.

In Figure 4.6(a), we compute the average (pairwise) CATEs of receiving the MET&CBT-5 ($\widehat{\tau}_{2,1}$) and the SCY ($\widehat{\tau}_{3,1}$) treatment programs over the traditional community program, respectively, within each of the three clusters. In Figure 4.6(b), we also present the density plots for each pairwise CATE across different clusters. Our result suggests that there is a moderate degree of treatment effect heterogeneity.

4.6 Discussion

Causal clustering is a novel methodological framework that provides an effective, and potentially more intuitive way of analyzing treatment effect heterogeneity by leveraging tools in clustering analysis. Based on what we propose, one may benefit from flexible unsupervised machine learning tools to uncover subgroup structure and ascertain subgroups with similar conditional treatment effects, even with multiple treatments and outcome-wide studies. We showed that k-means, density-based, and hierarchical clustering algorithms can be successfully adopted into our framework, and also developed an efficient k-means causal clustering based on nonparametric efficiency theory that attains fast convergence rates and asymptotic normality.

There are a couple of caveats to our developments that are worth mentioning. First, as mentioned in Section 4.4.3, our kernel bandwidth choice problem is not completely solved. Although it is always safe to assume small α, it would be better if we could also rely on a data-driven method to pick the optimal bandwidth. Unfortunately, unlike standard tuning parameter selection problems there is no clear way to estimate the risk, and thus we cannot rely on cross-validation. Addressing the optimal bandwidth choice problem in a data-driven way would be an interesting topic to pursue in future work. Second, albeit in a different context, some of previous work that also studied estimation of non-smooth parameters in the causal inference literature (for example, [135, 93] on optimal treatment regime and [72] on classification of compilers) required only the margin condition to guarantee fast $\sqrt{n}$ rates and asymptotic normality, whereas we required both the margin condition and kernel smoothing approach, which is typically not the case in the nonparametric literature. We conjecture that the reason why we require the both conditions arises from the increased complexity in our target parameter $R(c)$; the special non-smooth function Π_C might have brought the additional complexity which either the margin condition or the smoothing approach alone is not enough to deal with. We leave the formal discussion on this topic for future work.

Our study leads to many opportunities for important future work. For example, we plan to apply causal clustering tools in an optimal treatment regime framework, considering the optimal rule among those that map clusters to treatment decisions. It will be also useful to evaluate optimality of our estimators by computing minimax bounds. Furthermore, it would be interesting to consider clustering directly on counterfactual outcomes Y^a instead of μ_a. To this end, one may pursue to explore a link between previous work on clustering on partially observed data and clustering on Y^a, since clustering on Y^a can be framed as missing data problem in a vector form in fixed dimensions. However, this may require extra assumptions both for identification and estimation, which would lead to another interesting future work.

Chapter 5

Conclusion

In this thesis, I have extended methods in causal inference to novel, non-standard effects with complex data structures by adapting techniques in statistical machine learning and semiparametric theory. Methodologies developed in this thesis pursue a more nuanced way to explore causal effects beyond the ATE and simple data structure, making efficient use of the information in data while avoiding unnecessary assumptions about the underlying data generating mechanism. Many of my research questions can be framed as developing optimal nonparametric estimators of complex statistical functionals; in them I explore how to effectively harness advanced machine learning tools (e.g., techniques in unsupervised learning) to address crucial issues in modern causal inference.

Although I have already enumerated many promising future directions which can be potentially expanded from what is done in this thesis at the end of each chapter, I would like to highlight a few topics that I particularly plan to pursue in the near future. First, I expect many results in this thesis to play an important role to provide new insight into learning how to best assign treatment when effects are non-standard, in the context of optimal treatment regime estimation. For example, it will be important to consider how to effectively construct specific treatment decision rules with observational data when positivity is likely violated, effects of interest are non-standard, or we have considerable degree of heterogeneity in treatment effects. I plan to apply tools developed in this thesis in an optimal treatment regime framework. Second, it would be interesting to extend the given results to the other identification setups, i.e., instrumental variables, mediation, etc. Third, throughout some parts of the thesis a kernel smoothing approach has been used in an attempt to develop efficient nonparametric estimators for non-smooth functionals. However, due to the highly nontrivial nature of our original target functional, it is usually not obvious how to select the kernel bandwidth; unlike standard tuning parameter selection problems since we do not have access to all ground truth data there is no clear way to estimate the risk, and thus we cannot rely on cross-validation. It will be useful to develop a general data-driven approach for optimal bandwidth selection in the context of semiparametric causal inference with nonparametric functional estimation. Finally, since the methods and results developed in this thesis can accommodate more complex data structures and subtle effects, they would lead to many opportunities for interesting applied work with various real-world data. I plan to apply some of the tools developed in this thesis in such applied work during my postdoctoral experience.

I expect my work on causal inference with non-standard effects and complex data structures will produce substantial contributions to the literature and to statistical practice for modern causal inference. The R code for all developed methods will be publicly available, allowing researchers across many fields to go beyond simple effects and learn more valuable information about causality.

References

[1] Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267.

[2] Arceneaux, K., Gerber, A. S., and Green, D. P. (2006). Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, 14(1):37–62.

[3] Aronow, P. M. (2016). Local average causal effects and superefficiency. *arXiv preprint arXiv:1601.01413*.

[4] Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

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

[6] Balcan, M.-F., Blum, A., and Srebro, N. (2008). A theory of learning with similarity functions. *Machine Learning*, 72(1-2):89–112.

[7] Balcan, M.-F., Liang, Y., and Gupta, P. (2014). Robust hierarchical clustering. *The Journal of Machine Learning Research*, 15(1):3831–3871.

[8] Bartlett, P. L., Linder, T., and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information theory*, 44(5):1802–1813.

[9] Belloni, A., Chernozhukov, V., Chetverikov, D., and Wei, Y. (2015). Uniformly valid post-regularization confidence regions for many functional parameters in z-estimation framework. *arXiv preprint arXiv:1512.07619*.

[10] Biau, G., Devroye, L., and Lugosi, G. (2008). On the performance of clustering in hilbert spaces. *IEEE Transactions on Information Theory*, 54(2):781–790.

[11] Bickel, P., Klaassen, C., Ritov, Y., and Wellner, J. (1998). *Efficient and adaptive estimation for semiparametric models*. Springer.

[12] Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

[13] Bobrowski, O., Mukherjee, S., Taylor, J. E., et al. (2017). Topological consistency via kernel estimation. *Bernoulli*, 23(1):288–328.

[14] Boos, D. D. and Stefanski, L. A. (2013). *Essential statistical inference: theory and methods*, volume 120. Springer Science & Business Media.

[15] Burgette, L., Griffin, B. A., and McCaffrey, D. (2017). Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package. *R package. Rand Corporation*.

[16] Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. (2013). On the Bootstrap for Persistence Diagrams and Landscapes. *Model. Anal. Inform. Sist.*, 20:111–120.

[17] Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2017). Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, 112(520):1684–1696.

[18] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

[19] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. K. (2016). Double machine learning for treatment and causal parameters. Technical report, cemmap working paper.

[20] Chernozhukov, V., Chetverikov, D., Kato, K., et al. (2014). Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597.

[21] Chernozhukov, V. and Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261.

[22] Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

[23] Dasgupta, S. and Long, P. M. (2005). Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences*, 70(4):555–569.

[24] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L1 View*. New York: John Wiley & Sons.

[25] Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

[26] Díaz, I. and Hejazi, N. (2019). Causal mediation analysis for stochastic interventions. *arXiv preprint arXiv:1901.02776*.

[27] Díaz, I. and van der Laan, M. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549.

[28] Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM.

[29] Doove, L. L., Dusseldorp, E., Van Deun, K., and Van Mechelen, I. (2014). A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. *Advances in Data Analysis and Classification*, 8(4):403–425.

[30] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26.

[31] Eriksson, B., Dasarathy, G., Singh, A., and Nowak, R. (2011). Active clustering: Robust and efficient hierarchical clustering using adaptively selected similarities. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 260–268.

[32] Ertefaie, A. and Strawderman, R. L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105(4):963–977.

[33] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

[34] Eysenbach, G., Group, C.-E., et al. (2011). Consort-ehealth: improving and standardizing evaluation reports of web-based and mobile health interventions. *Journal of medical Internet research*, 13(4).

[35] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., Singh, A., et al. (2014). Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339.

[36] Ferraty, F. and Vieu, P., editors (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Verlag.

[37] Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276.

[38] Foster, D. J. and Syrgkanis, V. (2019). Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*.

[39] Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880.

[40] Gine, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.*, 18(2):851–869.

[41] Graf, S. and Luschgy, H. (2007). *Foundations of quantization for probability distributions*. Springer.

[42] Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511.

[43] Grimmer, J., Messing, S., and Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434.

[44] Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

[45] Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

[46] Haneuse, S. and Rotnitzky, A. (2013). Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine*, 32(30):5260–5277.

[47] Hanushek, E. A. (2016). Equality of Educational Opportunity. Technical report.

[48] Hartigan, J. A. (1975). Clustering algorithms.

[49] Haviland, A. M., Jones, B. L., and Nagin, D. S. (2011). Group-based trajectory modeling extended to account for nonrandom participant attrition. *Sociological Methods & Research*, 40(2):367–390.

[50] Hayden, E. C. (2009). Personalized cancer therapy gets closer.

[51] Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, pages 561–570.

[52] Hernan, M. A. and Robins, J. M. (2019). *Causal inference*. CRC Boca Raton, FL:.

[53] Hinneburg, A., Keim, D. A., et al. (1998). An efficient approach to clustering in large multimedia databases with noise. In *KDD*, volume 98, pages 58–65.

[54] Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

[55] Imai, K., Ratkovic, M., et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

[56] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[57] Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

[58] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

[59] Jiao, J., Han, Y., and Weissman, T. (2016). Minimax estimation of the l 1 distance. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 750–754. IEEE.

[60] Kandasamy, K., Krishnamurthy, A., Poczos, B., Wasserman, L., and Robins, J. M. (2014). Influence Functions for Machine Learning: Nonparametric Estimators for Entropies, Divergences and Mutual Informations. *ArXiv e-prints*.

[61] Kandasamy, K. and Yu, Y. (2016). Additive approximations in high dimensional nonparametric regression via the salsa. In *International Conference on Machine Learning*, pages 69–78.

[62] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):881–892.

[63] Kennedy, Edward H, B. S. and Wasserman, L. (2020). *Modern Causal inference*. Preprint.

[64] Kennedy, E. H. (2014). Semiparametric theory. *Wiley StatsRef: Statistics Reference Online*, pages 1–7.

[65] Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer.

[66] Kennedy, E. H. (2018). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 0(ja):0–0.

[67] Kennedy, E. H. (2019). Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656.

[68] Kennedy, E. H. (2020a). Lecture notes for foundations of causal inference (36-731).

[69] Kennedy, E. H. (2020b). Lecture notes for modern causal inference (36-732).

[70] Kennedy, E. H. (2020c). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*.

[71] Kennedy, E. H. (2020d). Tutorials on nonparametric causal inference & functional estimation.

[72] Kennedy, E. H., Balakrishnan, S., and G'Sell, M. (2018). Sharp instruments for classifying compliers and generalizing causal effects. *arXiv preprint arXiv:1801.03635*.

[73] Kennedy, E. H., Lorch, S., and Small, D. S. (2019). Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):121–143.

[74] Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245.

[75] Kent, B. P., Rinaldo, A., and Verstynen, T. (2013). Debacl: A python package for interactive density-based clustering. *arXiv preprint arXiv:1307.8136*.

[76] Kim, J., Shin, J., Rinaldo, A., and Wasserman, L. (2019). Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3398–3407, Long Beach, California, USA. PMLR.

[77] Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. A. (2015). Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S):1220.

[78] Kosorok, M. (2008). *Introduction to empirical processes and semiparametric inference*. Springer series in statistics. Springer.

[79] Kravitz, R. L., Duan, N., and Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82(4):661–687.

[80] Kumar, M. and Patel, N. R. (2007). Clustering data with measurement errors. *Computational Statistics & Data Analysis*, 51(12):6084–6101.

[81] Kumar, S., Nilsen, W. J., Abernethy, A., Atienza, A., Patrick, K., Pavel, M., Riley, W. T., Shar, A., Spring, B., Spruijt-Metz, D., et al. (2013). Mobile health technology evaluation: the mhealth evidence workshop. *American journal of preventive medicine*, 45(2):228–236.

[82] Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2017). Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*.

[83] Laber, E. B., Meyer, N. J., Reich, B. J., Pacifici, K., Collazo, J. A., and Drake, J. M. (2018). Optimal treatment allocations in space and time for on-line control of an emerging infectious disease. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(4):743–789.

[84] Leskovec, J., Rajaraman, A., and Ullman, J. D. (2020). *Mining of massive data sets*. Cambridge university press.

[85] Levrard, C. (2014). Non asymptotic bounds for vector quantization. *ArXiv e-prints*.

[86] Levrard, C. (2018). Quantization/clustering: when and why does k-means work? *arXiv preprint arXiv:1801.03742*.

[87] Levrard, C. et al. (2013). Fast rates for empirical vector quantization. *Electronic Journal of Statistics*, 7:1716–1746.

[88] Li, S., Stampfer, M. J., Williams, D. R., and VanderWeele, T. J. (2016). Association of religious service attendance with mortality among women. *JAMA internal medicine*, 176(6):777–785.

[89] Linder, T., Lugosi, G., and Zeger, K. (1994). Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40(6):1728–1740.

[90] Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366.

[91] Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

[92] Luedtke, A. R., Sofrygin, O., van der Laan, M. J., and Carone, M. (2017). Sequential double robustness in right-censored longitudinal models. *arXiv preprint arXiv:1705.02459*.

[93] Luedtke, A. R. and Van Der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713.

[94] MA, H. and JM, R. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7):578–586.

[95] McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414.

[96] Moore, K. L., Neugebauer, R., van der Laan, M. J., and Tager, I. B. (2012). Causal inference in epidemiological studies with strong confounding. *Statistics in medicine*, 31(13):1380–1404.

[97] Mumford, S. L., Silver, R. M., Sjaarda, L. A., Wactawski-Wende, J., Townsend, J. M., Lynch, A. M., Galai, N., Lesher, L. L., Faraggi, D., Perkins, N. J., et al. (2016). Expanded findings from a randomized controlled trial of preconception low-dose aspirin and pregnancy loss. *Human Reproduction*, 31(3):657–665.

[98] Murphy, S. A., van der Laan, M. J., Robins, J. M., and Group, C. P. P. R. (2001). Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423.

[99] Nie, X. and Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.

[100] Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

[101] Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, pages 135–140.

[102] Pollard, D. et al. (1982). A central limit theorem for *k*-means clustering. *The Annals of Probability*, 10(4):919–926.

[103] Powell, D. (2013). A new framework for estimation of quantile treatment effects: Nonseparable disturbance in the presence of covariates.

[104] Qian, T., Russell, M. A., Collins, L. M., Klasnja, P., Lanza, S. T., Yoo, H., and Murphy, S. A. (2020). The micro-randomized trial for developing digital interventions: Data analysis methods. *arXiv preprint arXiv:2004.10241*.

[105] Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics.

[106] Raskutti, G., Wainwright, M. J., and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(Feb):389–427.

[107] Rinaldo, A. and Wasserman, L. (2010). Generalized density clustering. *Ann. Statist.*, 38(5):2678–2722.

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

[109] Robins, J., Tchetgen, E. T., Li, L., and van der Vaart, A. (2009). Semiparametric minimax rates. *Electronic journal of statistics*, 3:1305.

[110] Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer.

[111] Robins, J. M. and Hernán, M. A. (2008). Estimation of the causal effects of time-varying exposures. In *Longitudinal data analysis*, pages 547–593. Chapman and Hall/CRC.

[112] Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

[113] Robins, J. M. and Rotnitzky, A. (2001). Comment on the bickel and kwon article,“inference for semiparametric models: Some questions and an answer”. *Statistica Sinica*, 11(4):920–936.

[114] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

[115] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, 90(429):106–121.

[116] Rothe, C. (2010). Nonparametric estimation of distributional policy effects. *Journal of Econometrics*, 155(1):56–70.

[117] Rubin, D. and van der Laan, M. J. (2006). Extending marginal structural models through local, penalized, and additive learning.

[118] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

[119] Schisterman, E. F., Silver, R. M., Lesher, L. L., Faraggi, D., Wactawski-Wende, J., Townsend, J. M., Lynch, A. M., Perkins, N. J., Mumford, S. L., and Galai, N. (2014). Preconception low-dose aspirin and pregnancy outcomes: results from the eager randomised trial. *The Lancet*, 384(9937):29–36.

[120] Schisterman, E. F., Silver, R. M., Perkins, N. J., Mumford, S. L., Whitcomb, B. W., Stanford, J. B., Lesher, L. L., Faraggi, D., Wactawski-Wende, J., Browne, R. W., et al. (2013). A randomised trial to evaluate the effects of low-dose aspirin in gestation and reproduction: design and baseline characteristics. *Paediatric and perinatal epidemiology*, 27(6):598–609.

[121] Serafini, A., Murphy, T. B., and Scrucca, L. (2020). Handling missing data in model-based clustering. *arXiv preprint arXiv:2006.02954*.

[122] Shalit, U., Johansson, F. D., and Sontag, D. (2016). Estimating individual treatment effect: generalization bounds and algorithms. *arXiv preprint arXiv:1606.03976*.

[123] Su, Y., Reedy, J., and Carroll, R. J. (2018). Clustering in general measurement error models. *Statistica Sinica*, 28(4):2337.

[124] Sun, B. and Tchetgen, E. J. T. (2014). On inverse probability weighting for nonmonotone missing at random data. *arXiv preprint arXiv:1411.5310*.

[125] Szabo, Z., Sriperumbudur, B., Póczos, B., and Gretton, A. (2014). Learning Theory for Distribution Regression. *Journal of Machine Learning Research*.

[126] Tchetgen, E. J. T., Wang, L., and Sun, B. (2016). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. *arXiv preprint arXiv:1607.02631*.

[127] Tseng, G. C. and Wong, W. H. (2005). Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, 61(1):10–16.

[128] Tsiatis, A. (2006a). *Semiparametric Theory and Missing Data*. Springer Verlag New York.

[129] Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

[130] Tsiatis, A. A. (2006b). *Semiparametric Theory and Missing Data*. New York: Springer.

[131] Tsybakov, A. B. (2010). *Introduction to Nonparametric Estimation*. Springer Series in Statistics, 1st edition.

[132] Vaart, A. W. v. d. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[133] van der Laan & James M Robins, M. J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Verlag New York.

[134] van der Laan, M. J. and Luedtke, A. R. (2014). Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome.

[135] van der Laan, M. J. and Luedtke, A. R. (2015). Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of causal inference*, 3(1):61–95.

[136] Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1).

[137] van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.

[138] Van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

[139] van der Vaart, A. (2000). *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press.

[140] van der Vaart, A. (2002). *Semiparametric statistics*, pages 331–457. Number 1781 in Lecture Notes in Math. Springer. MR1915446.

[141] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[142] Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

[143] VanderWeele, T. J. (2017). Outcome-wide epidemiology. *Epidemiology (Cambridge, Mass.)*, 28(3):399.

[144] VanderWeele, T. J., Li, S., Tsai, A. C., and Kawachi, I. (2016). Association between religious service attendance and lower suicide rates among us women. *JAMA psychiatry*, 73(8):845–851.

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

[146] Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.

[147] Xie, Y., Brand, J. E., and Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347.

[148] Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

[149] Yang, Y., Tokdar, S. T., et al. (2015). Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674.

[150] Young, J. G., Hernán, M. A., and Robins, J. M. (2014). Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic methods*, 3(1):1–19.

[151] Zaini, M. A., Lim, C., Low, W., and Harun, F. (2005). Effects of nutritional status on academic performance of malaysian primary school children. *Asia Pacific Journal of Public Health*, 17(2):81–87.

[152] Zha, H., He, X., Ding, C., Gu, M., and Simon, H. D. (2002). Spectral relaxation for k-means clustering. In *Advances in neural information processing systems*, pages 1057–1064.

[153] Zhang, W., Le, T. D., Liu, L., Zhou, Z.-H., and Li, J. (2017). Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 33(15):2372–2378.

[154] Zhang, Z., Chen, Z., Troendle, J. F., and Zhang, J. (2012). Causal inference on quantiles with an obstetric application. *Biometrics*, 68(3):697–706.

[155] Zheng, W. and Laan, M. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation 2010.

Appendix A

Supplementary Materials for Chapter 2

A.1 Algorithm

A.2 Empirical demonstration for Theorem 2.6.1

To empirically assess the above result in finite samples, we conduct two simple simulations under different setups; one in a randomized trial and the other in an observational study.

Simulation 1. (Randomized Trial) We set $p = 0.5$ in the simulation for both always-treated and never-treated units. We let $Y \mid \overline{A}_t \sim N\big(10 + \mid \overline{A}_t \mid_2, 1\big)$ truncated at ± two standard deviations. Given a value of δ, we generate datasets for $t = 1, ..., 50$, $n = 250$ for all t, and repeat the same simulation 100 times with the same data generation process. For positivity assumption to be valid, we always keep at least one always-treated or never-treated unit in each simulation. We compute the sample variance of each estimator and the relative efficiency. Figure A.1 shows the results along with the true lower bound on the relative efficiency given in Theorem 2.6.1 (the dotted line).

Simulation 2. (Observational Study) Although not directly covered by the setup from Theorem 2.6.1, it is also valuable to investigate the corresponding results in an observational study. To this end, we consider the following model

$$X_t = (X_{1,t}, X_{2,t}) \sim N(0, \mathbf{I})$$

$$\pi_t(H_t) = expit\Big(\mathbf{1}^\top X_t + 2 \sum_{s=t-2}^{t-1} (A_s - 1/2)\Big)$$

$$\big(Y \big| \overline{X}_t, \overline{A}_t\big) \sim N\big(\mu(\overline{X}_t, \overline{A}_t), 1\big)$$

Algorithm 4 Implementation of the proposed estimator (2.6)

Let δ be fixed and pick $t \leq T$. For each $k \in \{1, ..., K\}$, let $D_0 = \{Z_i : S_i \neq k\}$ and $D_1 = \{Z_i : S_i = k\}$ denote corresponding training and test data, respectively, and let $D = D_0 \bigcup D_1$.

1. For each time $t = 1, ..., t$ regress A_t on H_t using only observable samples at time t in D_0, then obtain predicted values $\widehat{\pi}_t(H_t)$ for only subject with $R_t = 1$ in D.

2. For each time $t = 1, ..., t$ regress R_{t+1} on (H_t, A_t) using only observable samples at time t in D_0, then obtain predicted values $\widehat{\omega}_t(H_t, A_t)$ for only subject with $R_t = 1$ in D.

3. For each time $t = 1, ..., t$, letting $W_s = \frac{\delta A_s + 1 - A_s}{\delta \widehat{\pi}_s(H_s) + 1 - \widehat{\pi}_s(H_s)} \cdot \frac{1}{\widehat{\omega}_s(H_s, A_s)}$ and construct following cumulative product weights for only subject with $R_{t+1} = 1$ in D_1:

   - $\widetilde{W}_t = \widehat{\omega}_t(H_t, A_t) \prod_{s=1}^{t} W_s$ for $1 \leq t < t$
   - $\widetilde{W}_t = \prod_{s=1}^{t} W_s$

4. For each time $t = t, t-1, ..., 1$, by setting $M_{t+1} = Y_t$:

   a. Regress M_{t+1} on (H_t, A_t) using only observable samples at time $t+1$ (i.e. only if $R_{t+1} = 1$) in D_0, then obtain predictions $\widehat{m}_t(H_t, 1)$ and $\widehat{m}_t(H_t, 0)$ for only subject with $R_t = 1$ in D.

   b. Construct pseudo-outcome $M_t = \frac{\widehat{m}_t(H_t,1)\delta\widehat{\pi}_t(H_t) + \widehat{m}_t(H_t,0)\{1-\widehat{\pi}_t(H_t)\}}{\delta\widehat{\pi}_t(H_t) + 1 - \widehat{\pi}_t(H_t)}$ for only subject with $R_t = 1$ in D.

5. Construct time-dependent weights $V_t = \frac{\{A_t - \widehat{\pi}_t(H_t)\}(1-\delta)}{\delta A_t + 1 - A_t}$ for only subject with $R_t = 1$ in D_1.

6. Compute $\sum_t \widetilde{W}_t V_t M_t + \widetilde{W}_t Y_t$ for only subject with $R_{t+1} = 1$ in D_1 and define $\widehat{\psi}_t^{(k)}(\delta)$ to be its average.

Output : $\widehat{\psi}_t(\delta) = \frac{1}{K} \sum_{k=1}^{K} \widehat{\psi}_t^{(k)}(\delta)$

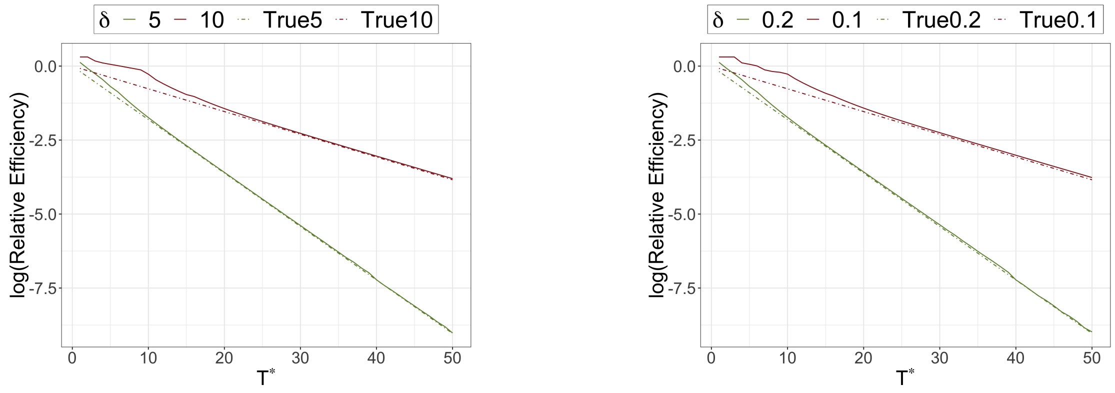


Fig. A.1 Relative efficiency curve in log-scale over time t for the case of always-treated unit where we use $\delta = 5, 10$ (Left) and for the case of never-treated unit where we use $\delta = 0.2, 0.1$ (Right). The true lower bound for each δ is represented as dotted line.

for all $t \leq T$ where we set $\mu(\overline{X}_t, \overline{A}_t) = 10 + A_t + A_{t-1} + |((\mathbf{1}^\top X_t + \mathbf{1}^\top X_{t-1})\mid$ and $\mathbf{1} = [1,1]^\top$. This simple simulation setup assumes that it is more (less) likely to receive a treatment if a subject has recently received (not received) treatments. The rest of the simulation specifications are the same as *Simulation 1*. The result is presented in Figure A.2.

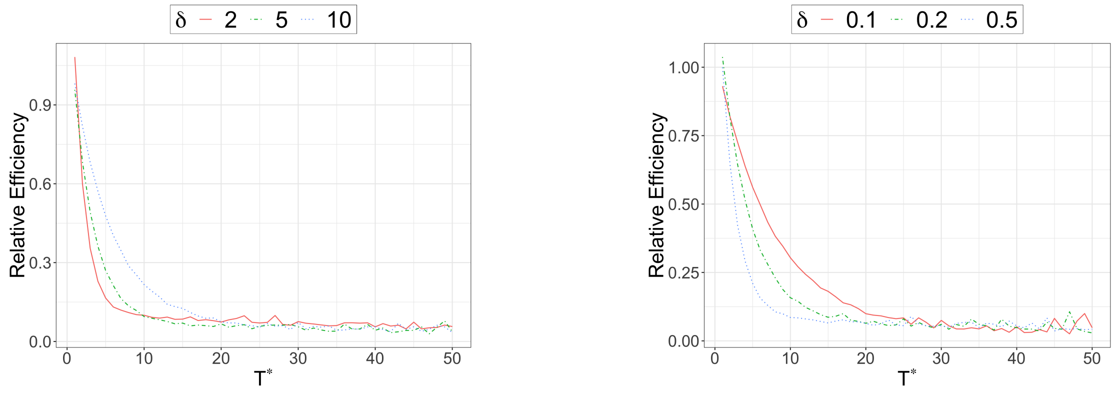


Fig. A.2 Relative efficiency curve over time t for the case of always-treated unit where we use $\delta = 2, 5, 10$ (Left) and for the case of never-treated unit where we use $\delta = 0.5, 0.2, 0.1$ (Right).

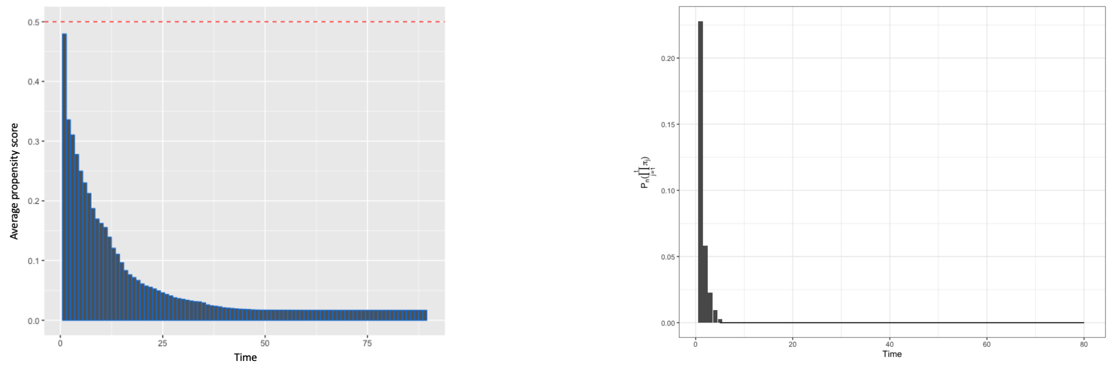
Overall, the simulation results support Theorem 2.6.1. Remarkably, even when we consider the setup for observational studies (the second simulation) we still observe almost exponential gains with incremental intervention effects.

A.3 Alternative approaches for the EAGeR data analysis

Here, we discuss why standard approaches might fail for our analysis of the EAGER dataset in Section 2.7.2 of the main text. Then, for the purpose of comparison, we alter our target effect and then apply some of other nonparametric approaches available in the literature. Then we compare the result with the one we obtained in Section 2.7.2.

A.3.1 Why standard model fails: positivity violation

All the standard models dealing with time-varying treatments, except on very rare occasions, require treatment positivity. However, as will be elaborated below, positivity is likely violated in the EAGER dataset. Many individuals turned out not to follow the given protocol of taking aspirin. This non-compliance only exacerbates over time. To illustrate this, we present the average propensity score over time in Figure A.3-(a). As shown in Figure A.3-(a), the average propensity score quickly drops to zero as t grows. As a result, at the end of the study it is almost impossible to find individuals who have been consistently taking aspirin at every timepoint. In other words, Figure A.3-(a) implies that it would be hard to imagine having all of the study participants take aspirin at each time.



(a)

(b)

Fig. A.3 (a) The average propensity score over the course of follow-up. We observe that due to the non-complinace, the average propensity score sharply decreases over time, which strongly hints at positivity violation in the EAGeR dataset. (b) $\mathbb{P}_n(\prod_{j=1}^{t}\widehat{\pi}_j)$ over the course of follow-up. When $t \geq 5$, $\mathbb{P}_n(\prod_{j=1}^{t}\widehat{\pi}_j)$ becomes less than 5×10^{-4}, which makes an IPW estimation in MSMs infeasible. We used Random Forests (via the `ranger` package in R) to estimate π_t.

Even if positivity is only nearly violated, it can pose a serious problem in attempting to estimate our target causal effect. One of the most widely-used approaches to handle time-varying treatments is marginal structural models (MSMs) [112]. In practice, MSMs are often estimated via inverse probability weighting (IPW). The following quantity appears in the IPW (also in the doubly robust) moment condition

$$h(\overline{A}_T)\left\{\frac{Y - m(\overline{A}_T;\beta)}{\prod_{t=1}^{T}\widehat{\pi}_t}\right\},$$

for any choice of h (with matching dimensions) where $\widehat{\pi}_t(a_t) = \widehat{\mathbb{P}}(A_t = a_t \mid H_t)$. However, Figure A.3-(b) indicates that on average a cumulative product of propensity score sharply drops to zero even with moderate t. This would make standard estimation techniques such as IPW to fail as $\mathbb{P}_n(\prod_{j=1}^{T}\widehat{\pi}_j)$ easily blows up.

Specifically, when we parametrically model the effect curve by $\mathbb{E}[Y^{\overline{a}_T}] = m(\overline{a}_T;\beta) = \beta_0 + \sum_{t=1}^{T}\beta_{1t}a_t$ so that the coefficient for exposure can vary with time, then an inverse-weighted MSM estimator which is the solution to

$$\mathbb{P}_n\left[h(\overline{A}_T)\left\{\frac{Y - m(\overline{A}_T;\beta)}{\prod_{t=1}^{T}\widehat{\pi}_t}\right\}\right] = 0$$

indeed fails and no coefficient estimates can be found even for moderate value of T, e.g. $T =\sim 10$. Thus, it appears that positivity violation in our dataset precludes the standard MSM-based approach. We remark that these limitations are not at all unique to the analysis of our EAGeR dataset, but instead are common to many observational MSM-based analyses as well as other recent approaches [e.g., 92].

A.3.2 Alternative approach

Due to the positivity violation, the estimation result, if any, via standard approaches will remain dubious at best. Therefore, we alter our target contrast from the standard ATE to the mean outcome we would have observed in a population if “observed" versus none (not all versus none) were treated, which is defined by

$$\tau_{\text{obs}}(T) \equiv \mathbb{E}\left[Y^{\bar{A}_T=\overline{a^{\text{obs}}},\bar{R}_T=\bar{\mathbf{1}}}\right] - \mathbb{E}\left[Y^{\bar{A}_T=\bar{\mathbf{0}},\bar{R}_T=\bar{\mathbf{1}}}\right], \tag{A.1}$$

where $\overline{a^{\text{obs}}}$ denotes an observed history of aspirin consumption. This new estimand would tell us how the mean outcome would have changed if no one in the population had taken aspirin throughout the study and we can avoid estimating the problematic counterfactual $\mathbb{E}\left[Y^{\bar{A}_T=\bar{\mathbf{1}},\bar{R}_T=\bar{\mathbf{1}}}\right]$. However, by construction this solution entails the fundamental limitation as we have sacrificed the causal effect of original interest.

We use the g-computation [1] (plug-in) estimator [108] and the sequential doubly robust (SDR) estimator proposed by Luedtke et al. [92] which also allows right-censored data structures.

A.3.3 Estimation and inference

Estimation. First for the g-computation estimator, we estimate the following g-formula

$$\mathbb{E}\left[Y^{\bar{A}_T=\bar{a}_T,\bar{R}_T=\bar{\mathbf{0}}}\right]=\int\cdots\int\mathbb{E}\left[Y|\overline{X}_T,\overline{A}_T=\overline{a}_T,\bar{R}_T=\bar{\mathbf{1}}_T\right]\prod_{t=2}^{T}d\mathbb{P}(X_t|\overline{X}_{t-1},\overline{A}_{t-1}=\overline{a}_{t-1},\overline{R}_{t-1}=\overline{\mathbf{1}}_{t-1})$$
$$\times\, d\mathbb{P}(X_1,A_1=a_1,R_1=1)$$

via plug in estimators of the pseudo-outcome regression function each time step. Next, for the SDR estimator, we tailor Algorithm 2 of Luedtke et al. [92] for our right-censored data structures (everything remains the same except that we add the condition $\overline{R}_{t-1}=\overline{\mathbf{1}}_{t-1}$ on each pseudo-outcome regression function). For both methods, we use the same nonparametric ensemble we used in Section 2.7.2 of the main text as our regression model.

Inference. Confidence intervals are estimated by bootstrapping at 95% level for both of the estimators. Note that for the SDR estimator, we are guaranteed to consistently estimate standard errors (pointwisely) by bootstrapping due to the following asymptotically property,

$$\sqrt{n}(\widehat{\tau}_{\text{obs}}(t)-\tau_{\text{obs}}(t))\rightsquigarrow\mathcal{N}(0,Var(\phi_\tau(t)))$$

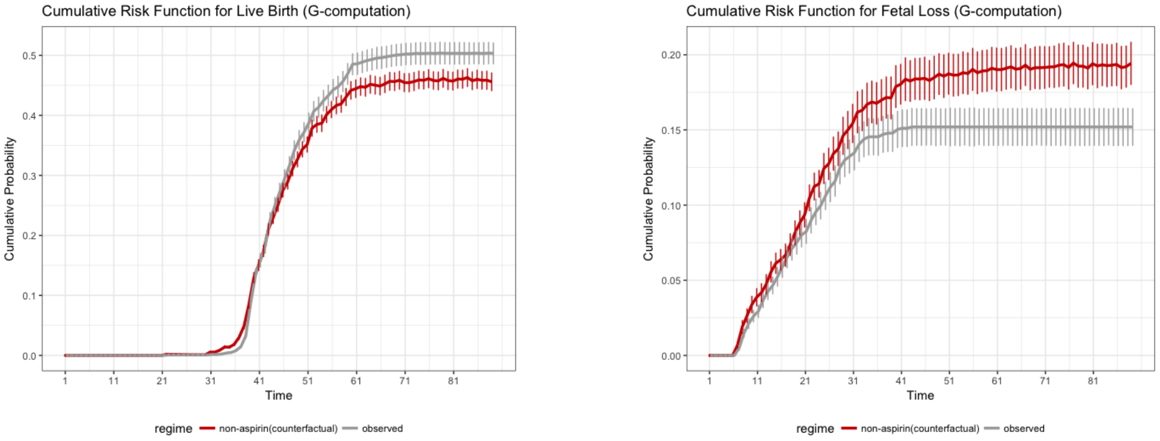
for all $t\leq T$, where $\phi_\tau(t)$ is the influence function of $\widehat{\tau}_{\text{obs}}(t)$. However, this is no longer guaranteed for the g-computation estimator.

A.3.4 Result

For the sake of completeness, we estimate each $\tau_{\text{obs}}(t)$ for all $t=2\sim 89$ and present the cumulative effects over time t. The results for the g-computation and the SDR estimators are presented in Figure A.4, A.5, respectively.

The result based on the g-computation estimator in Figure A.4 shows that the counterfactual mean outcomes for never-takers (individuals who have never taken aspirin throughout the study) are worse-off than the observed. Specifically, for the never-takers the probability of having live birth has been decreased and the probability of having fetal loss has increased. The result seems to be statistically significant at $T=89$.

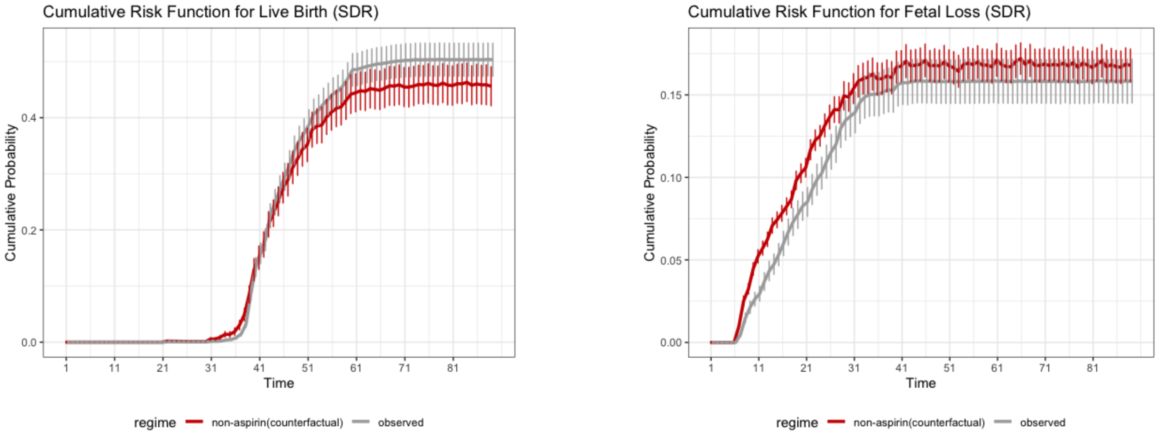
[1] We also tried a weighting estimator but omitted the result here, since it gives almost the same result with wider confidence band.



Live birth

Pregnancy loss

Fig. A.4 Cumulative risk curve for live birth and pregnancy loss via the regression based g-computation estimator. Pointwise 95% confidence interval is estimated by bootstrapping with 1000 resampling.



Live birth

Pregnancy loss

Fig. A.5 Cumulative risk curve for live birth and pregnancy loss via the sequential doubly robust (SDR) estimator. Pointwise 95% confidence interval is estimated by bootstrapping with 1000 resampling.

On the other hand, the result based on the SDR estimator in Figure A.5 indicates that although the mean effects for the never-takers still appear to be worse off than the observed, they look no longer statistically significant. Hence in this case we cannot draw any firm conclusion about the effect of aspirin on pregnancy outcome.

It might be tempting to take the results from Figure A.4 as it seems to deliver more clear messages. However, we do not know if our variance estimates there are correct.

Also, particularly considering the sample size (n=1024), we are likely to suffer from slow estimation rates of our regression model. These issues can be mitigated in doubly robust estimators as in the SDR estimator. Thus, we should rather resort to the results presented in Figure A.5, which basically tells us that the effect of low-dose aspirin is insignificant and remains dubious, at the very least, based on the causal effect defined in (A.1).

After all, it should be noted that due to the positivity violation we end up limiting ourselves to the more narrow notion of causal effects (i.e. observed versus none) which is different from the ATE type estimands that are typically of utmost interest for policy makers. The causal effect in (A.1) might not be practically meaningful as to aspirin prescription for pregnant since we are in general much more interested in the always-taker group than the never-taker group.

A.4 Technical Results and Proofs

A.4.1 Lemma for the identifying expression in Theorem 2.3.1

To identify our target parameter $\psi_t(\delta) = \mathbb{E}\left(Y_t^{\overline{Q}_t(\delta)}\right)$, we need the following lemma.

Lemma A.4.1. *Under (A2-M) and (A3), and for all $t \leq T$, we have following equvalence properties:*

a. $d\mathbb{P}(A_t|H_t) = d\mathbb{P}(A_t|H_t, R_t = 1)$

b. $d\mathbb{P}(X_t|A_{t-1}, H_{t-1}) = d\mathbb{P}(X_t|A_{t-1}, H_{t-1}, R_t = 1)$

c. $\mathbb{E}[Y|\overline{X}_t, \overline{A}_t] = \mathbb{E}[Y|\overline{X}_t, \overline{A}_t, R_{t+1} = 1]$

Lemma A.4.1 thus shows that the above important quantities conditional on the observed data are equivalent to corresponding quantities conditioned on the full data. In the identifying expression we can only use quantities directly estimated from observed history, so the above equivalence relations play a key role.

Proof. Proof is done based on induction. We proceed one by one as follows.

- $\boldsymbol{d\mathbb{P}(A_t|H_t) = d\mathbb{P}(A_t|H_t, R_t = 1)}$

First note that

$$\begin{aligned} d\mathbb{P}(A_t,H_t) &= d\mathbb{P}(\overline{X}_t,\overline{A}_t) = d\mathbb{P}(\underline{X}_2,\underline{A}_2 \mid X_1,A_1)d\mathbb{P}(X_1,A_1) \\ &= d\mathbb{P}(\underline{X}_2,\underline{A}_2 \mid X_1,A_1,R_2=1)d\mathbb{P}(X_1,A_1,R_1=1) \\ &= d\mathbb{P}(\underline{X}_3,\underline{A}_3 \mid \overline{X}_2,\overline{A}_2,R_2=1)\frac{d\mathbb{P}(X_1,A_1,R_1=1)}{d\mathbb{P}(X_1,A_1,R_2=1)}d\mathbb{P}(\overline{X}_2,\overline{A}_2,R_2=1) \\ &= d\mathbb{P}(\underline{X}_3,\underline{A}_3 \mid \overline{X}_2,\overline{A}_2,R_3=1)\frac{d\mathbb{P}(X_1,A_1,R_1=1)}{d\mathbb{P}(X_1,A_1,R_2=1)}d\mathbb{P}(\overline{X}_2,\overline{A}_2,R_2=1) \\ &= d\mathbb{P}(X_t,A_t \mid \overline{X}_{t-1},\overline{A}_{t-1},R_t=1)\prod_{s=1}^{t-2}\frac{d\mathbb{P}(\overline{X}_s,\overline{A}_s,R_s=1)}{d\mathbb{P}(\overline{X}_s,\overline{A}_s,R_{s+1}=1)}d\mathbb{P}(\overline{X}_{t-1},\overline{A}_{t-1},R_{t-1}=1) \\ &= d\mathbb{P}(\overline{X}_t,\overline{A}_t,R_t=1)\prod_{s=1}^{t-1}\frac{d\mathbb{P}(\overline{X}_s,\overline{A}_s,R_s=1)}{d\mathbb{P}(\overline{X}_s,\overline{A}_s,R_{s+1}=1)} \end{aligned}$$

where the first equality follows by definition, the second by definition of conditional probability, the third by assumption (A2-M), the fourth again by definition of conditional probability, the fifth by assumption (A2-M), and the sixth by repeating the same step $t-1$ times. The last expression is obtained by simply rearranging terms using the definition of conditional probability.

Now introduce the following shorthand notation:

$$\boldsymbol{\Pi}_{\mathbb{P}}(t-1) \equiv \prod_{s=1}^{t-1}\frac{d\mathbb{P}(\overline{X}_s,\overline{A}_s,R_s=1)}{d\mathbb{P}(\overline{X}_s,\overline{A}_s,R_{s+1}=1)}$$

so we can write $d\mathbb{P}(A_t,H_t) = d\mathbb{P}(\overline{X}_t,\overline{A}_t,R_t=1)\boldsymbol{\Pi}_{\mathbb{P}}(t-1)$.

Then, similarly we have

$$d\mathbb{P}(H_t) = d\mathbb{P}(\overline{X}_t,\overline{A}_{t-1}) = d\mathbb{P}(\overline{X}_t,\overline{A}_{t-1},R_t=1)\boldsymbol{\Pi}_{\mathbb{P}}(t-1).$$

Hence, finally we obtain

$$\begin{aligned} d\mathbb{P}(A_t \mid H_t) &= \frac{d\mathbb{P}(A_t,H_t)}{d\mathbb{P}(H_t)} = \frac{d\mathbb{P}(\overline{X}_t,\overline{A}_t,R_t=1)}{d\mathbb{P}(\overline{X}_t,\overline{A}_{t-1},R_t=1)} \\ &= \frac{d\mathbb{P}(A_t,H_t,R_t=1)}{d\mathbb{P}(H_t,R_t=1)} \\ &= d\mathbb{P}(A_t|H_t,R_t=1) \end{aligned}$$

where the second equality comes from the above results. The proof naturally leads to subsequent result of $\boldsymbol{dQ_t(A_t|H_t) = dQ_t(A_t|H_t, R_t = 1)}$.

- $\boldsymbol{d\mathbb{P}(X_t|A_{t-1}, H_{t-1}) = d\mathbb{P}(X_t|A_{t-1}, H_{t-1}, R_t = 1)}$

By definition $d\mathbb{P}(X_t|A_{t-1}, H_{t-1}) = d\mathbb{P}(H_t)/d\mathbb{P}(A_{t-1}, H_{t-1})$, and from previous part it immediately follows

$$d\mathbb{P}(H_t) = d\mathbb{P}(\overline{X}_t, \overline{A}_{t-1}, R_t = 1)\boldsymbol{\Pi}_{\mathbb{P}}(t-1),$$
$$d\mathbb{P}(A_{t-1}, H_{t-1}) = d\mathbb{P}(\overline{X}_{t-1}, \overline{A}_{t-1}, R_{t-1} = 1)\boldsymbol{\Pi}_{\mathbb{P}}(t-2).$$

Hence, we have

$$\begin{aligned}\frac{d\mathbb{P}(H_t)}{d\mathbb{P}(A_{t-1}, H_{t-1})} &= \frac{d\mathbb{P}(\overline{X}_t, \overline{A}_{t-1}, R_t = 1)}{d\mathbb{P}(\overline{X}_{t-1}, \overline{A}_{t-1}, R_t = 1)} \\ &= d\mathbb{P}(X_t \mid \overline{H}_{t-1}, \overline{A}_{t-1}, R_t = 1)\end{aligned}$$

which yields the desired result.

- $\boldsymbol{\mathbb{E}[Y|\overline{X}_t, \overline{A}_t] = \mathbb{E}[Y|\overline{X}_t, \overline{A}_t, R_{t+1} = 1]}$

By definition $\mathbb{E}[Y|\overline{X}_t, \overline{A}_t] = \int y d\mathbb{P}(y|\overline{X}_t, \overline{A}_t)$, and thereby it suffices to show that $d\mathbb{P}(Y|\overline{X}_t, \overline{A}_t) = d\mathbb{P}(Y|\overline{X}_t, \overline{A}_t, R_{t+1})$.

By the same logic we use for the first proof, we have

$$d\mathbb{P}(Y, \overline{X}_t, \overline{A}_t) = d\mathbb{P}(Y, \overline{X}_t, \overline{A}_t, R_t = 1)\boldsymbol{\Pi}_{\mathbb{P}}(t-1)$$

and also

$$d\mathbb{P}(\overline{X}_t, \overline{A}_t) = d\mathbb{P}(\overline{X}_t, \overline{A}_t, R_t = 1)\boldsymbol{\Pi}_{\mathbb{P}}(t-1).$$

Thus it follows by what are shown above displays together with assumption (A2-M) that

$$d\mathbb{P}(Y \mid \overline{X}_t, \overline{A}_t) = d\mathbb{P}(Y \mid \overline{X}_t, \overline{A}_t, R_t = 1) = d\mathbb{P}(Y \mid \overline{X}_t, \overline{A}_t, R_{t+1} = 1).$$

Hence, we have shown that all the identities hold. □

A.4.2 Proof of Theorem 2.4.1

Identifying expression for the efficient influence function

In the next lemma, we provide an identifying expression for the efficient influence function for our incremental effect $\psi_t(\delta)$ under a nonparametric model, which allows the data-generating process $\mathbb{P}$ to be infinite-dimensional.

Lemma A.4.2. *Define*

$$
\begin{aligned}
& m_s(h_s,a_s,R_{s+1}=1) \\
& \quad = \int_{\mathscr{R}_s} \mu(h_t,a_t,R_{t+1}=1) \prod_{k=s+1}^{t} dQ_k(a_k \mid h_k,R_k=1)\, d\mathbb{P}(x_k|h_{k-1},a_{k-1},R_k=1)
\end{aligned}
$$

for $s=0,...,t-1$, $\forall t \leq T$, where we write $\mathscr{R}_s = (\overline{\mathscr{X}}_t \times \overline{\mathscr{A}}_t) \setminus (\overline{\mathscr{X}}_s \times \overline{\mathscr{A}}_s)$ and $\mu(h_t,a_t,R_{t+1}=1) = \mathbb{E}(Y_t \mid H_t=h_t, A_t=a_t, R_{t+1}=1)$. For $s=t$ and $s=t+1$, we set $m_s(\cdot)=\mu(h_t,a_t,R_{t+1}=1)$ and $m_{t+1}(\cdot)=Y$. Moreover, let $\frac{\mathbb{1}(H_s=h_s,R_s=1)}{d\mathbb{P}(h_s,R_s=1)}\phi_s(H_s,A_s,R_s=1;a_s)$ denote the efficient influence function for $dQ_s(a_s|h_s,R_s=1)$.

Then, the efficient influence function for $m_0=\psi_t(\delta)$ is given by

$$
\begin{aligned}
& \sum_{s=0}^{t} \left\{ \int_{\mathscr{A}_{s+1}} m_{s+1}(H_{s+1},A_{s+1},R_{s+2}=1)\, dQ_{s+1}(a_{s+1}|H_{s+1},R_{s+1}=1) - m_s(H_s,A_s,R_{s+1}=1) \right\} \\
& \qquad \times \mathbb{1}(R_{s+1}=1) \left(\prod_{k=0}^{s} \frac{dQ_k(A_k \mid H_k,R_k=1)}{d\mathbb{P}(A_k \mid H_k,R_k=1)} \frac{1}{d\mathbb{P}(R_{k+1}=1 \mid H_k,A_k,R_k=1)} \right) \\
& + \sum_{s=1}^{t} \mathbb{1}(R_s=1) \left(\prod_{k=0}^{s-1} \frac{dQ_k(A_k \mid H_k,R_k=1)}{d\mathbb{P}(A_k \mid H_k,R_k=1)} \frac{1}{d\mathbb{P}(R_{k+1}=1 \mid H_k,A_k,R_k=1)} \right) \\
& \qquad \times \int_{\mathscr{A}_s} m_s(H_s,a_s,R_{s+1}=1)\, \phi_s(H_s,A_s,R_s=1;a_s)\, d\nu(a_s)
\end{aligned}
$$

where we define $dQ_{t+1}=1$, $m_{t+1}(\cdot)=Y$, and $dQ_0(a_0|h_0)/d\mathbb{P}(a_0|h_0)=1$, and ν is a dominating measure for the distribution of A_s.

The proof of Lemma A.4.2 involves derivation of efficient influence function for general stochastic interventions that depend on the both observational propensity scores and right-censoring process. In the proof, we delineate how we can apply chain rule arguments to derive efficient influence functions for complicated functionals from much simpler functional forms. We further simplify and render the above efficient influence function to estimable form in next theorem.

The basic proof structure follows the work of [67]. We begin by presenting the following three additional lemmas to prove Lemma A.4.2.

Lemma A.4.3. *For $\forall t$, the efficient influence function for*

$$dQ_t(a_t \mid h_t, R_t = 1) = \frac{a_t \delta \pi_t(h_t) + (1-a_t)\{1-\pi_t(h_t)\}}{\delta \pi_t(h_t) + 1 - \pi_t(h_t)}$$

which is defined in (2.2) is given by $\frac{\mathbb{1}(H_t=h_t, R_t=1)}{d\mathbb{P}(h_t, R_t=1)}\phi_t(H_t, A_t, R_t = 1; a_t)$, where $\phi_t(H_t, A_t, R_t = 1; a_t)$ equals

$$\frac{(2a_t - 1)\delta\{A_t - \pi_t(H_t)\}}{(\delta\pi_t(H_t) + 1 - \pi_t(H_t))^2}$$

where $\pi_t(h_t) = \mathbb{P}(A_t = 1 \mid H_t = h_t, R_t = 1)$.

Lemma A.4.4. *Suppose $\overline{Q}_T$ is not depending on $\mathbb{P}$. Recall that for $\forall t \le T$,*

$$m_s(h_s, a_s, R_{s+1} = 1) \quad = \int_{\mathscr{R}_s} \mu(h_t, a_t, R_{t+1} = 1) \prod_{k=s+1}^{t} dQ_k(a_k \mid h_k, R_k = 1) d\mathbb{P}(x_k | h_{k-1}, a_{k-1}, R_k = 1)$$

for $s = 0, ..., t-1$, where we write $\mathscr{R}_s = (\overline{\mathscr{X}}_t \times \overline{\mathscr{A}}_t) \setminus (\overline{\mathscr{X}}_s \times \overline{\mathscr{A}}_s)$ and $\mu(h_t, a_t, R_{t+1} = 1) = \mathbb{E}(Y_t \mid H_t = h_t, A_t = a_t, R_{t+1} = 1)$. Note that from definition of m_s it immeidately follows $m_s = \int_{\mathscr{X}_s \times \mathscr{A}_s} m_{s+1} dQ_{s+1}(a_{s+1} \mid h_{s+1}, R_{s+1} = 1) d\mathbb{P}(x_{s+1} | h_s, a_s, R_{s+1} = 1)$.

Now the efficient influence function for $\psi^*(\overline{Q}_t) = m_0$ is

$$\sum_{s=0}^{t} \left\{ \int_{\mathscr{A}_{s+1}} m_{s+1}(H_{s+1}, A_{s+1}, R_{s+2} = 1) dQ_{s+1}(a_{s+1} | H_{s+1}, R_{s+1} = 1) - m_s(H_s, A_s, R_{s+1} = 1) \right\}$$
$$\times \mathbb{1}(R_{s+1} = 1) \left(\prod_{k=0}^{s} \frac{dQ_k(A_k \mid H_k, R_k = 1)}{d\mathbb{P}(A_k \mid H_k, R_k = 1)} \frac{1}{d\mathbb{P}(R_{k+1} = 1 \mid H_k, A_k, R_k = 1)} \right)$$

where we define $dQ_{t+1} = 1$, $m_{t+1}(\cdot) = Y_t$, and $dQ_0(a_0|h_0)/d\mathbb{P}(a_0|h_0) = 1$.

Lemma A.4.5. *Suppose $\overline{Q}_T$ depends on $\mathbb{P}$ and let $\frac{\mathbb{1}(H_t=h_t, R_t=1)}{d\mathbb{P}(h_t, R_t=1)}\phi_t(H_t, A_t, R_t = 1; a_t)$ denote the efficient influence function for $dQ_t(a_t|h_t, R_t = 1)$ defined in Lemma A.4.3 for all t. Then the efficient influence function for $\psi_t(\delta)$ is given as*

$$\varphi^*(\overline{Q}_t)$$
$$+ \sum_{s=1}^{t} \mathbb{1}(R_s = 1) \left(\prod_{k=0}^{s-1} \frac{dQ_k(A_k \mid H_k, R_k = 1)}{d\mathbb{P}(A_k \mid H_k, R_k = 1)} \frac{1}{d\mathbb{P}(R_{k+1} = 1 \mid H_k, A_k, R_k = 1)} \right)$$
$$\times \int_{\mathscr{A}_s} m_s(H_s, a_s, R_{s+1} = 1) \phi_s(H_s, A_s, R_s = 1; a_s) d\nu(a_s)$$

where $\varphi^*(\overline{Q}_t)$ is the efficient influence function from Lemma A.4.4 and ν is a dominating measure for the distribution of A_s.

The proof of Lemma A.4.3, A.4.4 and A.4.5 are basically results of a series of chain rules, after specifying efficient influence functions for terms that commonly appear. The full proofs are not particularly illuminating considering its length. Thus we omit a proof of Lemma A.4.3 and only include a brief sketch for proofs of Lemma A.4.4 and A.4.5 below, which can be useful to develop results for more general stochastic interventions.

Proof of Lemma A.4.4 and Lemma A.4.5

Let $\mathscr{IF}:\psi\to\phi$ denote a map to the efficient influence function ϕ for a functional ψ. First without proof, we specify efficient influence functions for mean and conditional mean which serve two basic ingredients for our proof. For mean value of a random variable Z, we have

$$\mathscr{IF}\big(\mathbb{E}[Z]\big)=Z-\mathbb{E}[Z],$$

and for conditional mean with a pair of random variables $(X,Y)\sim\mathbb{P}$ where X is discrete, we have

$$\mathscr{IF}\big(\mathbb{E}[Y|X=x]\big)=\frac{\mathbb{1}(X=x)}{\mathbb{P}(X=x)}\Big\{Y-\mathbb{E}[Y\mid X=x]\Big\}.$$

These results can be directly obtained from either (2.5) or (**??**) in section 2.4.

Proof. It is sufficient to prove for the case $t=2$ since it is straightforward to extend the proof for general $t\le T$ by induction. For $t=2$, it is enough to compute the following four terms.

$$\begin{aligned}
\text{A)}\ &\int_{\mathscr{H}_2\times\mathscr{A}_2}\mathscr{IF}\Big(\mu(h_2,a_2,R_3=1)\Big)\prod_{s=1}^{2}dQ_s(a_s\mid h_s,R_s=1)d\mathbb{P}(x_s|h_{s-1},a_{s-1},R_s=1)\\
&=\int_{\mathscr{H}_2\times\mathscr{A}_2}\frac{\mathbb{1}\{(H_2,A_2,R_3)=(h_2,a_2,1)\}}{d\mathbb{P}(h_2,a_2,R_3=1)}\Big\{Y-\mu(h_2,a_2,R_3=1)\Big\}\\
&\qquad\times\prod_{s=1}^{2}dQ_s(a_s\mid h_s,R_s=1)d\mathbb{P}(x_s|h_{s-1},a_{s-1},R_s=1)\\
&=\int_{\mathscr{H}_2\times\mathscr{A}_2}\mathbb{1}\big\{(H_2,A_2,R_3)=(h_2,a_2,1)\big\}\big\{Y-\mu(h_2,a_2,R_3=1)\big\}\\
&\qquad\times\prod_{s=1}^{2}\frac{dQ_s(a_s\mid h_s,R_s=1)}{d\mathbb{P}(a_s\mid h_s,R_s=1)}\frac{1}{d\mathbb{P}(R_{s+1}=1\mid h_s,a_s,R_s=1)}\\
&=\{Y-\mu(H_2,A_2,R_3=1)\}\mathbb{1}(R_3=1)\prod_{s=1}^{2}\frac{dQ_t(A_s\mid H_s,R_s=1)}{d\mathbb{P}(A_s\mid H_s,R_s=1)}\frac{1}{d\mathbb{P}(R_{s+1}=1\mid H_s,A_s,R_s=1)}
\end{aligned}$$

B)
$$
\begin{aligned}
&\int_{\mathscr{H}_2\times\mathscr{A}_2}\mu(h_2,a_2,R_3=1)\mathscr{I}\mathscr{F}\Big(d\mathbb{P}(x_2|h_1,a_1,R_2=1)\Big)d\mathbb{P}(h_1)\prod_{s=1}^{2}dQ_s(a_s\mid h_s,R_s=1)\\
&=\int_{\mathscr{H}_2\times\mathscr{A}_2}\mu(h_2,a_2,R_3=1)\frac{\mathbb{1}\big\{(H_1,A_1,R_2)=(h_1,a_1,1)\big\}}{d\mathbb{P}(h_1,a_1,R_2=1)}\Big\{\mathbb{1}(X_2=x_2)-d\mathbb{P}(x_2|h_1,a_1,R_2=1)\Big\}\\
&\qquad\times d\mathbb{P}(h_1)\prod_{s=1}^{2}dQ_s(a_s\mid h_s,R_s=1)\\
&=\int_{\mathscr{H}_2\times\mathscr{A}_2}\mu(h_2,a_2,R_3=1)\frac{\mathbb{1}\big\{(H_1,A_1,R_2)=(h_1,a_1,1)\big\}\big\{\mathbb{1}(X_2=x_2)-d\mathbb{P}(x_2|h_1,a_1,R_2=1)\big\}}{d\mathbb{P}(R_2=1|h_1,a_1)d\mathbb{P}(a_1|h_1)d\mathbb{P}(h_1)}\\
&\qquad\times d\mathbb{P}(h_1)\prod_{s=1}^{2}dQ_s(a_s\mid h_s,R_s=1)\\
&=\int_{\mathscr{H}_2\times\mathscr{A}_2}\mu(h_2,a_2,R_3=1)dQ_2(a_2\mid h_2,R_2=1)\mathbb{1}\big\{(H_1,A_1,R_2)=(h_1,a_1,1)\big\}\\
&\qquad\times\big\{\mathbb{1}(X_2=x_2)-d\mathbb{P}(x_2|h_1,a_1,R_2=1)\big\}\frac{dQ_1(A_1\mid H_1)}{d\mathbb{P}(A_1\mid H_1)}\frac{1}{d\mathbb{P}(R_2=1\mid H_1,A_1)}\\
&=\Bigg\{\int_{\mathscr{H}_2\times\mathscr{A}_2\backslash\mathscr{H}_2}\mu(H_2,a_2,R_3=1)dQ_2(a_2\mid H_2,R_2=1)\\
&\qquad-\int_{\mathscr{H}_2\times\mathscr{A}_2\backslash\mathscr{H}_1\times\mathscr{A}_1}\mu(h_2,a_2,R_3=1)dQ_2(a_2\mid h_2,R_2=1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)\Bigg\}\\
&\qquad\times\mathbb{1}(R_2=1)\frac{dQ_1(A_1\mid H_1)}{d\mathbb{P}(A_1\mid H_1)}\frac{1}{d\mathbb{P}(R_2=1\mid H_1,A_1)}\\
&=\Bigg\{\int_{\mathscr{A}_2}\mu(H_2,a_2,R_3=1)dQ_2(a_2\mid H_2,R_2=1)-m_1(h_1,a_1,R_2=1)\Bigg\}\\
&\qquad\times\mathbb{1}(R_2=1)\frac{dQ_1(A_1\mid H_1)}{d\mathbb{P}(A_1\mid H_1)}\frac{1}{d\mathbb{P}(R_2=1\mid H_1,A_1)}
\end{aligned}
$$

C)
$$
\begin{aligned}
&\int_{\mathscr{H}_2\times\mathscr{A}_2}\mu(h_2,a_2,R_3=1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)\mathscr{I}\mathscr{F}\Big(d\mathbb{P}(h_1)\Big)\prod_{s=1}^{2}dQ_s(a_s\mid h_s,R_s=1)\\
&=\int_{\mathscr{H}_2\times\mathscr{A}_2}\mu(h_2,a_2,R_3=1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)\big\{\mathbb{1}(X_1=x_1)-d\mathbb{P}(x_1)\big\}\prod_{s=1}^{2}dQ_s(a_s\mid h_s,R_s=1)\\
&=\int_{\mathscr{H}_2\times\mathscr{A}_2\backslash\mathscr{H}_1}\mu(h_2,a_2,R_3=1)dQ_2(a_2\mid h_2,R_2=1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)dQ_1(a_1|h_1)-m_0\\
&=\int_{\mathscr{A}_1}m_1(h_1,a_1,R_2=1)dQ_1(a_1|h_1)-m_0
\end{aligned}
$$

D) Let ϕ_t denote the efficient influence function for $dQ_t(a_t|h_t, R_t = 1)$ as given in Lemma A.4.3. Now we have

$$\int_{\mathscr{H}_2\times\mathscr{A}_2} \mu(h_2,a_2,R_3=1)d\mathbb{P}(h_1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)\mathscr{I}\mathscr{F}\Big(dQ_1(a_1|h_1)dQ_2(a_2\mid h_2,R_2=1)\Big)$$

$$= \int_{\mathscr{H}_2\times\mathscr{A}_2} \mu(h_2,a_2,R_3=1)d\mathbb{P}(h_1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)\frac{\mathbb{1}\{(H_2,R_2)=(h_2,1)\}}{d\mathbb{P}(h_2,R_2=1)}\phi_2 dQ_1(a_1|h_1)$$

$$+ \int_{\mathscr{H}_2\times\mathscr{A}_2} \mu(h_2,a_2,R_3=1)d\mathbb{P}(h_1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)\frac{\mathbb{1}\{(H_1=h_1)\}}{d\mathbb{P}(h_1)}\phi_1 dQ_2(a_2\mid h_2,R_2=1)$$

$$= \int_{\mathscr{H}_2\times\mathscr{A}_2} \mu(h_2,a_2,R_3=1)\frac{\mathbb{1}\{(H_2,R_2)=(h_2,1)\}d\mathbb{P}(h_1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)dQ_1(a_1|h_1)}{d\mathbb{P}(x_2|h_1,a_1,R_2=1)d\mathbb{P}(R_2=1|h_1,a_1)d\mathbb{P}(a_1|h_1)d\mathbb{P}(h_1)}\phi_2$$

$$+ \int_{\mathscr{H}_2\times\mathscr{A}_2} \mu(h_2,a_2,R_3=1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)\mathbb{1}\{(H_1=h_1)\}\phi_1 dQ_2(a_2\mid h_2,R_2=1)$$

$$= \int_{\mathscr{H}_2\times\mathscr{A}_2\backslash\mathscr{H}_2} \mu(H_2,a_2,R_3=1)\mathbb{1}(R_2=1)\phi_2\frac{dQ_1(A_1\mid H_1)}{d\mathbb{P}(A_1\mid H_1)}\frac{1}{d\mathbb{P}(R_2=1\mid H_1,A_1)}$$

$$+ \int_{\mathscr{H}_2\times\mathscr{A}_2\backslash\mathscr{H}_1} \mu(h_2,a_2,R_3=1)dQ_2(a_2\mid h_2,R_2=1)d\mathbb{P}(x_2|h_1,a_1,R_2=1)\phi_1$$

$$= \left\{\frac{dQ_1(A_1\mid H_1)}{d\mathbb{P}(A_1\mid H_1)}\frac{1}{d\mathbb{P}(R_2=1\mid H_1,A_1)}\right\}\int_{\mathscr{A}_2}\mu(H_2,a_2,R_3=1)\phi_2 d\nu(a_2)\mathbb{1}(R_2=1)$$

$$+ \int_{\mathscr{A}_1} m_1(h_1,a_1,R_2=1)\phi_1 d\nu(a_1)$$

Note that we have set $dQ_0(a_0|h_0)/d\mathbb{P}(a_0|h_0) = 1$, and that we have $d\mathbb{P}(R_1 = 1) = 1$ and $\mathbb{1}(R_1 = 1) = 1$ by construction. Hence, putting part A), B), and C) together proves Lemma A.4.4 and part D) proves Lemma A.4.5. □

Preparation for Algebra: Some Lemmas

Next, we convert the identifying expression in Lemma A.4.2 into something we can estimate from observed data. To this end, we first present two Identity equations on the pseudo regression functions m_t defined in Lemma A.4.2 in the following lemma.

Lemma A.4.6. *Given m_t defined in Lemma A.4.2 for $\forall t \le T$ we have the following identities.*

a. $\mathbb{1}(R_{t+1}=1)m_t(H_t,A_t,R_{t+1}=1) = m_t(H_t,A_t,R_{t+1}=1)$

b. $\left(\frac{\mathbb{1}(R_{t+1}=1)}{d\mathbb{P}(R_{t+1}=1|H_t,A_t,R_t=1)}\right)m_t(H_t,A_t,R_{t+1}=1) = \mathbb{1}(R_{t+1}=1)m_t(H_t,A_t,R_{t+1}=1)$

Proof. First, note that from Remark 5,

$$m_t(H_t,A_t,R_{t+1}=1)$$

$$= \mathbb{E}\left[\frac{m_t(H_{t+1},a_{t+1},1)\delta\pi_{t+1}(H_{t+1}) + \{1-m_t(H_{t+1},0,1)\}\{1-\pi_{t+1}(H_{t+1})\}}{\delta\pi_{t+1}(H_{t+1})+1-\pi_{t+1}(H_{t+1})}\middle| H_t,A_t,R_{t+1}=1\right]$$

where we use shorthand notation $m_t(H_{t+1}, a_{t+1}, 1) = m_t(H_{t+1}, A_{t+1} = a_{t+1}, R_{t+2} = 1)$. In this proof, let $(m \cdot dQ)_{t+1}$ denote $\frac{m_t(H_{t+1},a_{t+1},1)\delta\pi_{t+1}(H_{t+1})+\{1-m_t(H_{t+1},0,1)\}\{1-\pi_{t+1}(H_{t+1})\}}{\delta\pi_{t+1}(H_{t+1})+1-\pi_{t+1}(H_{t+1})}$ which is the quotient inside above conditional expectation.

The identity in *part a* immediately follows from the definition of m_t.

For the identity in *part b*, we first note that by assumption (A2-M) it follows $d\mathbb{P}(x_s|h_{s-1}, a_{s-1}, R_s = 1) = d\mathbb{P}(x_s|h_{s-1}, a_{s-1}, R_{s-1} = 1)$ for every $s > 1$. Thus, we can write

$$m_t = \mathbb{E}\left[(m \cdot dQ)_{t+1} \middle| H_t, A_t, R_t = 1\right]$$

based on the definition of m_t. Now define another shorthand notation $h_{t+1}^{A_t,H_t} := (x_{t+1}, A_t, H_t)$ and $R_{t+1}^{R_t=1} := (R_{t+1}, R_t = 1)$. Then it follows that

$$\begin{aligned}
& m_t(H_t, A_t, R_{t+1} = 1) \\
&= \mathbb{E}\left[(m \cdot dQ)_{t+1} \middle| H_t, A_t, R_t = 1\right] \\
&= \mathbb{E}\left[\mathbb{E}\left\{(m \cdot dQ)_{t+1} \middle| H_{t+1}, A_{t+1}, R_{t+1}^{R_t=1}\right\} \middle| H_t, A_t, R_t = 1\right] \\
&= \int \mathbb{E}\left\{(m \cdot dQ)_{t+1} \middle| h_{t+1}^{A_t,H_t}, a_{t+1}, R_{t+1}^{R_t=1}\right\} d\mathbb{P}(a_{t+1} \mid h_{t+1}^{A_t,H_t}, R_{t+1}^{R_t=1}) d\mathbb{P}(x_{t+1}, R_{t+1} \mid H_t, A_t, R_t = 1) \\
&= \int \mathbb{E}\left\{(m \cdot dQ)_{t+1} \middle| h_{t+1}^{A_t,H_t}, a_{t+1}, R_{t+1}^{R_t=1}\right\} \\
&\qquad \times d\mathbb{P}(a_{t+1} \mid h_{t+1}^{A_t,H_t}, R_{t+1}^{R_t=1}) d\mathbb{P}(x_{t+1} \mid H_t, A_t, R_t = 1) d\mathbb{P}(R_{t+1} \mid H_t, A_t, R_t = 1) \\
&= \int \mathbb{E}\left\{(m \cdot dQ)_{t+1} \middle| h_{t+1}^{A_t,H_t}, a_{t+1}, R_{t+1}^{R_t=1}\right\} \\
&\qquad \times d\mathbb{P}(a_{t+1} \mid h_{t+1}^{A_t,H_t}, R_{t+1}^{R_t=1}) d\mathbb{P}(x_{t+1} \mid H_t, A_t, R_{t+1}^{R_t=1}) d\mathbb{P}(R_{t+1} \mid H_t, A_t, R_t = 1) \\
&= \mathbb{E}\left[(m \cdot dQ)_{t+1} \middle| H_t, A_t, R_{t+1}^{R_t=1}\right] d\mathbb{P}(R_{t+1} \mid H_t, A_t, R_t = 1)
\end{aligned}$$

, where both the fourth and the fifth equalities follow from assumption (A2-M). From this result, it is straightforward to see

$$\begin{aligned}
& \mathbb{1}(R_{t+1} = 1) m_t(H_t, A_t, R_{t+1} = 1) \\
&= \mathbb{1}(R_{t+1} = 1)\mathbb{E}\left[(m \cdot dQ)_{t+1} \middle| H_t, A_t, R_{t+1}^{R_t=1}\right] d\mathbb{P}(R_{t+1} \mid H_t, A_t, R_t = 1) \\
&= \mathbb{1}(R_{t+1} = 1)\mathbb{E}\left[(m \cdot dQ)_{t+1} \middle| H_t, A_t, R_{t+1} = 1\right] d\mathbb{P}(R_{t+1} = 1 \mid H_t, A_t, R_t = 1).
\end{aligned}$$

Finally assumption (A3) guarantees that we obtain

$$\left(\frac{\mathbb{1}(R_{t+1} = 1)}{d\mathbb{P}(R_{t+1} = 1 \mid H_t, A_t, R_t = 1)}\right) m_t(H_t, A_t, R_{t+1} = 1) = \mathbb{1}(R_{t+1} = 1) m_t(H_t, A_t, R_{t+1} = 1)$$

which is the desired identity. □

Finally, we are ready to give a proof of Theorem 2.4.1. In fact, it is nothing but rearranging terms in the given efficient influence function.

Proof of Theorem 2.4.1

Proof. First, we define following shorthand notations for the proof: for $\forall s \leq t$

$$dQ_s(A_s) \equiv dQ_s(A_s|H_s, R_s = 1), \qquad d\mathbb{P}_s(A_s) \equiv d\mathbb{P}(A_s \mid H_s, R_s = 1),$$

$$d\omega_s \equiv \omega_s(H_s, A_s) \equiv d\mathbb{P}(R_{s+1} = 1 \mid H_s, A_s, R'_s = 1),$$

$$m_s(H_s, a_s) \equiv m_s(H_s, a_s, R_{s+1} = 1)$$

With these notations we can rewrite the result of Lemma A.4.4 as below.

$$\begin{aligned}
&\sum_{s=0}^{t} \left\{ \int_{\mathscr{A}_{s+1}} m_{s+1}(H_{s+1}, a_{s+1}) dQ_{s+1}(a_{s+1}) - m_s(H_s, A_s) \right\} \mathbb{1}(R_{s+1} = 1) \left(\prod_{k=0}^{s} \frac{dQ_k(A_k)}{d\mathbb{P}_k(A_k)} \frac{1}{d\omega_k} \right) \\
&= \sum_{s=1}^{t} \left\{ \int_{\mathscr{A}_s} m_s(H_s, a_s) dQ_s(a_s) - m_s(H_s, A_s) \left[\mathbb{1}(R_{s+1} = 1) \frac{dQ_s(A_s)}{d\mathbb{P}_s(A_s)} \frac{1}{d\omega_s} \right] \right\} \\
&\qquad \times \mathbb{1}(R_s = 1) \left(\prod_{k=0}^{s-1} \frac{dQ_k(A_k)}{d\mathbb{P}_k(A_k)} \frac{1}{d\omega_k} \right) + \mathbb{1}(R_{t+1} = 1) \left(\prod_{s=1}^{t} \frac{dQ_s(A_s)}{d\mathbb{P}_s(A_s)} \frac{1}{d\omega_s} \right) Y_t - m_0.
\end{aligned}$$

Now, by the result of Lemma A.4.4 and A.4.5, we can represent the efficient influence function for $\psi_t(\delta)$ as

$$\begin{aligned}
&\sum_{s=1}^{t} \Bigg\{ \int_{\mathscr{A}_s} m_s(H_s, a_s) dQ_s(a_s) - m_s(H_s, A_s) \left[\mathbb{1}(R_{s+1} = 1) \frac{dQ_s(A_s)}{d\mathbb{P}_s(A_s)} \frac{1}{d\omega_s} \right] \\
&\qquad + \int_{\mathscr{A}_s} m_s(H_s, a_s) \phi_s(H_s, A_s, R_s = 1; a_s) d\nu(a_s) \Bigg\} \mathbb{1}(R_s = 1) \left(\prod_{k=0}^{s-1} \frac{dQ_k(A_k)}{d\mathbb{P}_k(A_k)} \frac{1}{d\omega_k} \right) \\
&\qquad + \mathbb{1}(R_{t+1} = 1) \left(\prod_{s=1}^{t} \frac{dQ_s(A_s)}{d\mathbb{P}_s(A_s)} \frac{1}{d\omega_s} \right) Y_t - m_0.
\end{aligned}$$

On the other hand, we have

$$\int_{\mathscr{A}_s} m_s(H_s, a_s) dQ_s(a_s) = \frac{m_s(H_s, 1)\delta\pi_s(H_s) + m_s(H_s, 0)\{1 - \pi_s(H_s)\}}{\delta\pi_s(H_s) + 1 - \pi_s(H_s)},$$

$$\frac{dQ_s(A_s)}{d\mathbb{P}_s(A_s)} = \frac{\delta A_s + 1 - A_s}{\delta\pi_s(H_s)+1-\pi_s(H_s)},$$

$$m_s(H_s,A_s)\frac{dQ_s(A_s)}{d\mathbb{P}_s(A_s)} = \frac{m_s(H_s,1,R_{s+1}=1)\delta A_s + m_s(H_s,0,R_{s+1}=1)(1-A_s)}{\delta\pi_s(H_s)+1-\pi_s(H_s)},$$

$$\int_{\mathscr{A}_s} m_s(H_s,a_s)\phi_s(H_s,A_s,R_s=1;a_s)d\nu(a_s) = \frac{\{m_s(H_s,1)-m_s(H_s,0)\}\delta(A_s-\pi_s(H_s))}{(\delta\pi_s(H_s)+1-\pi_s(H_s))^2}.$$

Now going back to the expression for the efficient influence function, note that by Lemma A.4.6 terms inside the summation before multiplied by $\mathbb{1}\left(R_s=1\right)\left(\prod_{k=0}^{s-1}\frac{dQ_k(A_k)}{d\mathbb{P}_k(A_k)}\frac{1}{d\omega_k}\right)$ simplify to

$$\begin{aligned}
&\int_{\mathscr{A}_s} m_s(H_s,a_s)dQ_s(a_s) - m_s(H_s,A_s)\left[\mathbb{1}\left(R_{s+1}=1\right)\frac{dQ_s(A_s)}{d\mathbb{P}_s(A_s)}\frac{1}{d\omega_s}\right] \\
&= \int_{\mathscr{A}_s}\mathbb{1}\left(R_{s+1}=1\right)m_s(H_s,a_s)dQ_s(a_s) - \mathbb{1}\left(R_{s+1}=1\right)m_s(H_s,A_s)\frac{dQ_s(A_s)}{d\mathbb{P}_s(A_s)} \\
&\quad + \int_{\mathscr{A}_s}\mathbb{1}\left(R_{s+1}=1\right)m_s(H_s,a_s)\phi_s(H_s,A_s,R_s'=1;a_s)d\nu(a_s) \\
&= \left[\frac{m_s(H_s,1)\delta\pi_s(H_s)+m_s(H_s,0)\{1-\pi_s(H_s)\}}{\delta\pi_s(H_s)+1-\pi_s(H_s)} + \frac{m_s(H_s,1)\delta A_s + m_s(H_s,0)(1-A_s)}{\delta\pi_s(H_s)+1-\pi_s(H_s)}\right. \\
&\qquad \left. + \frac{\{m_s(H_s,1)-m_s(H_s,0)\}\delta(A_s-\pi_s(H_s))}{(\delta\pi_s(H_s)+1-\pi_s(H_s))^2}\right]\mathbb{1}\left(R_{s+1}=1\right) \\
&= \quad \left[\frac{(\pi_s(H_s)-A_s)\{\delta m_s(H_s,1)-m_s(H_s,0)\}}{\delta\pi_s(H_s)+1-\pi_s(H_s)} + \frac{\{m_s(H_s,1)-m_s(H_s,0)\}\delta(A_s-\pi_s(H_s))}{(\delta\pi_s(H_s)+1-\pi_s(H_s))^2}\right]\mathbb{1}\left(R_{s+1}=1\right) \\
&= \left(\frac{\{A_s-\pi_s(H_s)\}\left(1-\delta\right)}{\delta\pi_s(H_s)+1-\pi_s(H_s)}\right)\left[\frac{m_s(H_s,1)\delta\pi_s(H_s)+m_s(H_s,0)\{1-\pi_s(H_s)\}}{\delta\pi_s(H_s)+1-\pi_s(H_s)}\right]\mathbb{1}\left(R_{s+1}=1\right)
\end{aligned}$$

By multiplying $\left[\frac{dQ_s(A_s)}{d\mathbb{P}_s(A_s)}\frac{1}{d\omega_s}\right]^{-1}$ to the last expression, we finally obtain an equivalent form of the efficient influence function for $\psi_t(\delta)$ as

$$\begin{aligned}
&\sum_{s=0}^{t}\left\{\frac{\{A_s-\pi_s(H_s)\}(1-\delta)}{\delta A_s+1-A_s}\right\}\left[\frac{m_s(H_s,1)\delta\pi_s(H_s)+m_s(H_s,0)\{1-\pi_s(H_s)\}}{\delta\pi_s(H_s)+1-\pi_s(H_s)}\right]\omega_s(H_s,A_s) \\
&\qquad \times\left(\prod_{k=1}^{s}\frac{\delta A_k+1-A_k}{\delta\pi_k(H_k)+1-\pi_k(H_k)}\cdot\frac{\mathbb{1}\left(R_{k+1}=1\right)}{\omega_k(H_k,A_k)}\right) + \prod_{s=1}^{t}\left\{\frac{\delta A_s+1-A_s}{\delta\pi_s(H_s)+1-\pi_s(H_s)}\cdot\frac{\mathbb{1}\left(R_{s+1}=1\right)}{\omega_s(H_s,A_s)}Y_t\right\} \\
&\quad -\psi_t(\delta).
\end{aligned}$$

□

A.4.3 Sequential regression formulation

The efficient influence function derived in the previous subsection involves pseudo-regression functions m, whose estimation in general might involve complicated conditional density estimation. However, as pointed out by Kennedy [67], one efficient strategy is to formulate a series of sequential regressions for m_s, as described in the subsequent remark in more detail.

Remark 5. *From the definition of m_s, it immediately follows that*

$$m_s = \int_{\mathscr{X}_s \times \mathscr{A}_s} m_{s+1} dQ_{s+1}(a_{s+1} \mid h_{s+1}, R_{s+1} = 1) d\mathbb{P}(x_{s+1} | h_s, a_s, R_{s+1} = 1).$$

Hence, we can find equivalent form of the functions $m_s(\cdot)$ in Theorem 2.4.1 as the following recursive regression:

$$\begin{aligned} & m_s(H_s, A_s, R_{s+1} = 1) \\ & = \mathbb{E}\left[\frac{m_{s+1}(H_{s+1}, a_{s+1}, 1)\delta\pi_{s+1}(H_{s+1}) + \{1 - m_{s+1}(H_{s+1}, 0, 1)\}\{1 - \pi_{s+1}(H_{s+1})\}}{\delta\pi_{s+1}(H_{s+1}) + 1 - \pi_{s+1}(H_{s+1})} \middle| H_s, A_s, R_{s+1} = 1 \right] \end{aligned}$$

for $s = 1, ..., t-1$, where we use shorthand notation $m_{s+1}(H_{s+1}, a_{s+1}, 1) = m_{s+1}(H_{s+1}, A_{s+1} = a_{s+1}, R_{t+2} = 1)$ and $m_s(H_s, A_s, 1) = \mu(H_s, A_s, R_{s+1} = 1)$.

Above sequential regression form is very practically useful when we estimate m_s, since it allows us to bypass all the conditional density estimations and instead use regression methods that are more readily available in statistical software.

A.4.4 EIF for $T = 1$

In the next corollary we provide the efficient influence function for the incremental effect in a single timepoint study ($T = 1$) whose identifying expression is given in Corollary 2.3.1.

Corollary A.4.1. *When $T = 1$, the efficient influence function for $\psi(\delta)$ in Corollary 2.3.1 is given by*

$$\mathbb{1}(R = 1) \left[\frac{\delta\pi(1|X)\phi_{1,R=1}(Z) + \pi(0|X)\phi_{0,R=1}(Z)}{\delta\pi(1|X) + \pi(0|X)} + \frac{\delta\{\mu(X,1,1) - \mu(X,0,1)\}(A - \pi(1|X))}{\{\delta\pi(1|X) + \pi(0|X)\}^2} \right]$$

where

$$\mu(x,a,1) = \mathbb{E}(Y \mid X = x, A = a, R = 1),$$

$$\pi(a|X) = d\mathbb{P}(A = a \mid X = x),$$

and

$$\phi_{a,R=1}(Z) = \frac{\mathbb{1}\,(A = a)\,\mathbb{1}\,(R = 1)}{\pi(a|X)\,\omega(X,a)}\{Y - \mu(X,a,1)\} + \mu(X,a,1)$$

which is the uncentered efficient influence function for $\mathbb{E}[\mu(X,a,1)]$.

The efficient influence function for the point exposure case has a simpler and more intuitive form. In fact, as stated in Corollary A.4.1, it is a weighted average of the two efficient influence functions $\phi_{0,R=1}, \phi_{1,R=1}$, plus a contribution term due to unknown propensity scores. An existence of the indicator function $\mathbb{1}\,(R = 1)$ proceeds from a likelihood of potential dropouts, and it implies that if a dropout occurs the outcome would not be available and consequently a contribution from the subject would not be taken into account.

A.4.5 Proof of Theorem 2.6.1

First we find an alternative form of the variance of each estimator, which eventually comes in handy for our proof. To this end, let $\widehat{\psi}_{c.ipw}(\overline{a'}_T)$ denote the standard IPW estimator of a classical deterministic intervention effect $\mathbb{E}\left[Y^{\overline{a'}_T}\right]$ under *i.i.d* assumption, i.e.

$$\widehat{\psi}_{c.ipw}(\overline{a'}_T) = \prod_{t=1}^{T}\left(\frac{\mathbb{1}\,(A_t = a'_t)}{\pi_t(a'_t|H_t)}\right)Y.$$

Hence $\widehat{\psi}_{c.ipw}(\overline{\mathbf{1}})$ is equivalent to $\widehat{\psi}_{at}$ in the main text. Now by definition we have

$$\begin{aligned} Var\left(\widehat{\psi}_{c.ipw}(\overline{a'}_T)\right) &= \mathbb{E}\left\{\left(\prod_{t=1}^{T}\frac{\mathbb{1}\,(A_t = a'_t)}{\pi_t(a'_t|H_t)^2}\right)Y^2\right\} - \left\{\mathbb{E}\left[\prod_{t=1}^{T}\frac{\mathbb{1}\,(A_t = a'_t)}{\pi_t(a'_t|H_t)}Y\right]\right\}^2 \\ &\equiv \mathbb{V}_{c.ipw.1}(\overline{a'}_T) - \mathbb{V}_{c.ipw.2}(\overline{a'}_T) \end{aligned}$$

where $\mathbb{V}_{c.ipw.1}(\overline{a'}_T)$ and $\mathbb{V}_{c.ipw.2}(\overline{a'}_T)$ are simply the first and second term in the first line of the expansion respectively.

By the same procedure to derive g-formula [108] it is easy to see

$$\begin{aligned}\mathbb{V}_{c.ipw.1}(\overline{a'}_T) &= \mathbb{E}\left\{\prod_{t=1}^{T}\left(\frac{\mathbb{1}\,(A_t = a'_t)}{\pi_t(a'_t|H_t)^2}\right)Y^2\right\}\\ &= \int_{\mathscr{X}} \mathbb{E}\left[Y^2 \mid \overline{X}_t, \overline{A}_t = \overline{a'}_t\right]\prod_{t=1}^{T}\frac{d\mathbb{P}(X_t \mid \overline{X}_{t-1}, \overline{A}_{t-1} = \overline{a'}_{t-1})}{\pi_t(a'_t|H_t)}\end{aligned}$$

where $\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_T$. Above result simply follows by iterative expectation conditioning on $\overline{X}_t$ and then another iterative expectation conditioning on H_t followed by the fact that $\mathbb{E}\left[\frac{\mathbb{1}(A_t=a'_t)}{\pi_t(a'_t|H_t)}|H_t\right] = 1$ for all t. We repeat this process T times, starting from $t = T$ all the way through $t = 1$.

Likewise, for $\widehat{\psi}_{inc}$ we have

$$\begin{aligned}Var(\widehat{\psi}_{inc}) &= \mathbb{E}\left\{\prod_{t=1}^{T}\left(\frac{\delta A_t + 1 - A_t}{\delta\pi_t(H_t) + 1 - \pi_t(H_t)}\right)^2 Y^2\right\} - \left\{\mathbb{E}\left[\prod_{t=1}^{T}\left(\frac{\delta A_t + 1 - A_t}{\delta\pi_t(H_t) + 1 - \pi_t(H_t)}\right)Y\right]\right\}^2\\ &\equiv \mathbb{V}_{inc.1} - \mathbb{V}_{inc.2}\end{aligned}$$

For the first term $\mathbb{V}_{inc.1}$, observe that

$$\begin{aligned}&\mathbb{E}\left\{\prod_{t=1}^{T}\left(\frac{\delta A_t + 1 - A_t}{\delta\pi_t(H_t) + 1 - \pi_t(H_t)}\right)^2 Y^2\right\}\\ &= \mathbb{E}\left\{\prod_{t=1}^{T-1}\left(\frac{\delta A_t + 1 - A_t}{\delta\pi_t(H_t) + 1 - \pi_t(H_t)}\right)^2 \mathbb{E}\left[\left(\frac{\delta A_T + 1 - A_T}{\delta\pi_T(H_T) + 1 - \pi_T(H_T)}\right)^2 Y^2 \middle| H_T\right]\right\}\\ &= \mathbb{E}\left\{\prod_{t=1}^{T-1}\left(\frac{\delta A_t + 1 - A_t}{\delta\pi_t(H_t) + 1 - \pi_t(H_t)}\right)^2 \mathbb{E}\left[\frac{\delta^2 Y^2}{(\delta\pi_T(H_T) + 1 - \pi_T(H_T))^2}\middle| H_T, A_T = 1\right]\pi_T(H_T)\right\}\\ &\quad + \mathbb{E}\left\{\prod_{t=1}^{T-1}\left(\frac{\delta A_t + 1 - A_t}{\delta\pi_t(H_t) + 1 - \pi_t(H_t)}\right)^2 \mathbb{E}\left[\frac{Y^2}{(\delta\pi_T + 1 - \pi_T)^2}\middle| H_T, A_T = 0\right](1 - \pi_T(H_T))\right\}\end{aligned}$$

where we apply the law of total expectation in the first equality and the law of total probability in the second.

After repeating the same process for $T-1$ times, for $t = T-1, ..., 1$, we obtain 2^T terms in the end where each of which corresponds to distinct treatment sequence $\overline{A}_T = \overline{a}_T$. Hence, we eventually have

$$\mathbb{V}_{inc.1} = \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} \int_{\mathscr{X}} \mathbb{E}\left[Y^2 \mid H_T, A_T = a_T\right] \prod_{t=1}^{T} \frac{\mathbb{1}\,(a_t = 1)\,\delta^2 \pi_t(H_t) + \mathbb{1}\,(a_t = 0)\,\{1 - \pi_t(H_t)\}}{(\delta \pi_t(H_t) + 1 - \pi_t(H_t))^2}$$
$$\times\, d\mathbb{P}(X_t \mid \overline{X}_{t-1}, \overline{A}_{t-1} = \overline{a}_{t-1}).$$

Recall that we assume $\pi_t(H_t) = p$ for all t as stated in Theorem 2.6.1. Hence we can write $\pi_t(a_t \mid H_t)$ as $\pi_t(a_t) = \mathbb{1}\,(a_t = 1)\,p + \mathbb{1}\,(a_t = 0)\,\{1 - p\}$.

Next we notice that to compute the upper bound of $RE(\widehat{\psi}_{c.ipw}(\overline{a}_T), \widehat{\psi}_{inc}) = \frac{\mathbb{V}_{inc.1} - \mathbb{V}_{inc.2}}{\mathbb{V}_{c.ipw.1}(\overline{a}_T) - \mathbb{V}_{c.ipw.2}(\overline{a}_T)}$ for always-treated unit (i.e. $\overline{a}_T = \overline{\mathbf{1}}$) it suffices to compute the quantity

$$\frac{\mathbb{V}_{inc.1}}{\mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}}) - \mathbb{V}_{c.ipw.2}(\overline{\mathbf{1}})}$$

since $0 < \mathbb{V}_{inc.2} < \mathbb{V}_{inc.1}$ by Jensen's inequality.

On the other hand, we have

$$\begin{aligned}\mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}}) - \mathbb{V}_{c.ipw.2}(\overline{\mathbf{1}}) &= \int_{\mathscr{X}} \mathbb{E}\left[Y^2 \mid \overline{X}_T, \overline{A}_T = \overline{a'}_T\right] \prod_{t=1}^{T} \frac{d\mathbb{P}(X_t \mid \overline{X}_{t-1}, \overline{A}_{t-1} = \overline{a'}_{t-1})}{p} - \left(\mathbb{E}[Y^{\overline{\mathbf{1}}}]\right)^2 \\ &= \left(\frac{1}{p}\right)^T \mathbb{E}\left[\left(Y^{\overline{\mathbf{1}}}\right)^2\right] - \left(\mathbb{E}\left[Y^{\overline{\mathbf{1}}}\right]\right)^2\end{aligned}$$

, and under the given boundedness assumption we see the ratio of the second term to the first term becomes quickly (at least exponentially) negligible as t increases. Hence we can write

$$\frac{1}{\mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}}) - \mathbb{V}_{c.ipw.2}(\overline{\mathbf{1}})} \leq \frac{1}{\mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}})} \left(1 + \frac{c\left(\mathbb{E}\left[Y^{\overline{\mathbf{1}}}\right]\right)^2}{(1/p)^T\, \mathbb{E}\left[\left(Y^{\overline{\mathbf{1}}}\right)^2\right]}\right)$$

for some constant c such that $\frac{1}{1 - \mathbb{V}_{c.ipw.2}(\overline{\mathbf{1}})/\mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}})} = \frac{1}{1 - p^T\left(\mathbb{E}\left[Y^{\overline{\mathbf{1}}}\right]\right)^2 / \mathbb{E}\left[\left(Y^{\overline{\mathbf{1}}}\right)^2\right]} \leq c$. Note that in our setting in which we have an infinitely large value of T, c can be almost any constant greater than one.

Putting above ingredients together, for sufficiently large t it follows that

$$RE(\widehat{\psi}_{c.ipw}(\overline{\mathbf{1}}), \widehat{\psi}_{inc}) \leq \frac{\mathbb{V}_{inc.1}}{\mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}})} \left(1 + \frac{c\left(\mathbb{E}\left[Y^{\overline{\mathbf{1}}}\right]\right)^2}{(1/p)^T\, \mathbb{E}\left[\left(Y^{\overline{\mathbf{1}}}\right)^2\right]}\right),$$

where we have

$$
\begin{aligned}
\frac{\mathbb{V}_{inc.1}}{\mathbb{V}_{c.ipw.1}(\bar{\mathbf{1}})} &= \frac{w(\bar{\mathbf{1}})\mathbb{V}_{c.ipw.1}(\bar{\mathbf{1}}) + \sum_{\bar{a}_T \neq \bar{\mathbf{1}}} w(\bar{a}_T;\delta,p)\mathbb{V}_{c.ipw.1}(\bar{a}_T)}{\mathbb{V}_{c.ipw.1}(\bar{\mathbf{1}})} \\
&= w(\bar{\mathbf{1}}) + \sum_{\bar{a}_T \neq \bar{\mathbf{1}}} w(\bar{a}_T;\delta,p) \prod_{t=1}^{T} \left(\frac{p}{\pi_t(a_t)} \frac{\mathbb{E}\left[(Y^2)^{\bar{a}_T} \right]}{\mathbb{E}\left[\left(Y^{\bar{\mathbf{1}}}\right)^2 \right]} \right) \\
&\leq \frac{b_u^2}{\mathbb{E}\left[\left(Y^{\bar{\mathbf{1}}}\right)^2 \right]} \left\{ w(\bar{\mathbf{1}}) + \sum_{\bar{a}_T \neq \bar{\mathbf{1}}} \left[\prod_{t=1}^{T} \frac{\mathbb{1}\,(a_t = 1)\,\delta^2 p^2 + \mathbb{1}\,(a_t = 0)\,(1-p)p}{(\delta p + 1 - p)^2} \right] \right\} \\
&= \frac{b_u^2}{\mathbb{E}\left[\left(Y^{\bar{\mathbf{1}}}\right)^2 \right]} \left\{ \frac{\delta^2 p^2 + p(1-p)}{(\delta p + 1 - p)^2} \right\}^T
\end{aligned}
$$

where the first equality follows by the fact that $\mathbb{V}_{inc.1} = \sum_{\bar{a}_T \in \overline{\mathscr{A}}_T} w(\bar{a}_T;\delta,p)\mathbb{V}_{c.ipw.1}(\bar{a}_T)$ derived in the proof of the first part, the second equality by the fact that $\mathbb{V}_{c.ipw.1}(\bar{a}_T) = \prod_{t=1}^{T} \frac{1}{\pi_t(a_t)} \mathbb{E}\left[(Y^2)^{\bar{a}_T} \right]$, the first inequality by definition of $w(\bar{a}_T;\delta,p)$ and the given boundedness assumption, and the last equality by binomial theorem. Therefore we obtain the upper bound as

$$
RE(\widehat{\psi}_{c.ipw}(\bar{\mathbf{1}}), \widehat{\psi}_{inc}) \leq \frac{b_u^2}{\mathbb{E}\left[\left(Y^{\bar{\mathbf{1}}}\right)^2 \right]} \left\{ \frac{\delta^2 p^2 + p(1-p)}{(\delta p + 1 - p)^2} \right\}^T \left(1 + \frac{c\left(\mathbb{E}\left[Y^{\bar{\mathbf{1}}}\right]\right)^2}{(1/p)^T\, \mathbb{E}\left[\left(Y^{\bar{\mathbf{1}}}\right)^2 \right]} \right).
$$

Next for the lower bound, first we note that

$$\begin{aligned}
\mathbb{V}_{inc.2} &= \left\{ \mathbb{E}\left[\prod_{t=1}^{T} \left(\frac{\delta A_t + 1 - A_t}{\delta p + 1 - p} \right) Y \right] \right\}^2 \\
&= \left\{ \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} \int_{\mathscr{X}} \mathbb{E}\left[Y \mid H_T, A_T = a_T \right] \left(\prod_{t=1}^{T} \frac{\mathbb{1}\left(a_t = 1\right) \delta p + \mathbb{1}\left(a_t = 0\right)(1-p)}{\delta p + 1 - p} \right) \right. \\
&\qquad\qquad \left. \times \, d\mathbb{P}(X_t \mid \overline{X}_{t-1}, \overline{A}_{t-1} = \overline{a}_{t-1}) \right\}^2 \\
&\leq b_u^2 \left[\sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} \prod_{t=1}^{T} \left(\frac{\mathbb{1}\left(a_t = 1\right) \delta p + \mathbb{1}\left(a_t = 0\right)(1-p)}{\delta p + 1 - p} \right) \right]^2 \\
&= b_u^2 \left(\frac{\delta p + 1 - p}{\delta p + 1 - p} \right)^{2T} = b_u^2
\end{aligned}$$

where the first equality follows by definition, the second equality by exactly same process used to find the expression for $\mathbb{V}_{inc.1}$, the first inequality by the boundedness assumption, and the third equality by binomial theorem.

However, we already know that

$$\mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}}) - \mathbb{V}_{c.ipw.2}(\overline{\mathbf{1}}) \leq \mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}}) = \left(\frac{1}{p} \right)^T \mathbb{E}\left[\left(Y^{\overline{1}} \right)^2 \right].$$

Hence putting these together we conclude

$$\begin{aligned}
RE(\widehat{\psi}_{c.ipw}(\overline{\mathbf{1}}), \widehat{\psi}_{inc}) &= \frac{\mathbb{V}_{inc.1} - \mathbb{V}_{inc.2}}{\mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}}) - \mathbb{V}_{c.ipw.2}(\overline{\mathbf{1}})} \\
&\geq \frac{\mathbb{V}_{inc.1} - b_u^2}{\mathbb{V}_{c.ipw.1}(\overline{\mathbf{1}})} \\
&= \frac{b_u^2}{\mathbb{E}\left[\left(Y^{\overline{1}} \right)^2 \right]} \left\{ \frac{\delta^2 p^2 + p(1-p)}{(\delta p + 1 - p)^2} \right\}^T - \frac{b_u^2}{\mathbb{E}\left[\left(Y^{\overline{1}} \right)^2 \right]} p^T.
\end{aligned}$$

At this point, we obtain upper and lower bound for $RE(\widehat{\psi}_{c.ipw}(\overline{\mathbf{1}}), \widehat{\psi}_{inc})$, which yields the result of part *ii*) having $C_T = \frac{b_u^2}{\mathbb{E}\left[\left(Y^{\overline{1}} \right)^2 \right]}$.

Proof for the case of $\overline{a'}_T = \overline{\mathbf{0}}$ (*never-treated unit*) is based on the almost same steps as the case of $\overline{a'}_T = \overline{\mathbf{1}}$ except for the rearragement of terms due to replacing $\left(\frac{1}{p} \right)^T$ by $\left(\frac{1}{1-p} \right)^T$

and so on. In fact, due to the generality of our proof structure, the exact same logic used for $\widehat{\psi}_{c.ipw}(\overline{\mathbf{1}})$ also applies to $\widehat{\psi}_{c.ipw}(\overline{\mathbf{0}})$ (and $\widehat{\psi}_{c.ipw}(\overline{a'}_T)$ for $\forall \overline{a'}_T \in \overline{\mathscr{A}_T}$). We present the result without the proof as below.

$$\begin{aligned} C'_T \left[\left\{ \frac{\delta^2 p(1-p) + (1-p)^2}{(\delta p + 1 - p)^2} \right\}^T - (1-p)^T \right] &\leq RE(\widehat{\psi}_{c.ipw}(\overline{\mathbf{0}}), \widehat{\psi}_{inc}) \\ &\leq C'_T \zeta'(T;p) \left\{ \frac{\delta^2 p(1-p) + (1-p)^2}{(\delta p + 1 - p)^2} \right\}^T \end{aligned}$$

where we define $C'_T = \frac{b_u^2}{\mathbb{E}\left[\left(Y^2\right)^{\overline{\mathbf{0}}}\right]}$ and $\zeta'(T;p) = \left(1 + \frac{c\left(\mathbb{E}\left[Y^{\overline{\mathbf{1}}}\right]\right)^2}{(1/(1-p))^T \mathbb{E}\left[\left(Y^{\overline{\mathbf{1}}}\right)^2\right]}\right)$.

A.4.6 Proof of Corollary 2.6.1

Now we provide following Lemma A.4.7 which becomes a key to prove Corollary 2.6.1.

Lemma A.4.7. *Assume that $\pi_t(H_t) = p$ for all $1 \leq t \leq T$ for $0 < p < 1$. Then we have following variance decomposition :*

$$Var(\widehat{\psi}_{inc}) = Var\left(\sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} \sqrt{w(\overline{a}_T; \delta, p)}\, \widehat{\psi}_{c.ipw}(\overline{a}_T) \right)$$

where for $\forall \overline{a}_T \in \overline{\mathscr{A}}_T$ the weight w is defined by

$$w(\overline{a}_T; \delta, p) = \prod_{t=1}^{T} \frac{\pi_t(a_t) \left\{ \mathbb{1}\left(a_t = 1\right) \delta^2 p + \mathbb{1}\left(a_t = 0\right)(1-p) \right\}}{(\delta \pi_t(H_t) + 1 - \pi_t(H_t))^2}.$$

Proof. From the last display for $\mathbb{V}_{inc.1}$, we have that

$$
\begin{aligned}
&\mathbb{V}_{inc.1} \\
&= \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} \int_{\mathscr{X}} \mathbb{E}\left[Y^2 \mid H_T, A_T = a_T\right] \prod_{t=1}^{T} \frac{\pi_t(a_t)\left(\mathbb{1}\left(a_t = 1\right)\delta^2 p + \mathbb{1}\left(a_t = 0\right)\{1-p\}\right)}{(\delta p + 1 - p)^2} \\
&\qquad \times \prod_{t=1}^{T} \frac{d\mathbb{P}(X_t \mid \overline{X}_{t-1}, \overline{A}_{t-1} = \overline{a}_{t-1})}{\pi_t(a_t)} \\
&= \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} w(\overline{a}_T; \delta, p) \int_{\mathscr{X}} \mathbb{E}\left[Y^2 \mid H_T, A_T = a_T\right] \prod_{t=1}^{T} \frac{d\mathbb{P}(X_t \mid \overline{X}_{t-1}, \overline{A}_{t-1} = \overline{a}_{t-1})}{\pi_t(a_t)} \\
&= \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} w(\overline{a}_T; \delta, p) \mathbb{V}_{c.ipw.1}(\overline{a}_T)
\end{aligned}
$$

where we let weight $w(\overline{a}_T; \delta, p)$ denote the product term $\prod_{t=1}^{T} \frac{\pi_t(a_t)\left(\mathbb{1}(a_t=1)\delta^2 p + \mathbb{1}(a_t=0)\{1-p\}\right)}{(\delta \pi_t(H_t) + 1 - \pi_t(H_t))^2}$.

Next, we observe that

$$
\begin{aligned}
\mathbb{V}_{inc.2} &= \left\{\mathbb{E}\left[\prod_{t=1}^{T}\left(\frac{\delta A_t + 1 - A_t}{\delta p + 1 - p}\right) Y\right]\right\}^2 \\
&= \left\{\mathbb{E}\left[\prod_{t=1}^{T}\left(\frac{\delta \mathbb{1}\left(A_t = 1\right)}{\delta p + 1 - p}\right) Y + \cdots + \prod_{t=1}^{T}\left(\frac{\mathbb{1}\left(A_t = 0\right)}{\delta p + 1 - p}\right) Y\right]\right\}^2 \\
&= \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} v_{inc.2}^2(\overline{A}_T; \overline{a}_T) + \sum_{\overline{a'}_T \neq \overline{a}_T} v_{inc.2}(\overline{A}_T; \overline{a}_T) v_{inc.2}(\overline{A}_T; \overline{a'}_T)
\end{aligned}
$$

where we have decomposed $\mathbb{V}_{inc.2}$ into $2^T \times 2^T$ terms by defining $v_{inc.2}(\overline{A}_T; \overline{a}_T)$ by

$$
v_{inc.2}(\overline{A}_T; \overline{a}_T) \equiv \mathbb{E}\left[\prod_{t=1}^{T}\left(\frac{\delta \mathbb{1}(a_t = 1) + \mathbb{1}(a_t = 0)}{\delta p + 1 - p}\right) \mathbb{1}(A_t = a_t) \cdot Y\right].
$$

Then for fixed $\overline{a}_T$ it is straightforward to see that

$$
\begin{aligned}
\frac{v_{inc.2}^2(\overline{A}_T; \overline{a}_T)}{w(\overline{a}_T; \delta, p)} &= \left\{\mathbb{E}\left[\prod_{t=1}^{T}\left(\frac{\{\delta \mathbb{1}(a_t = 1) + \mathbb{1}(a_t = 0)\} \mathbb{1}(A_t = a_t)}{\sqrt{\pi(a_t)\left(\mathbb{1}\left(a_t = 1\right)\delta^2 p + \mathbb{1}\left(a_t = 0\right)\{1-p\}\right)}}\right) Y\right]\right\}^2 \\
&= \left\{\mathbb{E}\left[\prod_{t=1}^{T}\left(\frac{\mathbb{1}(A_t = a_t)}{\pi(a_t)}\right) Y\right]\right\}^2 = \mathbb{V}_{c.ipw.2}(\overline{a}_T)
\end{aligned}
$$

Now putting this together, we obtain

$$\begin{aligned}
&\mathbb{V}_{inc.1} - \mathbb{V}_{inc.2} \\
&= \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} w(\overline{a}_T;\delta,p)\left\{\mathbb{V}_{c.ipw.1}(\overline{a}_T) - \mathbb{V}_{c.ipw.2}(\overline{a}_T)\right\} - \sum_{\overline{a'}_T \neq \overline{a}_T} v_{inc.2}(\overline{A}_T;\overline{a}_T)v_{inc.2}(\overline{A}_T;\overline{a'}_T) \\
&= \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} w(\overline{a}_T;\delta,p)Var\left(\widehat{\psi}_{c.ipw}(\overline{a}_T)\right) - \sum_{\overline{a'}_T \neq \overline{a}_T} v_{inc.2}(\overline{A}_T;\overline{a}_T)v_{inc.2}(\overline{A}_T;\overline{a'}_T).
\end{aligned}$$

However, from the second term in the last display one could notice that

$$\begin{aligned}
\frac{v_{inc.2}(\overline{A}_T;\overline{a}_T)v_{inc.2}(\overline{A}_T;\overline{a'}_T)}{\sqrt{w(\overline{a}_T;\delta,p)w(\overline{a'}_T;\delta,p)}} &= \mathbb{E}\left[\prod_{t=1}^{T}\left(\frac{\mathbb{1}(A_t = a_t)}{\pi(a_t)}\right)Y\right]\mathbb{E}\left[\prod_{t=1}^{T}\left(\frac{\mathbb{1}(A_t = a'_t)}{\pi(a'_t)}\right)Y\right] \\
&= -Cov(\widehat{\psi}_{c.ipw}(\overline{a}_T), \widehat{\psi}_{c.ipw}(\overline{a'}_T))
\end{aligned}$$

where the last equality follows by the fact that

$$\mathbb{E}\left\{\prod_{t=1}^{T}\left(\frac{\mathbb{1}(A_t = a_t)}{\pi(a_t)}\right)\prod_{t=1}^{T}\left(\frac{\mathbb{1}(A_t = a'_t)}{\pi(a'_t)}\right)Y^2\right\} = 0 \quad \text{for } \forall \overline{a'}_T \neq \overline{a}_T.$$

Hence finally we conclude that

$$\begin{aligned}
&Var(\widehat{\psi}_{inc}) = \mathbb{V}_{inc.1} - \mathbb{V}_{inc.2} \\
&= \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} w(\overline{a}_T;\delta,p)Var\left(\widehat{\psi}_{c.ipw}(\overline{a}_T)\right) \\
&\quad + \sum_{\substack{\overline{a}_T,\overline{a'}_T \in \overline{\mathscr{A}}_T \\ \overline{a'}_T \neq \overline{a}_T}} \sqrt{w(\overline{a}_T;\delta,p)w(\overline{a'}_T;\delta,p)}Cov(\widehat{\psi}_{c.ipw}(\overline{a}_T), \widehat{\psi}_{c.ipw}(\overline{a'}_T)) \\
&= \sum_{\overline{a}_T,\overline{a'}_T \in \overline{\mathscr{A}}_T} \sqrt{w(\overline{a}_T;\delta,p)w(\overline{a'}_T;\delta,p)}Cov(\widehat{\psi}_{c.ipw}(\overline{a}_T), \widehat{\psi}_{c.ipw}(\overline{a'}_T)).
\end{aligned}$$

□

In Lemma A.4.7 it should be noticed that the weight $w(\overline{a}_T;\delta,p)$ exponentially and monotonically decays to zero for $\forall \overline{a}_T \in \overline{\mathscr{A}}_T$.

Now we show that there always exists T_{min} such that $Var(\widehat{\psi}_{inc}) < Var(\widehat{\psi}_{c.ipw}(\overline{\mathbf{1}}))$ for all $T \geq T_{min}$. Let $\overline{\mathbf{1}} = [1,...,1]$. From Lemma A.4.7 it follows that

$$
\begin{aligned}
& Var(\widehat{\psi}_{inc}) - Var(\widehat{\psi}_{c.ipw}(\overline{\mathbf{1}})) \\
& = \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} w(\overline{a}_T;\delta,p) Var(\widehat{\psi}_{c.ipw}(\overline{a}_T)) - Var(\widehat{\psi}_{c.ipw}(\overline{\mathbf{1}})) \\
& \quad + \sum_{\substack{\overline{a}_T, \overline{a'}_T \in \overline{\mathscr{A}}_T \\ \overline{a'}_T \neq \overline{a}_T}} \sqrt{w(\overline{a}_T;\delta,p) w(\overline{a'}_T;\delta,p)} Cov(\widehat{\psi}_{c.ipw}(\overline{a}_T), \widehat{\psi}_{c.ipw}(\overline{a'}_T)) \\
& = \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} \prod_{t=1}^{T} \frac{\pi_t(a_t)\left\{\mathbb{1}(a_t = 1)\delta^2 p + \mathbb{1}(a_t = 0)(1-p)\right\}}{(\delta p + 1 - p)^2} \left(\prod_{t=1}^{T} \frac{1}{\pi_t(a_t)} \mathbb{E}\left[(Y^2)^{\overline{a}_T}\right] - \left(\mathbb{E}\left[Y^{\overline{a}_T}\right]\right)^2 \right) \\
& \quad - \left(\frac{1}{p}\right)^T \mathbb{E}\left[\left(Y^{\overline{\mathbf{1}}}\right)^2\right] + \left(\mathbb{E}\left[Y^{\overline{\mathbf{1}}}\right]\right)^2 - \sum_{\substack{\overline{a}_T, \overline{a'}_T \in \overline{\mathscr{A}}_T \\ \overline{a'}_T \neq \overline{a}_T}} \sqrt{w(\overline{a}_T;\delta,p) w(\overline{a'}_T;\delta,p)} \mathbb{E}\left[Y^{\overline{a}_T}\right] \mathbb{E}\left[Y^{\overline{a'}_T}\right] \\
& \leq b_u^2 \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} \left(\prod_{t=1}^{T} \frac{\mathbb{1}(a_t = 1)\delta^2 p + \mathbb{1}(a_t = 0)(1-p)}{(\delta p + 1 - p)^2} \right) - \left(\frac{1}{p}\right)^T \mathbb{E}\left[\left(Y^{\overline{\mathbf{1}}}\right)^2\right] + \left(\mathbb{E}\left[Y^{\overline{\mathbf{1}}}\right]\right)^2 \\
& \quad - \sum_{\substack{\overline{a}_T, \overline{a'}_T \in \overline{\mathscr{A}}_T \\ \overline{a'}_T \neq \overline{a}_T}} \sqrt{w(\overline{a}_T;\delta,p) w(\overline{a'}_T;\delta,p)} \mathbb{E}\left[Y^{\overline{a}_T}\right] \mathbb{E}\left[Y^{\overline{a'}_T}\right] + \sum_{\overline{a}_T \in \overline{\mathscr{A}}_T} w(\overline{a}_T;\delta,p) \left(\mathbb{E}\left[Y^{\overline{a}_T}\right]\right)^2 \\
& = b_u^2 \left\{ \left[\frac{\delta^2 p + 1 - p}{(\delta p + 1 - p)^2}\right]^T - \left(\frac{c_{\mathbf{1}}^{1/T}}{p}\right)^T \right\} \\
& \quad - \sum_{\overline{a}_T, \overline{a'}_T \in \overline{\mathscr{A}}_T} \sqrt{w(\overline{a}_T;\delta,p) w(\overline{a'}_T;\delta,p)} \mathbb{E}\left[Y^{\overline{a}_T}\right] \mathbb{E}\left[Y^{\overline{a'}_T}\right] + \left(\mathbb{E}\left[Y^{\mathbf{1}}\right]\right)^2 \\
& = b_u^2 \left\{ \left[\frac{\delta^2 p + 1 - p}{(\delta p + 1 - p)^2}\right]^T - \left(\frac{c_{\mathbf{1}}^{1/T}}{p}\right)^T - A(\delta,p) + B \right\}
\end{aligned}
$$

where $c_{\mathbf{1}} = \frac{\mathbb{E}\left[\left(Y^{\overline{\mathbf{1}}}\right)^2\right]}{b_u^2}$, $A(\delta,p) = \sum_{\overline{a}_T, \overline{a'}_T \in \overline{\mathscr{A}}_T} \sqrt{w(\overline{a}_T;\delta,p) w(\overline{a'}_T;\delta,p)} \frac{\mathbb{E}\left[Y^{\overline{a}_T}\right]}{b_u} \frac{\mathbb{E}\left[Y^{\overline{a'}_T}\right]}{b_u}$, and $B = \frac{\left(\mathbb{E}\left[Y^{\overline{\mathbf{1}}}\right]\right)^2}{b_u^2}$. The inequality comes from the boundedness condition. It can be immediately noted that $c_{\mathbf{1}}^{1/T} \to 1$ as $T \to \infty$ very quickly and monotonically. Also we note $|A(\delta,p)| \leq 1$ and $0 \leq B \leq 1$.

For $\delta > 1$, $\frac{\delta^2 p + 1 - p}{(\delta p + 1 - p)^2} < \frac{1}{p}$. Hence based on above observation, it follows that for sufficiently large T the last display is strictly less than zero. Consequently we conclude

$Var(\widehat{\psi}_{inc}) - Var(\widehat{\psi}_{c.ipw}(\overline{\mathbf{1}})) < 0$ for all $T \geq T_{min}$, which is the result of part $i)$. Likewise, we have the same conclusion for $\overline{\mathbf{0}}_T = [0,...,0]$ such that $Var(\widehat{\psi}_{inc}) - Var(\widehat{\psi}_{c.ipw}(\overline{\mathbf{0}}_T)) < 0$.

The value of T_{min} is determined by δ, p, and distribution of counterfactual outcome $Y^{\bar{a}_T}$. One rough upper bound of such T_{min} is

$$\min\left\{T : \left[\frac{\delta^2 p + 1 - p}{(\delta p + 1 - p)^2}\right]^T - \frac{c_1}{p^T} + 2 < 0\right\}$$

which could be obtained by the last display above and is always finite due to the fact $c_1 > 0$ by given assumption in the theorem. T_{min} should not be very large for moderately large value of δ unless c_1 is unreasonably small since the difference $\frac{1}{p^T} - \left[\frac{\delta^2 p+1-p}{(\delta p+1-p)^2}\right]^T$ also grows exponentially.

A.4.7 Proof of Theorem 2.5.1

First we need to define the following notations:

$$\|f\,|_{\mathscr{D},\mathscr{T}} \equiv \sup_{\delta\in\mathscr{D},t\in\mathscr{T}} |f(\delta,t)\,|$$

$$\widehat{\Psi}_n(\delta,t) \equiv \sqrt{n}\{\widehat{\psi}_t(\delta) - \psi_t(\delta)\}/\widehat{\sigma}(\delta,t)$$

$$\widetilde{\Psi}_n(\delta,t) \equiv \sqrt{n}\{\widehat{\psi}_t(\delta) - \psi_t(\delta)\}/\sigma(\delta,t)$$

$$\Psi_n(\delta;t) \equiv \mathbb{G}_n\{\widetilde{\varphi}(Z;\boldsymbol{\eta},\delta,t)\}$$

where we let $\mathscr{T} = \{1,...,T\}$, let $\mathbb{G}_n$ denote the empirical process on the full sample as usual, and let $\widetilde{\varphi}(Z;\boldsymbol{\eta},\delta,t) = \{\varphi(Z;\boldsymbol{\eta},\delta,t) - \psi(t;\delta)\}/\sigma(\delta;t)$ and let $\mathbb{G}$ be a mean-zero Gaussian process with covariance $\mathbb{E}[\mathbb{G}(\delta_1;t_1)\mathbb{G}(\delta_2;t_2)] = \mathbb{E}[\widetilde{\varphi}(Z;\boldsymbol{\eta},\delta_1,t_1)\widetilde{\varphi}(Z;\boldsymbol{\eta},\delta_2,t_2)]$ as defined in Theorem 2.5.1 in the main text.

The proof consists of two parts; in the first part we will show $\Psi_n(\cdot) \rightsquigarrow \mathbb{G}(\cdot)$ in $l^\infty(\mathscr{D},\mathscr{T})$ and in the second we will show $\|\widehat{\Psi}_n - \Psi_n\,|_{\mathscr{D},\mathscr{T}} = o_{\mathbb{P}}(1)$.

Part 1. A proof of the first statement immediately follows from the proof of Theorem 3 in Kennedy [67]. He showed the function class $\mathscr{F}_{\bar{\boldsymbol{\eta}}} = \{\varphi(\cdot;\bar{\boldsymbol{\eta}},\delta) : \delta \in \mathscr{D}\}$ is Lipschitz and thus has a finite bracketing integral for any fixed set of nuisance functions, and then applied Theorem 2.5.6 in Van Der Vaart and Wellner [142]. In our case, the function class

$\mathscr{F}_{\bar{\boldsymbol{\eta}}} = \{\varphi(\cdot;\bar{\boldsymbol{\eta}},\delta,t) : \delta \in \mathscr{D}, t \le T\}$ is still Lipschitz, since for $\forall t \in \{1,...,T\}$ we have

$$\left|\frac{\partial}{\partial\delta}\left[\frac{\{a_t-\pi_t(h_t)\}(1-\delta)}{\delta a_t+1-a_t}\right]\right| \le \frac{1}{\delta_l}+\frac{1}{4\delta_l^2}$$

$$\left|\frac{\partial}{\partial\delta}\left[\frac{m_t(h_t,1,1)\delta\pi_t(h_t)+m_t(h_t,0,1)\{1-\pi_t(h_t)\}}{\delta\pi_t(h_t)+1-\pi_t(h_t)}\cdot\omega_t(h_t,a_t)\right]\right| \le \frac{2C}{\delta_l^2}$$

$$\frac{\partial}{\partial\delta}\left[\frac{\delta a_t+1-a_t}{\delta\pi_t(h_t)+1-\pi_t(h_t)}\cdot\frac{1}{\omega_t(h_t,a_t)}\right] \le \frac{1}{c_\omega\delta_l^2}$$

where we use assumption 1) and 2) in the Theorem, and the identification assumption (A3) that there exist a constant c_ω such that $0 < \omega_t(h_t,a_t) < c_\omega \le 1$ and thus $\frac{1}{\omega_t(h_t,a_t)} \le \frac{1}{c_\omega}$ a.e. [$\mathbb{P}$]. Therefore, every $\varphi(\cdot;\bar{\boldsymbol{\eta}},\delta,t)$ is basically a finite sum of products of Lipschitz functions with bounded $\mathscr{D}$ and we conclude $\mathscr{F}_{\bar{\boldsymbol{\eta}}}$ is Lipschitz.

Hence our function class still has a finite bracketing integral for fixed $\bar{\boldsymbol{\eta}}$ and t, which concludes the first statement is true.

Part 2. Let $N = n/K$ be the sample size in any group $k = 1,...,K$, and denote the empirical process over group k units by $\mathbb{G}_n^k = \sqrt{N}(\mathbb{P}_n^k - \mathbb{P})$. From the result of Part 1 and the proof of Theorem 3 in Kennedy [67] we have

$$\begin{aligned}
&\widetilde{\Psi}_n(\delta;t) - \Psi_n(\delta;t)\\
&= \frac{\sqrt{n}}{K\sigma(\delta;t)}\sum_{k=1}^{K}\left[\frac{1}{\sqrt{N}}\mathbb{G}_n^k\{\varphi(Z;\hat{\boldsymbol{\eta}}_{-k},\delta,t)-\varphi(Z;\boldsymbol{\eta},\delta,t)\}+\mathbb{P}\{\varphi(Z;\hat{\boldsymbol{\eta}}_{-k},\delta,t)-\varphi(Z;\boldsymbol{\eta},\delta,t)\}\right]\\
&\equiv B_{n,1}(\delta;t)+B_{n,2}(\delta;t).
\end{aligned}$$

Now we analyze the above two pieces $B_{n,1}(\delta;t)$ and $B_{n,2}(\delta;t)$. Showing $B_{n,1}(\delta;t) = o_{\mathbb{P}}(1)$ follows the exact same steps done by Kennedy [67]. However, analysis on $B_{n,2}(\delta;t)$ is largely different.

To analyze $B_{n,2}(\delta;t)$, we follow the same notation with that of Kennedy [67]. First let $\psi(\mathbb{P};Q)$ denote the mean outcome under intervention Q for a population corresponding to observed data distribution $\mathbb{P}$. Next, let denote $\varphi^*(z;\boldsymbol{\eta},t)$ its *centered* efficient influence function when Q does not depend on $\mathbb{P}$, as given in Lemma A.4.4 and let denote $\zeta^*(z;\boldsymbol{\eta},t)$ the contribution to the efficient influence function $\varphi^*(z;\boldsymbol{\eta},t)$ due to estimating Q when it depends on $\mathbb{P}$, as given in Lemma A.4.5. Now by definition,

$$\varphi(Z;\boldsymbol{\eta},\delta,t) = \varphi^*(Z;\boldsymbol{\eta},t) + \psi(\mathbb{P};Q) + \zeta^*(Z;\boldsymbol{\eta},t),$$

and thereby after some rearrangement we obtain

$$\begin{aligned}\frac{1}{\sqrt{n}}B_{n,2}(\delta;t) &= \mathbb{P}\{\varphi(Z;\overline{\boldsymbol{\eta}},\delta,t)-\varphi(Z;\boldsymbol{\eta},\delta,t)\}\\ &= \int \varphi^*(z;\overline{\boldsymbol{\eta}},t)d\mathbb{P}(z)+\psi(\overline{\mathbb{P}};\overline{Q})-\psi(\mathbb{P};\overline{Q})\\ &\quad+\int \zeta^*(z;\overline{\boldsymbol{\eta}},t)d\mathbb{P}(z)+\psi(\mathbb{P};\overline{Q})-\psi(\mathbb{P};Q).\end{aligned}$$

Although one can relate $\overline{\boldsymbol{\eta}}$ to $\widehat{\boldsymbol{\eta}}_{-k}$ in above equation, it can be anything associated with new $\overline{\mathbb{P}}$ and $\overline{Q}$.

Hence, by analyzing the second order remainder terms of von Mises expansion for the efficient influence functions given in Lemma A.4.4 and A.4.5, we can evaluate the convergence rate of $B_{n,2}(\delta;t)$. The following two lemmas analyze those second order remainder terms in the presence of censoring process.

Lemma A.4.8. *Let $\psi(\mathbb{P};Q)$ be a mean outcome under intervention Q for a for a population corresponding to observed data distribution $\mathbb{P}$, and let $\varphi^*(z;\boldsymbol{\eta},t)$ denote its efficient influence function when Q does not depend on $\mathbb{P}$ for given t, as given in Lemma A.4.4. For another data distribution $\overline{\mathbb{P}}$, let $\overline{\boldsymbol{\eta}}$ denote the corresponding nuisance functions. Then we have von Mises type expansion*

$$\begin{aligned}\psi(\overline{\mathbb{P}};Q)-\psi(\mathbb{P};Q) &= \int \varphi^*(z;\overline{\boldsymbol{\eta}},t)d\mathbb{P}(z)\\ &\quad+\sum_{t=1}^{2}\sum_{s=1}^{t}\int(m_t^*-\overline{m}_t)\left(\prod_{r=1}^{t}dQ_r d\mathbb{P}_r\right)\left(\frac{d\pi_s-d\overline{\pi}_s}{d\overline{\pi}_s}\right)\left(\frac{d\omega_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)\\ &\quad+\sum_{t=1}^{2}\sum_{s=1}^{t}\int(m_t^*-\overline{m}_t)\left(\prod_{r=1}^{t}dQ_r d\mathbb{P}_r\right)\left(\frac{d\omega_s-d\overline{\omega}_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)\end{aligned}$$

where we define

$$\overline{m}_t=\int\overline{m}_{t+1}dQ_{t+1}d\overline{\mathbb{P}}_{t+1},\qquad m_t^*=\int\overline{m}_{t+1}dQ_{t+1}d\mathbb{P}_{t+1},$$

$$dQ_t=dQ_t(A_t\mid H_t),\qquad d\pi_t=d\mathbb{P}(A_t\mid H_t),\qquad d\mathbb{P}_t=d\mathbb{P}(X_t\mid H_{t-1},A_{t-1}),$$

$$d\omega_s=d\mathbb{P}(R_{s+1}=1\mid H_s,A_s,R_s=1),\qquad d\overline{\omega}_s=d\overline{\mathbb{P}}(R_{s+1}=1\mid H_s,A_s,R_s=1).$$

Proof. From Lemma A.4.4, we have

$$
\begin{aligned}
\mathbb{E}\{\varphi^*(Z;\overline{\boldsymbol{\eta}})\} &= \sum_{t=0}^{t} \mathbb{E}\left\{\left(\int \overline{m}_{t+1} dQ_{t+1} - \overline{m}_t\right) \mathbb{1}(R_{t+1}=1) \prod_{s=0}^{t}\left(\frac{dQ_s}{d\pi_s}\frac{1}{d\varpi_s}\right)\right\} \\
&= \sum_{t=0}^{t} \mathbb{E}\left\{\mathbb{E}\left[\left(\int \overline{m}_{t+1} dQ_{t+1} - \overline{m}_t\right) \mathbb{1}(R_{t+1}=1)\mathbb{1}(R_t=1) \prod_{s=0}^{t}\left(\frac{dQ_s}{d\pi_s}\frac{1}{d\varpi_s}\right)\middle| H_t, A_t, R_t\right]\right\} \\
&= \sum_{t=0}^{t} \mathbb{E}\Bigg\{\mathbb{E}\left[\left(\int \overline{m}_{t+1} dQ_{t+1} - \overline{m}_t\right) \mathbb{1}(R_t=1) \prod_{s=0}^{t}\left(\frac{dQ_s}{d\pi_s}\frac{1}{d\varpi_s}\right)\middle| H_t, A_t, R_t=1, R_{t+1}=1\right] \\
&\qquad\qquad \times d\mathbb{P}(R_{t+1}=1 \mid H_t, A_t, R_t=1)\Bigg\} \\
&= \sum_{t=0}^{t} \mathbb{E}\left\{\left(\int\int \overline{m}_{t+1} dQ_{t+1} d\mathbb{P}_{t+1} - \overline{m}_t\right) \mathbb{1}(R_t=1) d\omega_t \prod_{s=0}^{t}\left(\frac{dQ_s}{d\pi_s}\frac{1}{d\varpi_s}\right)\right\} \\
&= \sum_{t=0}^{t} \mathbb{E}\left\{(m_t^* - \overline{m}_t)\, d\omega_t \mathbb{1}(R_t=1) \prod_{s=0}^{t}\left(\frac{dQ_s}{d\pi_s}\frac{1}{d\varpi_s}\right)\right\} \\
&= \sum_{t=0}^{t} \int (m_t^* - \overline{m}_t)\, d\omega_t \prod_{s=0}^{t}\left\{\left(\frac{dQ_s}{d\pi_s}\frac{1}{d\varpi_s}\right) d\pi_s d\mathbb{P}_s d\omega_{s-1}\right\}
\end{aligned}
$$

where the first equality follows by the definition and linearity of expectation, the second by iterated expectation and the equivalence between $\mathbb{1}(R_{t+1}=1)$ and $\mathbb{1}(R_{t+1}=1, R_t=1)$ [2], the third by the law of total probability on conditional expectation [3], the fourth by the result of Lemma A.4.1 (i.e. $d\mathbb{P}_{t+1} = d\mathbb{P}(X_{t+1} \mid H_t, A_t, R_{t+1}=1)$) and by the definition, and the fifth simply by definition. To obtain the last equality, we first apply iterated expectation conditioning on (H_t, R_t), then do another iterated expectation conditioning on $(H_{t-1}, A_{t-1}, R_{t-1})$ followed by same steps from the second, the third and the fourth equalities, and repeat these processes for $t-2, ..., 1$.

[2] For $\forall t$ the event $\{R_t = 1\}$ implies $\{R_s = 1 \text{ for all } s \leq t\}$ by construction.

[3] For random variable X, Y, Z, it follows $\mathbb{E}[X|Y] = \sum_z \mathbb{E}[X|Y, Z=z]\mathbb{P}(Z=z|Y)$.

From the last expression, now we have

$$\begin{aligned}
&\sum_{t=0}^{t}\int(m_t^*-\overline{m}_t)\prod_{s=0}^{t}\left\{\left(\frac{dQ_s}{d\pi_s}\frac{d\omega_s}{d\overline{\omega}_s}\right)d\pi_s d\mathbb{P}_s\right\}\\
&\quad=\sum_{t=0}^{t}\int(m_t^*-\overline{m}_t)\frac{d\pi_t}{d\overline{\pi}_t}\frac{d\omega_t}{d\overline{\omega}_t}dQ_t d\mathbb{P}_t\prod_{s=0}^{t-1}\left\{\left(\frac{dQ_s}{d\pi_s}\frac{d\omega_s}{d\overline{\omega}_s}\right)d\pi_s d\mathbb{P}_s\right\}\\
&\quad=\sum_{t=0}^{t}\int(m_t^*-\overline{m}_t)\left(\frac{d\pi_t-d\overline{\pi}_t}{d\overline{\pi}_t}\right)\frac{d\omega_t}{d\overline{\omega}_t}dQ_t d\mathbb{P}_t\prod_{s=0}^{t-1}\left\{\left(\frac{dQ_s}{d\pi_s}\frac{d\omega_s}{d\overline{\omega}_s}\right)d\pi_s d\mathbb{P}_s\right\}\\
&\qquad+\sum_{t=0}^{t}\int(m_t^*-\overline{m}_t)\frac{d\omega_t}{d\overline{\omega}_t}dQ_t d\mathbb{P}_t\prod_{s=0}^{t-1}\left\{\left(\frac{dQ_s}{d\pi_s}\frac{d\omega_s}{d\overline{\omega}_s}\right)d\pi_s d\mathbb{P}_s\right\}\\
&\quad=\sum_{t=1}^{t}\int(m_t^*-\overline{m}_t)\left(\frac{d\pi_t-d\overline{\pi}_t}{d\overline{\pi}_t}\right)\frac{d\omega_t}{d\overline{\omega}_t}dQ_t d\mathbb{P}_t\prod_{s=0}^{t-1}\left\{\left(\frac{dQ_s}{d\pi_s}\frac{d\omega_s}{d\overline{\omega}_s}\right)d\pi_s d\mathbb{P}_s\right\}\\
&\qquad+\sum_{t=1}^{t}\int(m_t^*-\overline{m}_t)\left(\frac{d\omega_t-d\overline{\omega}_t}{d\overline{\omega}_t}\right)dQ_t d\mathbb{P}_t\prod_{s=0}^{t-1}\left\{\left(\frac{dQ_s}{d\pi_s}\frac{d\omega_s}{d\overline{\omega}_s}\right)d\pi_s d\mathbb{P}_s\right\}\\
&\qquad+\sum_{t=1}^{t}\int(m_t^*-\overline{m}_t)dQ_t d\mathbb{P}_t\prod_{s=0}^{t-1}\left\{\left(\frac{dQ_s}{d\pi_s}\frac{d\omega_s}{d\overline{\omega}_s}\right)d\pi_s d\mathbb{P}_s\right\}+(m_0^*-\overline{m}_0)
\end{aligned}$$

, where all the algebras are basically adding and subtracting the same term after some rearrangement. Note that we use the convention from earlier lemmas that all the quantities with negative times such as dQ_{-1} are set to one. If we repeat above process t times we obtain the following identity.

$$\begin{aligned}
&\sum_{t=0}^{t}\int(m_t^*-\overline{m}_t)\prod_{s=0}^{t}\left\{\left(\frac{dQ_s}{d\pi_s}\frac{d\omega_s}{d\overline{\omega}_s}\right)d\pi_s d\mathbb{P}_s\right\}\\
&\quad=\sum_{t=1}^{t}\sum_{s=1}^{t}\int(m_t^*-\overline{m}_t)\left(\prod_{r=s}^{t}dQ_r d\mathbb{P}_r\right)\left(\frac{d\pi_s-d\overline{\pi}_s}{d\overline{\pi}_s}\right)\frac{d\omega_s}{d\overline{\omega}_s}\prod_{r=1}^{s-1}\left\{\left(\frac{dQ_r}{d\pi_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)d\pi_r d\mathbb{P}_r\right\}\\
&\qquad+\sum_{t=1}^{t}\sum_{s=1}^{t}\int(m_t^*-\overline{m}_t)\left(\prod_{r=s}^{t}dQ_r d\mathbb{P}_r\right)\left(\frac{d\omega_s-d\overline{\omega}_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left\{\left(\frac{dQ_r}{d\pi_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)d\pi_r d\mathbb{P}_r\right\}\\
&\qquad+\sum_{t=1}^{t}\int(m_t^*-\overline{m}_t)\left(\prod_{s=1}^{t}dQ_s d\mathbb{P}_s\right)+(m_0^*-\overline{m}_0)
\end{aligned}$$

However, by last part of Lemma 5 in Kennedy [67] we have

$$\sum_{t=1}^{t}\int(m_t^*-\overline{m}_t)\left(\prod_{s=1}^{t}dQ_s d\mathbb{P}_s\right)=m_0-m_0^*.$$

Putting all these together, after some rearranging finally we have

$$
\begin{aligned}
\mathbb{E}\{\varphi^*(Z;\overline{\boldsymbol{\eta}})\} &= m_0 - \overline{m}_0 \\
&\quad + \sum_{t=1}^{t}\sum_{s=1}^{t}\int (m_t^* - \overline{m}_t)\left(\prod_{r=1}^{t} dQ_r d\mathbb{P}_r\right)\left(\frac{d\pi_s - d\overline{\pi}_s}{\overline{\pi}_s}\right)\left(\frac{d\omega_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right) \\
&\quad + \sum_{t=1}^{t}\sum_{s=1}^{t}\int (m_t^* - \overline{m}_t)\left(\prod_{r=1}^{t} dQ_r d\mathbb{P}_r\right)\left(\frac{d\omega_s - d\overline{\omega}_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)
\end{aligned}
$$

, which yields the formula we have in Lemma A.4.8. □

Lemma A.4.9. *Let $\zeta^*(z;\overline{\boldsymbol{\eta}},t)$ denote the contribution to the efficient influence function $\varphi^*(z;\boldsymbol{\eta},t)$ due to dependence between $\mathbb{P}$ and Q as given in Lemma A.4.5. Then for two different intervention distributions Q and $\overline{Q}$ whose corresponding densities are dQ_t and $d\overline{Q}_t$ respectively with respect to some dominating measure for $t = 1,...,t$, we have von Mises type expansion*

$$
\begin{aligned}
\psi(\mathbb{P};\overline{Q}) - \psi(\mathbb{P};Q) &= \int \zeta^*(z;\overline{\boldsymbol{\eta}},t)d\mathbb{P}(z) \\
&\quad + \sum_{t=1}^{t}\int \overline{\phi}_t d\pi_t(m_t - \overline{m}_t)d\nu d\mathbb{P}_t\prod_{s=0}^{t-1}\left(\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right)d\pi_s d\mathbb{P}_s d\omega_s \\
&\quad + \sum_{t=1}^{t}\sum_{s=1}^{t}\int \overline{\phi}_t d\pi_t m_t d\nu d\mathbb{P}_t\left(\prod_{r=0}^{t-1} d\overline{Q}_r d\mathbb{P}_r\right)\left(\frac{d\overline{\pi}_s - d\pi_s}{d\overline{\pi}_s}\right)\left(\frac{d\omega_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right) \\
&\quad + \sum_{t=1}^{t}\sum_{s=1}^{t}\int \overline{\phi}_t d\pi_t m_t d\nu d\mathbb{P}_t\left(\prod_{r=0}^{t-1} d\overline{Q}_r d\mathbb{P}_r\right)\left(\frac{d\overline{\omega}_s - d\omega_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right) \\
&\quad + \sum_{t=1}^{t}\int m_t\left(d\overline{Q}_t - dQ_t - \overline{\phi}_t d\pi_t d\nu\right)d\mathbb{P}_t\left(\prod_{s=0}^{t-1} d\overline{Q}_s d\mathbb{P}_s\right)
\end{aligned}
$$

where we define all the notation in the same way in Lemma A.4.8.

Proof. From Lemma 6 in Kennedy [67] and by Lemma A.4.1, we have

$$
\begin{aligned}
\Psi(\mathbb{P};\overline{Q}) - \Psi(\mathbb{P};Q) &= \int m_T\left(\prod_{t=1}^{T} d\overline{Q}_t d\mathbb{P}_t - \prod_{t=1}^{T} dQ_t d\mathbb{P}_t\right) \\
&= \sum_{t=1}^{t}\int m_t\left(d\overline{Q}_t - dQ_t\right)d\mathbb{P}_t\prod_{s=0}^{t-1} d\overline{Q}_s d\mathbb{P}_s.
\end{aligned}
$$

Next, for the expected contribution to the influence function due to estimating Q when it depends on $\mathbb{P}$, we have that

$$\begin{aligned}
\mathbb{E}[\zeta^*(Z;\boldsymbol{\eta})] &= \mathbb{E}\left[\sum_{t=1}^{t}\int \overline{\phi}_t \overline{m}_t d\nu \left(\prod_{s=0}^{t-1}\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right)\mathbb{1}(R_t=1)\right] \\
&= \sum_{t=1}^{t}\mathbb{E}\left[\int \overline{\phi}_t d\pi_t \overline{m}_t d\nu \left(\prod_{s=0}^{t-1}\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right)\mathbb{1}(R_t=1)\mathbb{1}(R_{t-1}=1)\right] \\
&= \sum_{t=1}^{t}\mathbb{E}\left\{\left[\int \overline{\phi}_t d\pi_t \overline{m}_t d\nu d\mathbb{P}_t \left(\prod_{s=0}^{t-1}\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right)\mathbb{1}(R_{t-1}=1)\right] d\mathbb{P}(R_t=1\mid H_{t-1},A_{t-1},R_{t-1}=1)\right\} \\
&= \sum_{t=1}^{t}\mathbb{E}\left\{\int \overline{\phi}_t d\pi_t \overline{m}_t d\nu d\mathbb{P}_t \prod_{s=0}^{t-1}\left(\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right) d\omega_{t-1}\mathbb{1}(R_{t-1}=1)\right\} \\
&= \sum_{t=1}^{t}\int \overline{\phi}_t d\pi_t \overline{m}_t d\nu d\mathbb{P}_t \prod_{s=0}^{t-1}\left(\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right) d\pi_s d\mathbb{P}_s d\omega_s
\end{aligned}$$

where the first equality by definition, the second by iterated expectation conditioning on (H_t,R_t) and equivalence between $\mathbb{1}(R_t=1)\mathbb{1}(R_{t-1}=1)$ and $\mathbb{1}(R_t=1)$, the third by iterated expectation conditioning on $(H_{t-1},A_{t-1},R_{t-1})$ and law of total probability, and the fifth by repeating the process T times. Details follow almost the same logic as in Lemma A.4.8.

Now, we further expand our last expression as

$$\begin{aligned}
&\sum_{t=1}^{t}\int \overline{\phi}_t d\pi_t \overline{m}_t d\nu d\mathbb{P}_t \prod_{s=0}^{t-1}\left(\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right) d\pi_s d\mathbb{P}_s d\omega_s \\
&\quad= \sum_{t=1}^{t}\int \overline{\phi}_t d\pi_t (\overline{m}_t - m_t) d\nu d\mathbb{P}_t \prod_{s=0}^{t-1}\left(\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right) d\pi_s d\mathbb{P}_s d\omega_s \\
&\qquad+ \sum_{t=1}^{t}\int \overline{\phi}_t d\pi_t m_t d\nu d\mathbb{P}_t \prod_{s=0}^{t-1}\left(\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right) d\pi_s d\mathbb{P}_s d\omega_s \\
&\quad= \sum_{t=1}^{t}\int \overline{\phi}_t d\pi_t (\overline{m}_t - m_t) d\nu d\mathbb{P}_t \prod_{s=0}^{t-1}\left(\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right) d\pi_s d\mathbb{P}_s d\omega_s \\
&\qquad+ \sum_{t=1}^{t}\sum_{s=1}^{t}\int \overline{\phi}_t d\pi_t m_t d\nu d\mathbb{P}_t \left(\prod_{r=0}^{t-1} d\overline{Q}_r d\mathbb{P}_r\right)\left(\frac{d\pi_s - d\overline{\pi}_s}{d\overline{\pi}_s}\right)\left(\frac{d\omega_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right) \\
&\qquad+ \sum_{t=1}^{t}\sum_{s=1}^{t}\int \overline{\phi}_t d\pi_t m_t d\nu d\mathbb{P}_t \left(\prod_{r=0}^{t-1} d\overline{Q}_r d\mathbb{P}_r\right)\left(\frac{d\omega_s - d\overline{\omega}_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right) \\
&\qquad+ \sum_{t=1}^{t}\int \overline{\phi}_t d\pi_t m_t d\nu d\mathbb{P}_t \left(\prod_{s=0}^{t-1} d\overline{Q}_s d\mathbb{P}_s\right)
\end{aligned}$$

where the first equality follows by adding and subtracting the second term, an the second by the same steps used in Lemma A.4.8.

With the last term in the last expression above, it follows

$$\begin{aligned}\Psi(\mathbb{P};\overline{Q}) - \Psi(\mathbb{P};Q) - \sum_{t=1}^{t} \int \overline{\phi}_t d\pi_t m_t d\nu d\mathbb{P}_t \left(\prod_{s=0}^{t-1} d\overline{Q}_s d\mathbb{P}_s \right) \\ = \sum_{t=1}^{t} \int m_t \left(d\overline{Q}_t - dQ_t - \overline{\phi}_t d\pi_t d\nu \right) d\mathbb{P}_t \left(\prod_{s=0}^{t-1} d\overline{Q}_s d\mathbb{P}_s \right).\end{aligned}$$

Putting these all together, finally we have

$$\begin{aligned}\Psi(\mathbb{P};\overline{Q}) - \Psi(\mathbb{P};Q) &= \mathbb{E}[\zeta^*(Z;\overline{\boldsymbol{\eta}})] \\ &+ \sum_{t=1}^{t} \int \overline{\phi}_t d\pi_t (m_t - \overline{m}_t) d\nu d\mathbb{P}_t \prod_{s=0}^{t-1} \left(\frac{d\overline{Q}_s}{d\overline{\pi}_s} \frac{1}{d\overline{\omega}_s} \right) d\pi_s d\mathbb{P}_s d\omega_s \\ &+ \sum_{t=1}^{t} \sum_{s=1}^{t} \int \overline{\phi}_t d\pi_t m_t d\nu d\mathbb{P}_t \left(\prod_{r=0}^{t-1} d\overline{Q}_r d\mathbb{P}_r \right) \left(\frac{d\overline{\pi}_s - d\pi_s}{d\overline{\pi}_s} \right) \left(\frac{d\omega_s}{d\overline{\omega}_s} \right) \prod_{r=1}^{s-1} \left(\frac{d\pi_r}{d\overline{\pi}_r} \frac{d\omega_r}{d\overline{\omega}_r} \right) \\ &+ \sum_{t=1}^{t} \sum_{s=1}^{t} \int \overline{\phi}_t d\pi_t m_t d\nu d\mathbb{P}_t \left(\prod_{r=0}^{t-1} d\overline{Q}_r d\mathbb{P}_r \right) \left(\frac{d\overline{\omega}_s - d\omega_s}{d\overline{\omega}_s} \right) \prod_{r=1}^{s-1} \left(\frac{d\pi_r}{d\overline{\pi}_r} \frac{d\omega_r}{d\overline{\omega}_r} \right) \\ &+ \sum_{t=1}^{t} \int m_t \left(d\overline{Q}_t - dQ_t - \overline{\phi}_t d\pi_t d\nu \right) d\mathbb{P}_t \left(\prod_{s=0}^{t-1} d\overline{Q}_s d\mathbb{P}_s \right)\end{aligned}$$

which is the result of the lemma. □

Finally, the next Lemma concludes the proof of the second statement and thus completes the proof of the Theorem 2.5.1. In fact, it is this lemma that substantiates why having all nuisance functions estimated at rate of $n^{-1/4}$ can be one sufficient condition.

Lemma A.4.10. *Remainders of the von Mises expansion from Lemma A.4.8 and A.4.9 are both diminishing at rate of* $n^{-\frac{1}{2}}$ *uniformly in* δ*, if*

$$\left(\underset{\delta \in \mathscr{D}}{sup} \| m_{\delta,t} - \widehat{m}_{\delta,t} \mid + \mid \pi_t - \widehat{\pi}_t \mid \right) \left(\mid \pi_s - \overline{\pi}_s \| + \mid \omega_s - \overline{\omega}_s \| \right) = o_{\mathbb{P}} \left(\frac{1}{\sqrt{n}} \right),$$

for $\forall s \leq t \leq T$*.*

Proof. The remainder term of the Von Mises type expansion from Lemma A.4.8 equals

$$\sum_{t=1}^{t}\sum_{s=1}^{t}\int(m_t^*-\overline{m}_t)\left(\prod_{r=1}^{t}dQ_rd\mathbb{P}_r\right)\left(\frac{d\pi_s-d\overline{\pi}_s}{d\overline{\pi}_s}\right)\left(\frac{d\omega_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)$$

$$+\sum_{t=1}^{t}\sum_{s=1}^{t}\int(m_t^*-\overline{m}_t)\left(\prod_{r=1}^{t}dQ_rd\mathbb{P}_r\right)\left(\frac{d\omega_s-d\overline{\omega}_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)$$

$$=\sum_{t=1}^{t}\sum_{s=1}^{t}\int\Big\{(\overline{m}_{t+1}-m_{t+1})dQ_{t+1}d\mathbb{P}_{t+1}+(m_t-\overline{m}_t)\Big\}\left(\prod_{r=1}^{t}dQ_rd\mathbb{P}_r\right)\left(\frac{d\pi_s-d\overline{\pi}_s}{d\overline{\pi}_s}\right)\left(\frac{d\omega_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega}{d\overline{\omega}}\right.$$

$$+\sum_{t=1}^{t}\sum_{s=1}^{t}\int\Big\{(\overline{m}_{t+1}-m_{t+1})dQ_{t+1}d\mathbb{P}_{t+1}+(m_t-\overline{m}_t)\Big\}\left(\prod_{r=1}^{t}dQ_rd\mathbb{P}_r\right)\left(\frac{d\omega_s-d\overline{\omega}_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)$$

$$\lesssim\sum_{t=1}^{t}\Big(\mid\overline{m}_{t+1}-m_{t+1}\mid+\mid m_t-\overline{m}_t\mid\Big)\sum_{s=1}^{t}\Big(\mid\pi_s-\overline{\pi}_s\|+\mid\omega_s-\overline{\omega}_s\|\Big)$$

where we obtain the first inequality simply by adding and subtracting m_t.

For the remainder term from Lemma A.4.9, first note that by Lemma A.4.1 the following results stated in Kennedy [67] also holds for our case:

$$\int\overline{\phi}_td\pi_t=\frac{\delta(2a_t-1)(\pi_t-\overline{\pi}_t)}{(\delta\overline{\pi}_t+1-\overline{\pi}_t)^2},$$

$$d\overline{Q}_t-dQ_t-\int\overline{\phi}_td\pi_t=\frac{\delta(\delta-1)(2a_t-1)(\overline{\pi}_t-\pi_t)^2}{(\delta\overline{\pi}_t+1-\overline{\pi}_t)^2(\delta\pi_t+1-\pi_t)}.$$

where we additionally condition $R_t=1$ for $\pi_t,\overline{\pi}_t$ in our case. Hence, it immediately follows that the remainder from Lemma A.4.9 is

$$\sum_{t=1}^{t}\int\overline{\phi}_td\pi_t(m_t-\overline{m}_t)d\nu d\mathbb{P}_t\prod_{s=0}^{t-1}\left(\frac{d\overline{Q}_s}{d\overline{\pi}_s}\frac{1}{d\overline{\omega}_s}\right)d\pi_sd\mathbb{P}_sd\omega_s$$

$$+\sum_{t=1}^{t}\sum_{s=1}^{t}\int\overline{\phi}_td\pi_tm_td\nu d\mathbb{P}_t\left(\prod_{r=0}^{t-1}d\overline{Q}_rd\mathbb{P}_r\right)\left(\frac{d\overline{\pi}_s-d\pi_s}{d\overline{\pi}_s}\right)\left(\frac{d\omega_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)$$

$$+\sum_{t=1}^{t}\sum_{s=1}^{t}\int\overline{\phi}_td\pi_tm_td\nu d\mathbb{P}_t\left(\prod_{r=0}^{t-1}d\overline{Q}_rd\mathbb{P}_r\right)\left(\frac{d\overline{\omega}_s-d\omega_s}{d\overline{\omega}_s}\right)\prod_{r=1}^{s-1}\left(\frac{d\pi_r}{d\overline{\pi}_r}\frac{d\omega_r}{d\overline{\omega}_r}\right)$$

$$+\sum_{t=1}^{t}\int m_t\left(d\overline{Q}_t-dQ_t-\overline{\phi}_td\pi_td\nu\right)d\mathbb{P}_t\left(\prod_{s=0}^{t-1}d\overline{Q}_sd\mathbb{P}_s\right)$$

$$\lesssim\sum_{t=1}^{t}\mid\pi_t-\overline{\pi}_t\mid\left\{\mid m_t-\overline{m}_t\mid+\sum_{s=1}^{t}\Big(\mid\pi_s-\overline{\pi}_s\|+\mid\omega_s-\overline{\omega}_s\|\Big)+\mid\pi_t-\overline{\pi}_t\mid\right\}.$$

Therefore, supported by the condition 4) in Theorem 2.5.1, if we have

$$\left(\underset{\delta\in\mathscr{D}}{sup}\|m_{\delta,t} - \widehat{m}_{\delta,t} \mid + \mid \pi_t - \widehat{\pi}_t \mid\right)\left(\mid \pi_s - \overline{\pi}_s\| + \mid \omega_s - \overline{\omega}_s\|\right) = o_{\mathbb{P}}(\frac{1}{\sqrt{n}}),$$

for $\forall s \leq t \leq t$, both of the remainders from Lemma A.4.8 and A.4.9 are diminishing at rate of $n^{-\frac{1}{2}}$ uniformly in δ. □

A.4.8 Rationality of using multiplier bootstrap from [67]

As in the proof of Theorem 2.5.1, we let

Proof.

$$\|f\mid_{\mathscr{D},\mathscr{T}} \equiv \sup_{\delta\in\mathscr{D},t\in\mathscr{T}} |f(\delta,t)\mid$$

and define the processes

$$\widehat{\Psi}_n(\delta,t) \equiv\sqrt{n}\{\widehat{\psi}_t(\delta) - \psi_t(\delta)\}/\widehat{\sigma}(\delta,t)$$

$$\widehat{\Psi}^*_n(\delta,t) \equiv \mathbb{G}_n[\varepsilon\{\varphi(Z;\hat{\boldsymbol{\eta}}_{-S},\delta,t) - \widehat{\psi}_t(\delta)\}/\widehat{\sigma}(\delta,t)]$$

$$\Psi^*_n(\delta,t) \equiv \mathbb{G}_n[\varepsilon\{\varphi(Z;\boldsymbol{\eta},\delta,t) - \psi(t;\delta)\}/\sigma(\delta,t)]$$

where we let the star superscripts denote multiplier bootstrap processes defined in Theorem 4 of Kennedy [67] and let $\mathbb{G}$ be a mean-zero Gaussian process with covariance $\mathbb{E}[\mathbb{G}(\delta_1;t)\mathbb{G}(\delta_2;t)] = \mathbb{E}[\widetilde{\varphi}(Z;\boldsymbol{\eta},\delta_1,t_1)\widetilde{\varphi}(Z;\boldsymbol{\eta},\delta_2,t_2)]$ as defined in Theorem 2.5.1 in the main text.

From above setup and the result of Theorem 2.5.1 it only requires to show

$$\left|\mathbb{P}\left(\mid \widehat{\Psi}_n \mid_{\mathscr{D},\mathscr{T}} \leq \hat{c}_\alpha\right) - \mathbb{P}\left(\mid \widehat{\Psi}^*_n \mid_{\mathscr{D},\mathscr{T}} \leq \hat{c}_\alpha\right)\right| = o(1),$$

since $\mathbb{P}\left(\mid \widehat{\Psi}^*_n \mid_{\mathscr{D},\mathscr{T}} \leq \hat{c}_\alpha\right) = 1-\alpha$ by definition. The proof is very straightforward since we already have shown $\|\widehat{\Psi}_n - \Psi_n \mid_{\mathscr{D},\mathscr{T}} = o_{\mathbb{P}}(1)$ in the proof of Theorem 2.5.1, which implies that $\left|\mid \widehat{\Psi}^*_n \mid_{\mathscr{D},\mathscr{T}} - \mid \Psi^*_n \mid_{\mathscr{D},\mathscr{T}}\right| = o_{\mathbb{P}}(1)$. Furthermore since we are adding only finite number of discrete timepoints into the function class used in the proof of Theorem 4 in Kennedy [67], Lemma 2.3. in Chernozhukov et al. [20] and Corollary 2.2 in Belloni et al. [9] are still valid in our case and thereby the exact same argument used in the proof of Theorem 4 in Kennedy [67] follows to conclude the above statement.

□

Appendix B

Supplementary Materials for Chapter 3

B.1 Simulations: supplementary materials

Summary of variables

Variable labels in the following tables are exactly match with the one in the SEDA archive at *Stanford Center for Education Policy Analysis* [1]. In the end, I have on average 1035 samples (the number of districts) each year. In the following tables, one can find list of covariates (X) used in the simulation and also description of treatment (A) and outcome (Y) variables.

Group	Variable Labels	Varies by grade?	Varies by year?
Baseline Covariates	`fips, baplus_all, poverty517_all, singmom_all, snap_all, samehouse_all, unemp_all, inc50all, giniall, baplus_mal, baplus_fem, pov_mal, pov_fem, teenbirth_all`	No	No
Time-varying Covariates	`nsch, speced, tottch, aides, diffstutch_hspwht, diffstutch_blkwht`	No	Yes

Table B.1 Summary of covariates

[1] https://cepa.stanford.edu/seda/papers

Group	Variable Labels	Description

Treatment (A)	`flunch_hsp`
	`, flunch_blk`
		Percent free lunch in average {Black, Hispanic} student's school

Outcome (Y)	`White.Black.ELA.Gap,`
	`White.Hispanic.ELA.Gap,`
	`White.Black.Math.Gap,`
	`White.Hispanic.Math.Gap`
		Test score gaps between white and {Black, Hispanic} students in English/Language Arts (ELA) and Math standardized assessment outcomes in grades 3 to 8. We render it binary by using their average (i.e. 1 if > mean).

Table B.2 Summary of treatment and outcome

Additional results

For completeness of section 3.5.3, we attach additional simulation results for year 2010-2012. Original data in SEDA ranges from 2009-2013, but we found that there are some unusual outliers and a number of samples is particularly also very small for year 2013. Hence we exclude year 2013 and conduct the same simulation with rest of the years. Here $\hat{\psi}_{\text{DD}}$ indicates estimated value of our proposed estimator (difference-in-distribution) for observational study.

	White-Black		White-Hispanic	
Estimator	Math	ELA	Math	ELA
$\hat{\psi}_{\text{pi}}$	-0.057 $(-0.098,-0.016)$	-0.041 $(-0.075,-0.006)$	-0.041 $(-0.087,0.006)$	-0.028 $(-0.061,0.006)$
$\hat{\phi}_{\text{IPW}}$	-0.040 $(-0.089,0.009)$	-0.032 $(-0.055,-0.009)$	-0.032 $(-0.065,0.002)$	-0.033 $(-0.069,0.004)$
$\hat{\psi}_{\text{DR}}$	-0.053 $(-0.064,-0.045)$	-0.055 $(-0.067,-0.041)$	-0.029 $(-0.040,-0.019)$	-0.050 $(-0.076,-0.025)$
$\hat{\psi}_{\text{DD}}$	**0.812** (0.768,0.856)	**0.783** (0.740,0.822)	**0.658** (0.620,0.695)	**0.802** (0.775,0.830

Table B.3 Estimated causal effect of free lunch on test gaps in 2010 (with 95% CI)

	White-Black		White-Hispanic	
Estimator	Math	ELA	Math	ELA
$\hat{\psi}_{\text{pi}}$	-0.032 $(-0.076,0.012)$	-0.029 $(-0.047,-0.010)$	-0.052 $(-0.097,-0.007)$	-0.025 $(-0.054,0.005)$
$\hat{\phi}_{\text{IPW}}$	-0.012 $(-0.038,0.005)$	-0.019 $(-0.039,0.002)$	-0.022 $(-0.059,0.015)$	0.011 $(-0.025,0.048)$
$\hat{\psi}_{\text{DR}}$	-0.003 $(-0.009,0.016)$	-0.060 $(-0.070,-0.050)$	-0.036 $(-0.049,-0.023)$	0.038 (0.026,0.048)
$\hat{\psi}_{\text{DD}}$	**0.752** (0.725,0.780)	**0.538** (0.510,0.555)	**0.702** (0.665,0.730)	**0.359** (0.334,0.385)

Table B.4 Estimated causal effect of free lunch on test gaps in 2011 (with 95% CI)

	White-Black		White-Hispanic	
Estimator	Math	ELA	Math	ELA
$\hat{\psi}_{\text{pi}}$	-0.039 $(-0.087,0.016)$	-0.030 $(-0.060,0.001)$	-0.046 $(-0.077,-0.016)$	-0.039 $(-0.072,-0.006)$
$\hat{\phi}_{\text{IPW}}$	-0.037 $(-0.091,0.018)$	-0.022 $(-0.067,0.023)$	-0.019 $(-0.055,0.035)$	0.019 $(-0.018,0.057)$
$\hat{\psi}_{\text{DR}}$	-0.068 $(-0.079,-0.057)$	-0.066 $(-0.077,-0.056)$	-0.080 $(-0.092,-0.069)$	0.022 $(-0.001,0.045)$
$\hat{\psi}_{\text{DD}}$	**0.798** (0.749,0.847)	**0.723** (0.681,0.745)	**0.658** (0.629,0.694)	**0.746** (0.723,0.770)

Table B.5 Estimated causal effect of free lunch on test gaps in 2012 (with 95% CI)

B.2 Proofs

In every proof, all the constants are only defined locally unless a connection to the one in the main paper is explicitly stated.

B.2.1 Proof of Proposition 3.3.1

Proof. We let $T_{h,y}(Y) = \frac{1}{h^d}K\left(\frac{\|y-Y\|_2}{h}\right)$ as previously. Then

$$\begin{aligned}
\mathbb{E}[\widehat{q_h^a} \mid A_i = a, \forall i] &= \mathbb{E}\left[\frac{\mathbb{1}(n_a > 0)}{n_a}\sum_{j=1}^{n} T_{h,y}(Y_j)\mathbb{1}(A_j = a) \mid A_i = a, \forall i\right] \\
&= \frac{1}{n_a}\sum_{j=1}^{n} \mathbb{1}(A_j = a)\mathbb{E}[T_{h,y}(Y_j) \mid A_i = a, \forall i] \\
&= \frac{1}{n_a}\sum_{j=1}^{n_a} \mathbb{E}[T_{h,y}(Y_j) \mid A_j = a] \\
&= \frac{1}{n_a}\sum_{j=1}^{n_a} \mathbb{E}[T_{h,y}(Y_j^a) \mid A_j = a] \\
&= \frac{1}{n_a}\sum_{j=1}^{n_a} \mathbb{E}[T_{h,y}(Y_j^a)] = \mathbb{E}[T_{h,y}(Y^a)],
\end{aligned}$$

where the fourth and the fifth equalities follow by assumption (C1) and (C2) respectively. □

B.2.2 Proof of Lemma 3.3.1

Lemma B.2.1. *Under the assumption (A1), (A2),*

$$\mathbb{E}\left[|\hat{q}_h^a(y) - q_h^a(y)|\right] \leq \frac{C_{K,q_{\max}}}{\sqrt{n\pi_a h^d}},$$

where $C_{K,q_{\max}}$ is a constant depending only on $\|K\|_2$ and $q_{\max}$.

Proof. Recall that $\hat{q}^a_h(y)$ is defined by

$$\hat{q}^a_h(y) = \frac{\sum_{i=1}^n T_{h,y}(Y_i)\mathbb{1}(A_i = a)}{\sum_{i=1}^n \mathbb{1}(A_i = a)}\mathbb{1}(n_a > 0)$$

where we let $T_{h,y}(Y) = \frac{1}{h_a^d}K\left(\frac{||y-Y||_2}{h_a}\right)$ as previously. Then $(\hat{q}^a_h(y) - q^a_h(y))^2$ is expanded as

$$\begin{aligned}(\hat{q}^a_h(y) - q^a_h(y))^2 &= \left(\frac{\sum_{i=1}^n T_{h,y}(Y_i)\mathbb{1}(A_i = a)}{\sum_{i=1}^n \mathbb{1}(A_i = a)}\mathbb{1}(n_a > 0) - q^a_h(y)\right)^2 \\ &= \left(\frac{\sum_{i=1}^n (T_{h,y}(Y_i) - q^a_h(y))\mathbb{1}(A_i = a)}{\sum_{i=1}^n \mathbb{1}(A_i = a)}\right)^2 \mathbb{1}(n_a > 0) + q^a_h(y)^2\mathbb{1}(n_a = 0).\end{aligned} \tag{B.1}$$

Consider the first term of (B.1). Conditioned on $\tilde{A} =: A_1, ..., A_n$, its expectation can be expanded as

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{\sum_{i=1}^n (T_{h,y}(Y_i) - q^a_h(y))\mathbb{1}(A_i = a)}{\sum_{i=1}^n \mathbb{1}(A_i = a)}\right)^2 \mathbb{1}(n_a > 0) \mid \tilde{A}\right] \\ &= \frac{\mathbb{E}\left[\left(\sum_{i=1}^n (T_{h,y}(Y_i) - q^a_h(y))\mathbb{1}(A_i = a)\right)^2 \mid \tilde{A}\right]}{(\sum_{i=1}^n \mathbb{1}(A_i = a))^2}\mathbb{1}(n_a > 0) \\ &= \frac{\sum_{i=1}^n \mathbb{E}\left[(T_{h,y}(Y_i) - q^a_h(y))^2 \mid A_i = a\right]}{(\sum_{i=1}^n \mathbb{1}(A_i = a))^2}\mathbb{1}(n_a > 0) \\ &+ \frac{\sum_{i \neq j} \mathbb{E}\left[(T_{h,y}(Y_i) - q^a_h(y))(T_{h,y}(Y_j) - q^a_h(y))\mathbb{1}(A_i = a)\mathbb{1}(A_j = a) \mid \tilde{A}\right]}{(\sum_{i=1}^n \mathbb{1}(A_i = a))^2}\mathbb{1}(n_a > 0).\end{aligned} \tag{B.2}$$

Then it follows

$$\begin{aligned}\mathbb{E}\left[(T_{h,y}(Y_i) - q^a_h(y))^2 \mid A_i = a\right] &= \mathbb{E}\left[(T_{h,y}(Y^a_i) - q^a_h(y))^2\right] \\ &= Var(T_{h,y}(Y_i)) \leq \frac{q_{\max}\|K\|_2^2}{h^d}\end{aligned} \tag{B.3}$$

where the first equality follows by (C1) and (C2) and the last inequality by Proposition 1.1 of [131].

Similarly, for any $i \neq j$, we have

$$\begin{aligned}\mathbb{E}\big[(T_{h,y}(Y_i)-q_h^a(y))(T_{h,y}(Y_j)-q_h^a(y))\mathbb{1}(A_i=a)\mathbb{1}(A_j=a)\mid \tilde{A}\big] &= \mathbb{E}\big[(T_{h,y}(Y_i)-q_h^a(y))(T_{h,y}(Y_j)-q_h^a(y))\mid A \\ &= \mathbb{E}\big[(T_{h,y}(Y_i^a)-q_h^a(y))(T_{h,y}(Y_j^a)-q_h^a(y))\big] \\ &= 0, \qquad\qquad\qquad\qquad (B.4)\end{aligned}$$

where the last equality follows by the identity $\mathbb{E}[T_{h,y}(Y^a)] = q_h^a(y)$. Hence applying (B.3) and (B.4) to (B.2) gives

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{\sum_{i=1}^n (T_{h,y}(Y_i)-q_h^a(y))\mathbb{1}(A_i=a)}{\sum_{i=1}^n \mathbb{1}(A_i=a)}\right)^2 \mathbb{1}(n_a>0)\mid \tilde{A}\right] \\ &\leq \frac{\sum_{i=1}^n \frac{q_{\max}\|K\|_2^2}{h^d}\mathbb{1}(A_i=a)}{(\sum_{i=1}^n \mathbb{1}(A_i=a))^2}\mathbb{1}(n_a>0) \\ &= \frac{q_{\max}\|K\|_2^2}{h^d}\frac{\mathbb{1}(n_a>0)}{\sum_{i=1}^n \mathbb{1}(A_i=a)}. \end{aligned} \tag{B.5}$$

Now, by the law of total expectation we have

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{\sum_{i=1}^n (T_{h,y}(Y_i)-q_h^a(y))\mathbb{1}(A_i=a)}{\sum_{i=1}^n \mathbb{1}(A_i=a)}\right)^2 \mathbb{1}(n_a>0)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{\sum_{i=1}^n (T_{h,y}(Y_i)-q_h^a(y))\mathbb{1}(A_i=a)}{\sum_{i=1}^n \mathbb{1}(A_i=a)}\right)^2 \mathbb{1}(n_a>0)\mid \tilde{A}\right]\right] \\ &\leq \frac{q_{\max}\|K\|_2^2}{h^d}\mathbb{E}\left[\frac{\mathbb{1}(n_a>0)}{\sum_{i=1}^n \mathbb{1}(A_i=a)}\right] \\ &\leq \frac{2q_{\max}\|K\|_2^2}{(n+1)h^d\pi_a}, \end{aligned} \tag{B.6}$$

where the last inequality follows by Lemma 4.1 from [**?**]. Thus we have obtained the upper bound for the first term of (B.1).

Next, the second term of (B.1) can be bounded simply as

$$\mathbb{E}\left[q_h^a(y)^2\mathbb{1}(n_a=0)\right] = q_h^a(y)^2\mathbb{P}(n_a=0) \leq \frac{q_{\max}\|K\|_2^2}{h^d}(1-\pi_a)^n. \tag{B.7}$$

Finally, applying (B.6) and (B.7) to (B.1) gives $L_2(\mathbb{P})$ bound for $\hat{q}_h^a(y) - q_h^a(y)$ as

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{q}_h^a(y) - q_h^a(y)\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n (T_{h,y}(Y_i) - q_h^a(y))\mathbb{1}(A_i = a)}{\sum_{i=1}^n \mathbb{1}(A_i = a)}\right)^2 \mathbb{1}(n_a > 0)\right] + \mathbb{E}\left[q_h^a(y)^2 \mathbb{1}(n_a = 0)\right] \\
&\leq \frac{q_{\max}\|K\|_2^2}{h^d}\left(\frac{2}{(n+1)\pi_a} + (1-\pi_a)^n\right) \\
&\leq \frac{q_{\max}\|K\|_2^2}{h^d}\left(\frac{2}{n\pi_a} + \exp(-n\pi_a)\right) \\
&\leq \frac{3q_{\max}\|K\|_2^2}{n\pi_a h^d}.
\end{aligned}$$

Applying Jensen's inequality gives the bound for $\hat{q}_h^a(y) - q_h^a(y)$ by

$$\begin{aligned}
\mathbb{E}\left[|\hat{q}_h^a(y) - q_h^a(y)|\right] &\leq \sqrt{\mathbb{E}\left[\left(\hat{q}_h^a(y) - q_h^a(y)\right)^2\right]} \\
&\leq \frac{\sqrt{3q_{\max}}\|K\|_2}{\sqrt{n\pi_a h^d}} \\
&\leq \frac{C_{K,q_{\max}}}{\sqrt{n\pi_a h^d}},
\end{aligned}$$

where $C_{K,q_{\max}} = \sqrt{3q_{\max}}\|K\|_2$ is a constant depending only on $\|K\|_2$ and $q_{\max}$.

□

Lemma B.2.2.

$$\mathbb{E}\left[D(\hat{Q}_h^a, Q_h^a)\right] \leq \frac{C_{K,q_{\max},\mathbb{D}}}{\sqrt{n\pi_a h^d}},$$

where $C_{K,q_{\max},\mathbb{D}}$ is a constant depending only on $\|K\|_2$, $q_{\max}$, $\lambda_d(\mathbb{D})$.

Proof. Applying Lemma B.2.1 and Fubini Theorem gives

$$\begin{aligned}
\mathbb{E}\left[D(\hat{Q}_h^a, Q_h^a)\right] &= \mathbb{E}\left[\int_{\mathbb{D}} |\hat{q}_h^a(u) - q_h^a(u)| du\right] \\
&= \int_{\mathbb{D}} \mathbb{E}\left[|\hat{q}_h^a(u) - q_h^a(u)|\right] du \\
&\leq \lambda_d(\mathbb{D}) \sup_{u\in\mathbb{D}} \mathbb{E}\left[|\hat{q}_h^a(u) - q_h^a(u)|\right] \\
&\leq \frac{C_{K,q_{\max}}\lambda_d(\mathbb{D})}{\sqrt{n\pi_a h^d}} = \frac{C_{K,q_{\max},\mathbb{D}}}{\sqrt{n\pi_a h^d}},
\end{aligned}$$

where $C_{K,q_{\max},\mathbb{D}} = C_{K,q_{\max}}\lambda_d(\mathbb{D})$.

□

B.2.3 Proof of Theorem 3.3.1

Claim B.2.1. *For distributions Q_1, Q_2, Q_3, Q_4,*

$$|D(Q_1,Q_2)-D(Q_3,Q_4)| \leq D(Q_1,Q_3)+D(Q_2,Q_4).$$

Proof. Since D is distance measure, by triangle inequality it follows

$$\begin{aligned} D(Q_1,Q_2) &\leq D(Q_1,Q_3)+D(Q_3,Q_4)+D(Q_4,Q_2),\\ D(Q_3,Q_4) &\leq D(Q_3,Q_1)+D(Q_1,Q_2)+D(Q_2,Q_4), \end{aligned}$$

and consequently we obtain

$$|D(Q_1,Q_2)-D(Q_3,Q_4)| \leq D(Q_1,Q_3)+D(Q_2,Q_4).$$

□

Theorem B.2.1. *Under the assumptions (A1) and (A2),*

$$\mathbb{E}\left[|D(\widehat{Q^1_{h_1}},\widehat{Q^0}_h)-D(Q^1_{h_1},Q^0_{h_0})|\right] \leq C_{K,q_{\max},\mathbb{D}}\left(\frac{1}{\sqrt{n\pi_1 h_1^d}}+\frac{1}{\sqrt{n\pi_0 h_0^d}}\right),$$

where $C_{K,q_{\max},\mathbb{D}}$ is a constant depending only on $\|K\|_2$, $q_{\max}$, $\lambda_d(\mathbb{D})$.

Proof. Applying Claim B.2.1 gives

$$|D(\widehat{Q^1_{h_1}},\widehat{Q^0_{h_0}})-D(Q^1_{h_1},Q^0_{h_0})| \leq D(\widehat{Q^1_{h_1}},Q^1_{h_1})+D(\widehat{Q^0_{h_0}},Q^0_{h_0}).$$

Hence under (A1) and (A2), taking expectation and applying Lemma B.2.2 gives

$$\begin{aligned} \mathbb{E}\left[|D(\widehat{Q^1_{h_1}},\widehat{Q^0_{h_0}})-D(Q^1_{h_1},Q^0_{h_0})|\right] &\leq \mathbb{E}\left[D(\widehat{Q^1_{h_1}},Q^1_{h_1})\right]+\mathbb{E}\left[D(\widehat{Q^0_{h_0}},Q^0_{h_0})\right]\\ &\leq C_{K,q_{\max},\mathbb{D}}\left(\frac{1}{\sqrt{n\pi_1 h_1^d}}+\frac{1}{\sqrt{n\pi_0 h_0^d}}\right). \end{aligned}$$

□

B.2.4 Proof of Theorem 3.3.2

Theorem B.2.2. *Under the assumptions (A1) and (A2),*

$$\begin{aligned}
&\mathbb{E}_{\mathscr{P}}\left[\left|\frac{1}{N}\sum_{i=1}^{N} D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\right|\right]\\
&\leq \frac{C_{K,q_{\max},\mathbb{D}}}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{n_i\pi_{1,i}h_1^d}}+\frac{1}{\sqrt{n_i\pi_{0,i}h_0^d}}\right)+\frac{\sigma_{\mathscr{P}}}{\sqrt{N}},
\end{aligned}$$

where $C_{K,q_{\max},\mathbb{D}}$ is from Theorem B.2.1 and $\sigma_{\mathscr{P}}=\sqrt{Var_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]}$ depends only on $\mathscr{P}$.

Proof. First, note that $\frac{1}{N}\sum_{i=1}^{N} D(\widehat{Q^1_i},\widehat{Q^0_i})-\mathbb{E}_{\mathscr{P}}[D(Q^1,Q^0)]$ can be expanded as

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\\
&=\frac{1}{N}\sum_{i=1}^{N}\left(D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)-D(Q^1_{h_1},Q^0_{h_0})\right)+\frac{1}{N}\sum_{i=1}^{N}\left(D(Q^1_{h_1},Q^0_{h_0})-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\right).
\end{aligned}\tag{B.8}$$

For the first term of (B.8), by law of total expectation and Theorem B.2.1, we have

$$\begin{aligned}
&\mathbb{E}_{\mathscr{P}}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)-D(Q^1_{h_1},Q^0_{h_0})\right)\right|\right]\\
&=\mathbb{E}_{\mathscr{P}}\left[\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)-D(Q^1_{h_1},Q^0_{h_0})\right)\right| |\mathbb{P}_1,\dots,\mathbb{P}_n\right]\right]\\
&\leq\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathscr{P}}\left[\mathbb{E}\left[\left|D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)-D(Q^1_{h_1},Q^0_{h_0})\right| |\mathbb{P}_i\right]\right]\\
&\leq\frac{C_{K,q_{\max},\mathbb{D}}}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{n_i\pi_{1,i}h_1^d}}+\frac{1}{\sqrt{n_i\pi_{0,i}h_0^d}}\right),
\end{aligned}\tag{B.9}$$

For the second term of (B.8), Jensen inequality and applying Lemma B.2.4 gives the bound as

$$\begin{aligned}
&\mathbb{E}_{\mathscr{P}}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(D(Q^1_{h_1},Q^0_{h_0})-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\right)\right|\right]\\
&\leq\sqrt{\mathbb{E}_{\mathscr{P}}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\left(D(Q^1_{h_1},Q^0_{h_0})-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\right)\right)^2\right]}\\
&\leq\frac{\sqrt{Var_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]}}{\sqrt{N}}. \qquad \text{(B.10)}
\end{aligned}$$

Hence applying (B.9) and (B.10) to (B.8) gives the bound for $\frac{1}{N}\sum_{i=1}^{N}D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]$ as

$$\begin{aligned}
&\mathbb{E}_{\mathscr{P}}\left[\left|\frac{1}{N}\sum_{i=1}^{N}D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\right|\right]\\
&\leq\mathbb{E}_{\mathscr{P}}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(D((\widehat{Q^1_{h_1}})_i,(\widehat{Q^0_{h_0}})_i)-D(Q^1_{h_1},Q^0_{h_0})\right)\right|\right]\\
&\quad+\mathbb{E}_{\mathscr{P}}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(D(Q^1_{h_1},Q^0_{h_0})-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\right)\right|\right]\\
&\leq\frac{C_{K,q_{\max},\mathbb{D}}}{N}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{n_i\pi_{1,i}h_1^d}}+\frac{1}{\sqrt{n_i\pi_{0,i}h_0^d}}\right)+\frac{\sigma_{\mathscr{P}}}{\sqrt{N}},
\end{aligned}$$

where $\sigma_{\mathscr{P}}=\sqrt{Var_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]}$.

□

B.2.5 Proof of Theorem 3.3.3

Part A: Lemmas on robustness to model misspecification and sample splitting

First we prove following two lemmas that come in handy for rest of the proof.

Lemma B.2.3. *(Robustness to model misspecification) As in assumption (B1), let $\overline{\pi}_a$ and $\overline{\mu}_a$ denote fixed functions to which $\widehat{\pi}_a$ and $\widehat{\mu}_a$ asymptotically converge in the sense that $\|\widehat{\pi}_a-\overline{\pi}_a\|=o_{\mathbb{P}}(1)$ and $\|\widehat{\mu}^a-\overline{\mu}^a\|=o_{\mathbb{P}}(1)$, where $\overline{\pi}_a$ and $\overline{\mu}_a$ are not necessarily true* *functions π_a and μ_a. Also recall that $q_h^a(y) = \mathbb{E}\left\{\mathbb{E}\left[T_{h,y}(Y)\middle|X, A=a\right]\right\}$ as defined in (3.6). Then under the set of causal assumptions for observational study, it follows*

$$\begin{aligned}
q_h^a(y) &= \mathbb{E}\left\{\mu_a(X)\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{1}(A=a)}{\overline{\pi}_a(X)}\Big(T_{h,y}(Y) - \mu_A(X)\Big) + \mu_a(X)\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big(T_{h,y}(Y) - \overline{\mu}_A(X)\Big) + \overline{\mu}_a(X)\right\}
\end{aligned} \tag{B.11}$$

Proof. First equality in (B.11) immediately comes from the definition. Let's start with the second equality in which π_a is not correctly specified. It is not hard to obtain the following:

$$\begin{aligned}
&\mathbb{E}\left\{\frac{\mathbb{1}(A=a)}{\overline{\pi}_a(X)}\Big(T_{h,y}(Y) - \mu_A(X)\Big) + \mu_a(X)\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[\frac{\mathbb{1}(A=a)}{\overline{\pi}_a(X)}\Big(T_{h,y}(Y) - \mu_A(X)\Big) + \mu_a(X)\middle|A, X\right]\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{1}(A=a)}{\overline{\pi}_a(X)}\Big(\mathbb{E}[T_{h,y}(Y)|A, X] - \mu_A(X)\Big) + \mu_a(X)\Big|\right\} \\
&= \mathbb{E}\left\{\frac{\mathbb{1}(A=a)}{\overline{\pi}_a(X)}\Big(\mu_A(X) - \mu_A(X)\Big) + \mu_a(X)\Big|\right\} \\
&= \mathbb{E}\left\{\mu_a(X)\right\}
\end{aligned}$$

, where the first equality comes from the law of total expectation. Next, we show the third equality of (B.11) where μ_a is not correctly specified. In fact, it follows that

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big(T_{h,y}(Y)-\overline{\mu}_A(X)\Big)+\overline{\mu}_a(X)\right\}\\
&=\mathbb{E}\left\{\mathbb{E}\left[\frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big(T_{h,y}(Y)-\overline{\mu}_A(X)\Big)+\overline{\mu}_a(X)\Big|A,X\right]\right\}\\
&=\mathbb{E}\left\{\frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big(\mathbb{E}[T_{h,y}(Y)|A,X]-\overline{\mu}_A(X)\Big)+\overline{\mu}_a(X)\right\}\\
&=\mathbb{E}\left\{\frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big(\mu_A(X)-\overline{\mu}_A(X)\Big)+\overline{\mu}_a(X)\right\}\\
&=\mathbb{E}\left\{\frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big(\mu_a(X)-\overline{\mu}_a(X)\Big)+\overline{\mu}_a(X)\right\}\\
&=\mathbb{E}\left\{\mathbb{E}\left[\frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big(\mu_a(X)-\overline{\mu}_a(X)\Big)+\overline{\mu}_a(X)\Big|X\right]\right\}\\
&=\mathbb{E}\left\{\frac{E[A=a|X]}{\pi_a(X)}\Big(\mu_a(X)-\overline{\mu}_a(X)\Big)+\overline{\mu}_a(X)\right\}\\
&=\mathbb{E}\left\{\mu_a(X)-\overline{\mu}_a(X)+\overline{\mu}_a(X)\right\}\\
&=\mathbb{E}\left\{\mu_a(X)\right\}
\end{aligned}
$$

, where we use the law of total expectation in the first and the fifth equality and use the fact that $E[A=a|X]=\pi_a(X)$ in the sixth equality. □

Note that the assumption used in Lemma B.2.3 is even much weaker than (B1). Indeed, $\overline{\mu}_a$ and $\overline{\pi}_a$ can be anything to which $\widehat{\pi}_a$ and $\widehat{\mu}_a$ converge regardless of the convergence rates.

We then show the following Lemma, which is a slight modification from [**?** , Lemma 2]. As mentioned in the main text, for a function f, we use the notation $\|f\|_q=(\int|f(z)|^q d\mathbb{P}(z))^{\frac{1}{q}}$ be the $L_q(\mathbb{P})$-norm of f.

Lemma B.2.4. *Let $\mathbb{P}_n$ denote the empirical measure over $(Z_1,\ldots,Z_n)$, which is i.i.d. from $\mathbb{P}$. Let $\hat{f}$ be a real-valued function constructed in a separate independent sample. Let $\mathbb{P}(\hat{f})=\int\hat{f}(z)d\mathbb{P}(z)$ and let $\mathbb{E}$ be over $(Z_1,\ldots,Z_n)$, then we have*

$$
\sqrt{\mathbb{E}\left[\left((\mathbb{P}_n-\mathbb{P})\hat{f}\right)^2\right]}\leq\sqrt{\frac{Var\left[\hat{f}\right]}{n}}\leq\frac{\|\hat{f}\|_2}{\sqrt{n}}.
$$

Proof. $\left((\mathbb{P}_n-\mathbb{P})\hat{f}\right)^2$ can be expanded as

$$\begin{aligned}\left((\mathbb{P}_n-\mathbb{P})\hat{f}\right)^2 &= \left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(Z_i)-\mathbb{E}\left[\hat{f}(Z_i)\right]\right)\right)^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\left(\hat{f}(Z_i)-\mathbb{E}\left[\hat{f}(Z_i)\right]\right)^2+\sum_{i\neq j}\left(\hat{f}(Z_i)-\mathbb{E}\left[\hat{f}(Z_i)\right]\right)\left(\hat{f}(Z_j)-\mathbb{E}\left[\hat{f}(Z_j)\right]\right).\end{aligned}$$

Then from independence of Z_i and Z_j,

$$\begin{aligned}&\mathbb{E}\left[\left((\mathbb{P}_n-\mathbb{P})\hat{f}\right)^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left(\hat{f}(Z_i)-\mathbb{E}\left[\hat{f}(Z_i)\right]\right)^2\right]+\sum_{i\neq j}\mathbb{E}\left[\hat{f}(Z_i)-\mathbb{E}\left[\hat{f}(Z_i)\right]\right]\mathbb{E}\left[\hat{f}(Z_j)-\mathbb{E}\left[\hat{f}(Z_j)\right]\right] \\ &= \frac{1}{n}Var\left[\hat{f}\right] \\ &\leq \frac{1}{n}\mathbb{E}\left[\hat{f}^2\right]=\frac{1}{n}\left\|\hat{f}\right\|_2^2.\end{aligned}$$

□

Just for further guide to notations, notice that $\mathbb{P}(f)$ is random only if $\hat{f}$ depends on samples, in which case $\mathbb{P}(\hat{f})\neq\mathbb{E}(\hat{f})$. Otherwise $\mathbb{P}$ and $\mathbb{E}$ can be use exchangeably.

Part B: Bounding $\widehat{\psi}_h^a-q_h^a$

For all $y\in\mathbb{R}^d$, let $T_{h,y}:\mathbb{R}^d\to\mathbb{R}$ be $T_{h,y}(y')=\frac{1}{h^d}K\left(\frac{\|y-y'\|_2}{h}\right)$, and let $\hat{f}_{h,y}^a:\mathbb{R}^d\times\{0,1\}\times\mathbb{R}\to\mathbb{R}$, $f_{h,y}^a:\mathbb{R}^d\times\{0,1\}\times\mathbb{R}\to\mathbb{R}$ be

$$\begin{aligned}\hat{f}_{h,y}^a(x',a',y') &= \frac{\mathbb{1}_a(a')}{\hat{\pi}_a(x')}\left(T_{h,y}(y')-\hat{\mu}_{a'}(x')\right)+\hat{\mu}_a(x'), \\ f_{h,y}^a(x',a',y') &= \frac{\mathbb{1}_a(a')}{\bar{\pi}_a(x')}\left(T_{h,y}(y')-\bar{\mu}_{a'}(x')\right)+\bar{\mu}_a(x'),\end{aligned}$$

Hereafter we proceed with shorthand notations $\widehat{\pi}_a,\widehat{\mu}_a,\overline{\pi}_a,\overline{\mu}_a$.

Claim B.2.2. *For all $y\in\mathbb{R}^d$, $\widehat{\psi}_h^a(y)-q_h^a(y)$ can be decomposed as*

$$\widehat{\psi}_h^a(y)-q_h^a(y)=(\mathbb{P}_n-\mathbb{P})\hat{f}_{h,y}^a+\mathbb{P}(\hat{f}_{h,y}^a-f_{h,y}^a).$$

Proof. By Lemma B.2.3, we have

$$\widehat{\psi}_h^a(y)-q_h^a(y)=\mathbb{P}_n\hat{f}_{h,y}^a-\mathbb{P}f_{h,y}^a=(\mathbb{P}_n-\mathbb{P})\hat{f}_{h,y}^a+\mathbb{P}(\hat{f}_{h,y}^a-f_{h,y}^a),$$

as long as at least one of $\overline{\mu}_a$ and $\overline{\pi}_a$ is correctly specified as μ_a and π_a respectively.

□

Lemma B.2.5. *Under the assumptions (A2), (B2), and (B3), and that at least one of $\overline{\mu}_a$ and $\overline{\pi}_a$ is correctly specified as μ_a and π_a, for all $y \in \mathbb{R}^d$,*

$$\mathbb{E}\left[|\widehat{\psi}_h^a(y) - q_h^a(y)|\right] \leq C_{h,K,\hat{\pi}_a,\hat{\mu}_a} \frac{1}{\sqrt{n}} + C_{\hat{\pi}_a} \|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2,$$

where $C_{h,K,\hat{\pi}_a,\hat{\mu}_a}$ is a constant depending only on h, $\|K\|_2$, $\left\|\frac{1}{\hat{\pi}_a}\right\|_\infty$, $\|\hat{\mu}_a\|_2$, and $C_{\hat{\pi}_a}$ is a constant depending only on $\left\|\frac{1}{\hat{\pi}_a}\right\|_\infty$.

Proof. From Claim B.2.2 it follows

$$\widehat{\psi}_h^a(y) - q_h^a(y) = (\mathbb{P}_n - \mathbb{P})\hat{f}_{h,y}^a + \mathbb{P}(\hat{f}_{h,y}^a - f_{h,y}^a). \tag{B.12}$$

For the first term of (B.12), under (A2) and (B2), note that $\left\|\hat{f}_{h,y}^a\right\|_{L_2}$ can be bounded as

$$\begin{aligned}
\left\|\hat{f}_{h,y}^a\right\|_2 &= \left\|\frac{\mathbb{1}_a}{\hat{\pi}_a}\left(T_{h,y} - \hat{\mu}_a\right) + \hat{\mu}_a\right\|_2 \\
&\leq \left\|\frac{\mathbb{1}_a}{\hat{\pi}_a}\right\|_\infty \left(\left\|T_{h,y}\right\|_2 + \|\hat{\mu}_a\|_2\right) + \|\hat{\mu}_a\|_2 \\
&\leq \left\|\frac{1}{\hat{\pi}_a}\right\|_\infty \left(h^{-d}\|K\|_2 + 2\|\hat{\mu}_a\|_2\right).
\end{aligned}$$

Hence under (B3), we apply Lemma B.2.4 and get the bound as

$$\begin{aligned}
\mathbb{E}\left[\left|(\mathbb{P}_n - \mathbb{P})\hat{f}_{h,y}^a\right|\right] &\leq \sqrt{\mathbb{E}\left[\left|(\mathbb{P}_n - \mathbb{P})\hat{f}_{h,y}^a\right|^2\right]} \\
&\leq \frac{\left\|\hat{f}_{h,y}^a\right\|_2}{\sqrt{n}} \\
&\leq \frac{1}{\sqrt{n}}\left\|\frac{1}{\hat{\pi}_a}\right\|_\infty \left(h^{-d}\|K\|_2 + 2\|\hat{\mu}_a\|_2\right).
\end{aligned} \tag{B.13}$$

For the second term of the decomposition (B.12), we have that

$$\begin{aligned}
\mathbb{P}(\hat{f}_h^a - f_h^a) &= \mathbb{P}\left[\frac{\mathbb{1}(A=a)}{\widehat{\pi}_a}(T_h - \widehat{\mu}_A) + \widehat{\mu}_a - \frac{\mathbb{1}(A=a)}{\overline{\pi}_a}(T_h - \overline{\mu}_A) - \overline{\mu}_a\right] \\
&= \mathbb{P}\left[\frac{\mathbb{1}(A=a)T_h}{\widehat{\pi}_a \pi_a}(\overline{\pi}_a - \widehat{\pi}_a) - \frac{\mathbb{1}(A=a)}{\widehat{\pi}_a}(\widehat{\mu}_A - \overline{\mu}_A) - \overline{\mu}_A \mathbb{1}(A=a)\frac{\overline{\pi}_a - \widehat{\pi}_a}{\widehat{\pi}_a \overline{\pi}_a} + \widehat{\mu}_a - \overline{\mu}_a\right] \\
&= \mathbb{P}\left[\frac{\mathbb{1}(A=a)}{\widehat{\pi}_a \overline{\pi}_a}(\overline{\pi}_a - \widehat{\pi}_a)(T_h - \mu_A) + (\widehat{\mu}_a - \overline{\mu}_a)\left(1 - \frac{\mathbb{1}(A=a)}{\widehat{\pi}_a}\right)\right] \\
&= \mathbb{P}\left[\frac{\mathbb{1}(A=a)}{\widehat{\pi}_a \overline{\pi}_a}(\overline{\pi}_a - \widehat{\pi}_a)(\mu_A - \mu_A) + (\widehat{\mu}_a - \overline{\mu}_a)\frac{(\widehat{\pi}_a - \overline{\pi}_a)}{\widehat{\pi}_a}\right] \\
&= \mathbb{P}\left[(\widehat{\mu}_a - \overline{\mu}_a)\frac{(\widehat{\pi}_a - \overline{\pi}_a)}{\widehat{\pi}_a}\right]
\end{aligned}$$

where the second inequality follows by adding and subtracting $\frac{\mathbb{1}(A=a)}{\widehat{\pi}_a}\overline{\mu}_A$ and the fourth by the law of total expectation conditioning on (X,A). By assumption (B2), we have $\left\|\frac{1}{\hat{\pi}_a}\right\|_\infty < \infty$. Hence by conditional Cauchy-Schwarz inequality finally we have

$$\begin{aligned}
\mathbb{P}(\hat{f}_h^a - f_h^a) &\leq \left\|\frac{1}{\hat{\pi}_a}\right\|_\infty \mathbb{P}\left[(\widehat{\mu}_a - \overline{\mu}_a)(\widehat{\pi}_a - \overline{\pi}_a)\right] \\
&\leq \left\|\frac{1}{\hat{\pi}_a}\right\|_\infty \|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2. \qquad \text{(B.14)}
\end{aligned}$$

Hence applying (B.13) and (B.14) to (B.12) leads to

$$\begin{aligned}
\mathbb{E}\left[|\widehat{\psi}_h^a(y) - q_h^a(y)|\right] &\leq \mathbb{E}\left[\left|(\mathbb{P}_n - \mathbb{P})\hat{f}_{h,y}^a\right|\right] + \mathbb{P}(\hat{f}_{h,y}^a - f_{h,y}^a) \\
&\leq \frac{1}{\sqrt{n}}\left\|\frac{1}{\hat{\pi}_a}\right\|_\infty \left(h^{-d}\|K\|_2 + 2\|\hat{\mu}_a\|_2\right) + \left\|\frac{1}{\hat{\pi}_a}\right\|_\infty \|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2 \\
&= C_{h,K,\hat{\pi}_a,\hat{\mu}_a}\frac{1}{\sqrt{n}} + C_{\hat{\pi}_a}\|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2,
\end{aligned}$$

where $C_{h,K,\hat{\pi}_a,\hat{\mu}_a} = \left\|\frac{1}{\hat{\pi}_a}\right\|_\infty \left(h^{-d}\|K\|_2 + 2\|\hat{\mu}_a\|_2\right)$ is a constant depending only on h, $\|K\|_2$, $\left\|\frac{1}{\hat{\pi}_a}\right\|_\infty$, $\|\hat{\mu}_a\|_2$, and $C_{\hat{\pi}_a} = \left\|\frac{1}{\hat{\pi}_a}\right\|_\infty$ is a constant depending only on $\left\|\frac{1}{\hat{\pi}_a}\right\|_\infty$.

□

Part C: Bounding L_1 risk of $D(\widehat{\psi}_h^1, \widehat{\psi}_h^0)$

Claim B.2.3. *Let $\mathscr{Y}_{h,R_K} = \{u \in \mathbb{R}^d$: there exists $y \in \mathscr{Y}$ with $\|\frac{u-y}{h}\| \leq R_K\}$. Then if $u \notin \mathscr{Y}_{h,R_K}$,*

$$\widehat{\psi}_h^a(u) = q_h^a(u) = 0.$$

Proof. Note that for all $u \notin \mathscr{Y}_{h,R_K}$ and $y \in \mathscr{Y}$, $K\left(\frac{u-y}{h}\right) = 0$. And hence $\widehat{\psi}_h^a(u) = q_h^a(u) = 0$ if $u \notin \mathscr{Y}_{h,R_K}$.

□

Lemma B.2.6. *Under the assumptions (A1), (A2), (B2), and (B3), we have*

$$\mathbb{E}\left[\left|D(\widehat{Q_h^a}, Q_h^a)\right|\right] \leq C_{h,K,\hat{\pi}_a,\hat{\mu}_a,\mathscr{Y}_{h,R_K}} \frac{1}{\sqrt{n}} + C_{\hat{\pi}_a,\mathscr{Y}_{h,R_K}} \|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2,$$

where $C_{h,K,\hat{\pi}_a,\hat{\mu}_a,\mathscr{Y}_{h,R_K}}$ is a constant depending only on h, $\|K\|_2$, $\left\|\frac{1}{\hat{\pi}_a}\right\|_\infty$, $\|\hat{\mu}_a\|_2$, $\lambda_d(\mathscr{Y}_{h,R_K})$, and $C_{\hat{\pi}_a,\mathscr{Y}_{h,R_K}}$ is a constant depending only on $\left\|\frac{1}{\hat{\pi}_a}\right\|_\infty$, $\lambda_d(\mathscr{Y}_{h,R_K})$, with $\mathscr{Y}_{h,R_k}$ from Claim B.2.3.

Proof. From our set up we have that

$$D(\widehat{Q}_h^a, Q_h^a) = \int |\hat{q}^a(u) - q^a(u)|\, du = \int |\widehat{\psi}_h^a(u) - q_h^a(u)|\, du.$$

Then from Claim B.2.3,

$$D(\widehat{Q}_h^a, Q_h^a) = \int_{\mathscr{Y}_{h,R_k}} |\widehat{\psi}_h^a(u) - q_h^a(u)|\, du.$$

Then applying Fubini's Theorem and Lemma B.2.5 provides the upper bound for $\mathbb{E}\left[\left|D(\widehat{Q_h^a}, Q_h^a)\right|\right]$ as

$$\begin{aligned}
\mathbb{E}\left[\left|D(\widehat{Q_h^a}, Q_h^a)\right|\right] &= \mathbb{E}\left[\int_{\mathscr{Y}_{h,R_k}} |\widehat{\psi}_h^a(u) - q_h^a(u)|\, du\right] \\
&= \int_{\mathscr{Y}_{h,R_k}} \mathbb{E}\left[|\widehat{\psi}_h^a(u) - q_h^a(u)|\right] du \\
&\leq \lambda_d(\mathscr{Y}_{h,R_k}) \sup_{u \in \mathscr{Y}_{h,R_k}} \left|\mathbb{E}\left[|\widehat{\psi}_h^a(u) - q_h^a(u)|\right]\right| \\
&\leq \lambda_d(\mathscr{Y}_{h,R_k}) C_{h,K,\hat{\pi}_a,\hat{\mu}_a} \frac{1}{\sqrt{n}} + \lambda(\mathscr{Y}_{h,R_k}) C_{\hat{\pi}_a} \|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2 \\
&= C_{h,K,\hat{\pi}_a,\hat{\mu}_a,\mathscr{Y}_{h,R_K}} \frac{1}{\sqrt{n}} + C_{\hat{\pi}_a,\mathscr{Y}_{h,R_K}} \|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2,
\end{aligned}$$

where $C_{h,K,\hat{\pi}_a,\hat{\mu}_a,\mathscr{Y}_{h,R_K}} = \lambda_d(\mathscr{Y}_{h,R_K}) C_{h,K,\hat{\pi}_a,\hat{\mu}_a}$ and $C_{\hat{\pi}_a,\mathscr{Y}_{h,R_K}} = \lambda_d(\mathscr{Y}_{h,R_K}) C_{\hat{\pi}_a}$.

□

Finally the following theorem concludes our proof for Theorem 3.3.3.

Theorem B.2.3. *Under the assumptions (A1), (A2), (B1), (B2), and (B3), we have*

$$\mathbb{E}\left[\left|D(\widehat{Q_h^1},\widehat{Q_h^0})-D\left(Q_h^1,Q_h^0\right)\right|\right]\leq C_{h,K,\hat{\pi}^1,\hat{\pi}^0,\hat{\mu}_1,\hat{\mu}_0,\mathscr{Y}_{h,R_K}}\frac{1}{\sqrt{n}}+C_{\hat{\pi}^1,\hat{\pi}^0,\mathscr{Y}_{h,R_K}}\|\widehat{\mu}_a-\overline{\mu}_a\|_2\|\widehat{\pi}_a-\overline{\pi}_a\|_2,$$

where $C_{h,K,\hat{\pi}^1,\hat{\pi}^0,\hat{\mu}_1,\hat{\mu}_0,\mathscr{Y}_{h,R_K}}$ is a constant depending only on h, $\|K\|_2$, $\left\|\frac{1}{\hat{\pi}^1}\right\|_\infty$, $\left\|\frac{1}{\hat{\pi}^0}\right\|_\infty$, $\|\hat{\mu}_1\|_2$, $\|\hat{\mu}_0\|_2$, $\lambda_d(\mathscr{Y}_{h,R_K})$, and $C_{\hat{\pi}^1,\hat{\pi}^0,\mathscr{Y}_{h,R_K}}$ is a constant depending only on $\left\|\frac{1}{\hat{\pi}^1}\right\|_\infty$, $\left\|\frac{1}{\hat{\pi}^0}\right\|_\infty$, $\lambda_d(\mathscr{Y}_{h,R_K})$.

In particular,

$$\mathbb{E}\left[\left|D(\widehat{Q_h^1},\widehat{Q_h^0})-D\left(Q_h^1,Q_h^0\right)\right|\right]=O\left(\frac{1}{\sqrt{n}}\right)+O_{\mathbb{P}}\left(s(n)r(n)\right).$$

Proof. $\mathbb{E}\left[\left|D(\widehat{Q_h^1},\widehat{Q_h^0})-D\left(Q_h^1,Q_h^0\right)\right|\right]$ can be bounded as

$$\mathbb{E}\left[\left|D(\widehat{Q_h^1},\widehat{Q_h^0})-D\left(Q_h^1,Q_h^0\right)\right|\right]\leq\mathbb{E}\left[\left|D(\widehat{Q_h^1},Q_h^1)\right|\right]+\mathbb{E}\left[\left|D(\widehat{Q_h^0},Q_h^0)\right|\right].$$

Then under (A1), (A2), (B1), (B2), (B3), applying Lemma B.2.6 gives the bound as

$$\begin{aligned}\mathbb{E}\left[\left|D(\widehat{Q_h^1},\widehat{Q_h^0})-D\left(Q_h^1,Q_h^0\right)\right|\right]&\leq C_{h,K,\hat{\pi}^1,\hat{\mu}_1,\mathscr{Y}_{h,R_K}}\frac{1}{\sqrt{n}}+C_{\hat{\pi}^1,\mathscr{Y}_{h,R_K}}\|\widehat{\mu}_a-\overline{\mu}_a\|_2\|\widehat{\pi}_a-\overline{\pi}_a\|_2\\&\quad+C_{h,K,\hat{\pi}^0,\hat{\mu}_0,\mathscr{Y}_{h,R_K}}\frac{1}{\sqrt{n}}+C_{\hat{\pi}^0,\mathscr{Y}_{h,R_K}}\|\widehat{\mu}_a-\overline{\mu}_a\|_2\|\widehat{\pi}_a-\overline{\pi}_a\|_2\\&\leq C_{h,K,\hat{\pi}^1,\hat{\pi}^0,\hat{\mu}_1,\hat{\mu}_0,\mathscr{Y}_{h,R_K}}\frac{1}{\sqrt{n}}+C_{\hat{\pi}^1,\hat{\pi}^0,\mathscr{Y}_{h,R_K}}\|\widehat{\mu}_a-\overline{\mu}_a\|_2\|\widehat{\pi}_a-\overline{\pi}_a\|_2,\end{aligned}$$

where $C_{h,K,\hat{\pi}^1,\hat{\pi}^0,\hat{\mu}_1,\hat{\mu}_0,\mathscr{Y}_{h,R_K}}=C_{h,K,\hat{\pi}^1,\hat{\mu}_1,\mathscr{Y}_{h,R_K}}+C_{h,K,\hat{\pi}^0,\hat{\mu}_0,\mathscr{Y}_{h,R_K}}$ and $C_{\hat{\pi}^1,\hat{\pi}^0,\mathbb{D}}=C_{\hat{\pi}^1,\mathscr{Y}_{h,R_K}}+C_{\hat{\pi}^0,\mathscr{Y}_{h,R_K}}$. □

B.2.6 Proof of Theorem 3.4.1

For $a\in\{0,1\}$, let $\widehat{\pi}_a=\frac{1}{n}\sum_{i=1}^n\mathbb{1}_a(A_i)=\frac{n_a}{n}$. For all $y\in\mathbb{R}^d$, let $f_{h,y}^a:\{0,1\}\times\mathbb{R}\to\mathbb{R}$ be

$$f_{h,y}^a(a',y')=\frac{1}{h^d}K\left(\frac{\|y-y'\|_2}{h}\right)\mathbb{1}_a(a'),$$

and let $\mathscr{F}^a:=\{f_{h,y}^a:y\in\mathscr{Y}_{h,R_K},a\in\mathscr{A}\}$ where $\mathscr{Y}_{h,R_K}=\{u\in\mathbb{R}^d:$ there exists $y\in\mathscr{Y}$ with $\left\|\frac{u-y}{h}\right\|\leq R_K\}$ which can be found in Claim B.2.3. Now we need the following Lemma B.2.7.

Lemma B.2.7. *Under the assumptions (A1), (A2'),*

$$\sqrt{n}(\mathbb{P}_n-\mathbb{P})\to\mathbb{G}\text{ weakly in }\ell_\infty(\mathscr{F}^a).$$

Proof. First, we note that by Assumption (A2') for all $y_1, y_2 \in \mathscr{Y}_{h,R_K}$,

$$\begin{aligned}\left|f^a_{h,y_1}(a',y') - f^a_{h,y_2}(a',y')\right| &\leq \frac{1}{h^d}\left|K\left(\frac{\|y_1 - y'\|_2}{h}\right) - K\left(\frac{\|y_2 - y'\|_2}{h}\right)\right| \\ &\leq \frac{L_K}{h^{d+1}}\|y_1 - y_2\|_2,\end{aligned}$$

and hence for any probability measure P on $\{0,1\} \times \mathbb{R}^d$,

$$\|f^a_{h,y_1} - f^a_{h,y_2}\|_{L_2(P)} \leq \|f^a_{h,y_1} - f^a_{h,y_2}\|_\infty \leq \frac{L_K}{h^{d+1}}\|y_1 - y_2\|_2.$$

Therefore $f^a_{h,y}$ is Lipschitz in parameter, and by Example 19.7 of van der Vaart [139] we have the bracketing numbers satisfy

$$\mathscr{N}_{[]}(\mathscr{F}^a, L_2(P), \varepsilon\frac{L_K}{h^{d+1}}) \leq C_{\mathscr{Y}_{h,R_K}}\left(\frac{\mathrm{diam}\Theta}{\varepsilon}\right)^d$$

for some constant $C_{\mathscr{Y}_{h,R_K}}$, where $\Theta = \mathscr{Y}_{h,R_K} \cup \mathscr{A}$. Since $\mathscr{Y}_{h,R_K}$ is compact subset of $\mathbb{R}^d$, diamΘ is bounded by some constant $C_\Theta < \infty$. Then we have the bracketing integral satisfying

$$\begin{aligned}J_{[]}(\mathscr{F}^a, L_2(P), 1) &= \int_0^1 \sqrt{\log \mathscr{N}_{[]}(\mathscr{F}, L_2(P), \varepsilon)}d\varepsilon \\ &\leq \int_0^1 \sqrt{\log\left(\frac{C_\Theta L_k/h^{d+1}}{\varepsilon}\right)^d}d\varepsilon \\ &\leq \int_0^1 \sqrt{d\log\left(\frac{1}{\varepsilon}\right) + d\log L_K d\varepsilon} < \infty.\end{aligned}$$

Hence, by Theorem 19.5 in van der Vaart [139] $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \to \mathbb{G}$ weakly in $\ell_\infty(\mathscr{F}^a)$. □

Proof of Theorem 3.4.1. Note that from Claim B.2.3, $\hat{q}^a_h(y) = q^a_h(y) = 0$ if $y \notin \mathscr{Y}_{h,R_K}$. Also we note that from the proof of Proposition B.2.1, $q^a_h = \frac{\mathbb{E}[T_{h,y}(Y)\mathbb{1}(A=a)]}{\mathbb{P}(A=a)} = \frac{\mathbb{P}f^a_{h,y}}{\mathbb{P}\mathbb{1}_a}$. Hence

$\sqrt{n}D(\hat{Q}_h^a, Q_h^a)$ can be expanded as

$$\begin{aligned}
\sqrt{n}D(\hat{Q}_h^a, Q_h^a) &= \int \sqrt{n}\,|\hat{q}_h^a(y) - q_h^a(y)|\,dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n}\,|\hat{q}_h^a(y) - q_h^a(y)|\,dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n}\left| \frac{\mathbb{P}_n f_{h,y}^a \mathbb{1}(\mathbb{P}_n \mathbb{1}_a > 0)}{\mathbb{P}_n \mathbb{1}_a} - \frac{\mathbb{P} f_{h,y}^a}{\mathbb{P}\mathbb{1}_a} \right| dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n}\left| \frac{1(\mathbb{P}_n \mathbb{1}_a > 0)}{\mathbb{P}_n \mathbb{1}_a} \left((\mathbb{P}_n - \mathbb{P}) f_{h,y}^a - \frac{(\mathbb{P}_n - \mathbb{P})\mathbb{1}_a \mathbb{P} f_{h,y}^a}{\mathbb{P}\mathbb{1}_a} \right) - 1(\mathbb{P}_n \mathbb{1}_a = 0) \frac{\mathbb{P} f_{h,y}^a}{\mathbb{P}\mathbb{1}_a} \right| dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \left| \frac{1(\widehat{\pi}_a > 0)}{\widehat{\pi}_a} \left(\sqrt{n}(\mathbb{P}_n - \mathbb{P}) f_{h,y}^a - q_h^a(y)\sqrt{n}(\mathbb{P}_n - \mathbb{P})\mathbb{1}_a \right) - q_h^a(y)\sqrt{n}1(\widehat{\pi}_a = 0) \right| dy.
\end{aligned}$$

Hence by letting $\Phi : \ell_\infty(\mathscr{F}_a) \times [0,1] \times \ell_\infty(\mathbb{R}) \to \mathbb{R}$ as $\Phi(\mu, \theta, q) = \frac{1(\theta > 0)}{\theta} \int_{\mathscr{Y}_{h,R_K}} |\mu f_{h,y}^a - q(y)\mu \mathbb{1}_a| dy$, then Φ is continuous on $\ell_\infty(\mathscr{F}_a) \times (0,1] \times \ell_\infty(\mathbb{R})$, and

$$\left| \sqrt{n}D(\hat{Q}_h^a, Q_h^a) - \Phi(\sqrt{n}(\mathbb{P}_n - \mathbb{P}), \widehat{\pi}_a, q_h^a) \right| \leq \sqrt{n}1(\widehat{\pi}_a = 0) \int_{\mathscr{Y}_{h,R_K}} q_h^a(y) dy.$$

Now, note that from strong law of large numbers, $\widehat{\pi}_a \to \pi_a > 0$ a.s.. Hence by Lemma B.2.7 and continuous mapping theorem (e.g., Kosorok [78, Theorem 7.7]) applied to $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \to \mathbb{G}$ and $\hat{\pi}_a \to \pi_a$, we have

$$\Phi(\sqrt{n}(\mathbb{P}_n - \mathbb{P}), \widehat{\pi}_a, q_h^a) \to \Phi(\mathbb{G}, \pi_a, q_h^a) = \frac{1}{\pi_a} \int \left| \mathbb{G} f_{h,y}^a - q_h^a(y) \mathbb{G} \mathbb{1}_a \right| dy \text{ in distribution.}$$

Also, note that

$$\mathbb{E}\left[\sqrt{n}1(\widehat{\pi}_a = 0) \right] = \sqrt{n}(1 - \pi_a)^n \to 0.$$

Then by Markov inequality, we have

$$\left| \sqrt{n}D(\hat{Q}_h^a, Q_h^a) - \Phi(\sqrt{n}(\mathbb{P}_n - \mathbb{P}), \widehat{\pi}_a, q_h^a) \right| \to 0 \text{in probability,}$$

and consequently

$$\sqrt{n}D(\hat{Q}_h^a, Q_h^a) \to \frac{1}{\pi_a} \int \left| \mathbb{G} f_{h,y}^a - q_h^a(y) \mathbb{G} \mathbb{1}_a \right| dy \text{ in distribution.}$$

□

B.2.7 Proof of Theorem 3.4.2

For all $y \in \mathbb{R}^d$, let $T_{h,y} : \mathbb{R}^d \to \mathbb{R}$ be $T_{h,y}(y') = \frac{1}{h^d} K\left(\frac{\|y-y'\|_2}{h}\right)$, and let $\hat{f}^a_{h,y} : \mathbb{R}^k \times \{0,1\} \times \mathbb{R}^d \to \mathbb{R}$, $f^a_{h,y} : \mathbb{R}^k \times \{0,1\} \times \mathbb{R}^d \to \mathbb{R}$ be

$$\begin{aligned}\hat{f}^a_{h,y}(x',a',y') &= \frac{\mathbb{1}_a(a')}{\hat{\pi}_a(x')}\left(T_{h,y}(y') - \hat{\mu}_{a'}(x')\right) + \hat{\mu}_a(x'),\\ f^a_{h,y}(x',a',y') &= \frac{\mathbb{1}_a(a')}{\bar{\pi}_a(x')}\left(T_{h,y}(y') - \bar{\mu}_{a'}(x')\right) + \bar{\mu}_a(x').\end{aligned}$$

Consider $\mathscr{F}^a = \{f^a_{h,y} : y \in \mathscr{Y}_{h,R_K}, a \in \mathscr{A}\}$ where $\mathscr{Y}_{h,R_K}$ is defined in Claim B.2.3.

Lemma B.2.8. *Under the assumptions (A1), (A2'), (B1), (B2'), (B3),*

$$\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \to \mathbb{G} \textit{ weakly in } \ell_\infty(\mathscr{F}^a).$$

Proof. By assumption (A2') and (B2') for all $y_1, y_2 \in \mathscr{Y}_{h,R_K}$,

$$\begin{aligned}\left|f^a_{h,y_1}(x',a',y') - f^a_{h,y_2}(x',a',y')\right| &\le \left\|\frac{1}{\bar{\pi}_a}\right\|_\infty \frac{1}{h^d}\left|K\left(\frac{\|y_1-y'\|_2}{h}\right) - K\left(\frac{\|y_2-y'\|_2}{h}\right)\right|\\ &\le \left\|\frac{1}{\bar{\pi}_a}\right\|_\infty \frac{L_K}{h^{d+1}}\|y_1-y_2\|_2\\ &\le B_\pi \frac{L_K}{h^{d+1}}\|y_1-y_2\|_2,\end{aligned}$$

for some constant $0 < B_\pi < \infty$, and hence for any probability measure P on $\{0,1\} \times \mathbb{R}^d$ we have

$$\|f^a_{h,y_1} - f^a_{h,y_2}\|_{L_2(P)} \le \|f^a_{h,y_1} - f^a_{h,y_2}\|_\infty \le B_\pi \frac{L_K}{h^{d+1}}\|y_1-y_2\|_2.$$

Therefore $f^a_{h,y}$ is Lipschitz in parameter, and by Example 19.7 of van der Vaart [139] we have the bracketing numbers satisfy

$$\mathscr{N}_{[]}(\mathscr{F}^a, L_2(P), \varepsilon\frac{B_\pi L_K}{h^{d+1}}) \le C_{\mathscr{Y}_{h,R_K}}\left(\frac{\mathrm{diam}\Theta}{\varepsilon}\right)^d$$

for some constant $C_{\mathscr{Y}_{h,R_K}}$, where $\Theta = \mathscr{Y}_{h,R_K} \cup \mathscr{A}$. Since $\mathscr{Y}_{h,R_K}$ is compact subset of $\mathbb{R}^d$, diamΘ is bounded by some constant $C_\Theta < \infty$. By the similar argument as in Lemma B.2.7, we have the bracketing integral satisfying

$$\begin{aligned}
J_{[]}(\mathscr{F}^a, L_2(P), 1) &= \int_0^1 \sqrt{\log \mathscr{N}_{[]}(\mathscr{F}^a, L_2(P), \varepsilon)} d\varepsilon \\
&\leq \int_0^1 \sqrt{\log \left(\frac{C_\Theta B_\pi L_k / h^{d+1}}{\varepsilon} \right)^d} d\varepsilon \\
&\leq \int_0^1 \sqrt{d \log \left(\frac{1}{\varepsilon} \right) + d \log (B_\pi L_K)} d\varepsilon < \infty.
\end{aligned}$$

Hence, by Theorem 19.5 in van der Vaart [139] $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \to \mathbb{G}$ weakly in $\ell_\infty(\mathscr{F}^a)$. Notice that since

$$\|\hat{f}^a_{h,y_1} - \hat{f}^a_{h,y_2}\|_{L_2(P)} \leq \|\hat{f}^a_{h,y_1} - \hat{f}^a_{h,y_2}\|_\infty \leq B_\pi \frac{L_K}{h^{d+1}} \|y_1 - y_2\|_2.$$

as well, the same conclusion also holds for $\hat{\mathscr{F}}^a = \{\hat{f}^a_{h,y} : y \in \mathscr{Y}_{h,R_K}, a \in \mathscr{A}\}$.

□

Proof of Theorem 3.4.2. Note that from Claim B.2.3, $\hat{\psi}^a_h(y) = q^a_h(y) = 0$ if $y \notin \mathscr{Y}_{h,R_K}$. Also under (B1), by Lemma B.2.3 we have $q^a_h(y) = \mathbb{P} f^a_{h,y}$. Hence $\sqrt{n} D(\hat{Q}^a_h, Q^a_h)$ can be expanded as

$$\begin{aligned}
\sqrt{n} D(\hat{Q}^a_h, Q^a_h) &= \int \sqrt{n} |\hat{\psi}^a_h(y) - q^a_h(y)| dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n} |\hat{\psi}^a_h(y) - q^a_h(y)| dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n} \left| \mathbb{P}_n \hat{f}^a_{h,y} - \mathbb{P} f^a_{h,y} \right| dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \left| \sqrt{n}(\mathbb{P}_n - \mathbb{P}) f^a_{h,y} + \sqrt{n} \mathbb{P}_n (\hat{f}^a_{h,y} - f^a_{h,y}) \right| dy. \\
&= \Phi(\sqrt{n}(\mathbb{P}_n - \mathbb{P})) + r_n,
\end{aligned}$$

where $\Phi : \ell_\infty(\mathscr{F}^a) \to \mathbb{R}$ is defined by $\Phi(\mu) = \int_{\mathscr{Y}_{h,R_K}} |\mu f^a_{h,y}| dy$. Then Φ is continuous on $\ell_\infty(\mathscr{F}^a)$. Hence by Lemma B.2.8 $\sqrt{n}(\mathbb{P}_n - \mathbb{P}) \to \mathbb{G}$ weakly in $\ell_\infty(\mathscr{F}^a)$ and the continuous mapping theorem [e.g., 78, Theorem 7.7] implies

$$\Phi(\sqrt{n}(\mathbb{P}_n - \mathbb{P})) \to \Phi(\mathbb{G}) = \int_{\mathscr{Y}_{h,R_K}} |\mathbb{G} f^a_{h,y}| dy \text{weakly in } \mathbb{R}.$$

For r_n, it follows that

$$\begin{aligned} r_n &= \sqrt{n} D(\hat{Q}^a_{h_a}, Q^a_{h_a}) - \Phi(\sqrt{n}(\mathbb{P}_n - \mathbb{P})) \\ &\leq \int_{\mathscr{Y}_{h,R_K}} |\sqrt{n}\mathbb{P}_n(\hat{f}^a_{h,y} - f^a_{h,y})| dy \\ &\leq \int_{\mathscr{Y}_{h,R_K}} |\sqrt{n}\mathbb{P}(\hat{f}^a_{h,y} - f^a_{h,y})| dy + \int_{\mathscr{Y}_{h,R_K}} |\sqrt{n}(\mathbb{P}_n - \mathbb{P})(\hat{f}^a_{h,y} - f^a_{h,y})| dy. \end{aligned}$$

Under the condition (B1) and (B2'), by the previous result of (B.14) we have

$$|\sqrt{n}\mathbb{P}(\hat{f}^a_y - f^a_y)| \leq \sqrt{n} \left\| \frac{1}{\hat{\pi}_a} \right\|_\infty \|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2 = o_{\mathbb{P}}(1),$$

and also by (B.14) together with Lemma B.2.4,

$$\int_{\mathscr{Y}_{h,R_K}} |\sqrt{n}(\mathbb{P}_n - \mathbb{P})(\hat{f}^a_{h,y} - f^a_{h,y})| dy = o_{\mathbb{P}}(1).$$

Consequently we have

$$r_n = o_{\mathbb{P}}(1),$$

and hence by Slutsky Theorem [78, Theorem 7.15], finally we have

$$\sqrt{n} D(\hat{Q}^a_{h_a}, Q^a_{h_a}) \to \int |\mathbb{G} f^a_{h,y}| dy \text{ weakly in } \mathbb{R}.$$

□

B.2.8 Bootstrap validity of Theorem 3.4.3 for Single-source randomized study

For $\theta = D(Q^1_{h_1}, Q^0_{h_0})$ and $\hat{\theta} = D(\widehat{Q^1_{h_1}}, \widehat{Q^0_{h_0}})$, by triangle inequality we have $|\hat{\theta} - \theta| \leq D(\widehat{Q^1_{h_1}}, Q^1_{h_1}) + D(\widehat{Q^0_{h_0}}, Q^0_{h_0})$, hence one of the sufficient condition for the confidence interval $\hat{C}_\alpha$ to be valid is

$$\liminf_{n\to\infty} \mathbb{P}\big(D(\widehat{Q^1_{h_1}}, Q^1_{h_1}) + D(\widehat{Q^0_{h_0}}, Q^0_{h_0}) \leq c_n\big) \geq 1 - \alpha.$$

And this is implied from

$$\liminf_{n\to\infty}\mathbb{P}\left(\sqrt{n}D(\widehat{Q^1_{h_1}},Q^1_{h_1})\leq\hat{z}^1_{\alpha/2}\right)\geq 1-\frac{\alpha}{2},$$
$$\liminf_{n\to\infty}\mathbb{P}\left(\sqrt{n}D(\widehat{Q^0_{h_0}},Q^0_{h_0})\leq\hat{z}^0_{\alpha/2}\right)\geq 1-\frac{\alpha}{2}.$$

Hence it suffice to show that $\sqrt{n}D(\widehat{Q^a_h},Q^a_h)$ and $\sqrt{n}D(\widehat{Q^a_h}^*,\widehat{Q^a_h})$ converges to the same distribution.

As in Section B.2.6, for $a\in\{0,1\}$, we let $\widehat{\pi}_a=\frac{1}{n}\sum_{i=1}^n\mathbb{1}_a(A_i)=\frac{n_a}{n}$. For all $y\in\mathbb{R}^d$, let $f^a_{h,y}:\{0,1\}\times\mathbb{R}^d\to\mathbb{R}$ be

$$f^a_{h,y}(a',y')=\frac{1}{h^d}K\left(\frac{\|y-y'\|_2}{h}\right)\mathbb{1}_a(a'),$$

and let $\mathscr{F}^a:=\{f^a_{h,y}:y\in\mathscr{Y}_{h,R_K},a\in\mathscr{A}\}$.

Theorem B.2.4. *Under the assumptions (A1), (A2'),*

$$\sqrt{n}D(\widehat{Q^a_h},Q^a_h)\to\frac{1}{\pi_a}\int\left|\mathbb{G}(f^a_{h,y})-q^a_h(y)\mathbb{G}(\mathbb{1}_a)\right|dy \textit{ weakly in } \mathbb{R}, \tag{B.15}$$
$$\sqrt{n}D((\widehat{Q^a_h})^*,\widehat{Q^a_h})\to\frac{1}{\pi_a}\int\left|\mathbb{G}(f^a_{h,y})-q^a_h(y)\mathbb{G}(\mathbb{1}_a)\right|dy \textit{ weakly in } \mathbb{R}, \tag{B.16}$$

where $\mathbb{G}$ is a centered Gaussian process with $Cov[\mathbb{G}(f),\mathbb{G}(g)]=\int fgd\mathbb{P}-\int fd\mathbb{P}\int gd\mathbb{P}$.

Proof. We already have (B.15) from the result of Theorem 3.4.1. Hence we are left to show (B.16), which can be done by Theorem (3.2.1) and repetition of the proof of Theorem 3.4.1.

Combining Lemma B.2.7 and Theorem (3.2.1) implies that

$$\sqrt{n}(\mathbb{P}^*_n-\mathbb{P}_n)\to\mathbb{G}\text{ weakly in }\ell_\infty(\mathscr{F}^a).$$

Then similarly as in the proof of Theorem 3.4.1 in Section B.2.6, $\sqrt{n}D((\widehat{Q_h^a})^*, \widehat{Q_h^a})$ can be expanded as

$$\begin{aligned}
\sqrt{n}D((\widehat{Q_h^a})^*, \widehat{Q_h^a}) &= \int \sqrt{n}\,|(\hat{q}_h^a)^*(y) - \hat{q}_h^a(y)|\,dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n} \left| \frac{\mathbb{P}_n^* f_{h,y}^a \mathbb{1}(\mathbb{P}_n^* \mathbb{1}_a > 0)}{\mathbb{P}_n^* \mathbb{1}_a} - \frac{\mathbb{P}_n f_{h,y}^a \mathbb{1}(\mathbb{P}_n \mathbb{1}_a > 0)}{\mathbb{P}_n \mathbb{1}_a} \right| dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n} \Bigg| \frac{\mathbb{1}(\mathbb{P}_n^* \mathbb{1}_a > 0)\mathbb{1}(\mathbb{P}_n \mathbb{1}_a > 0)}{\mathbb{P}_n^* \mathbb{1}_a} \Bigg\{ (\mathbb{P}_n^* - \mathbb{P}_n) f_{h,y}^a - \frac{(\mathbb{P}_n^* - \mathbb{P}_n)\mathbb{1}_a \mathbb{P}_n f_{h,y}^a \mathbb{1}(\mathbb{P}_n \mathbb{1}_a > 0)}{\mathbb{P}_n \mathbb{1}_a} \\
&\qquad + \frac{\mathbb{P}_n^* f_{h,y}^a \mathbb{1}(\mathbb{P}_n^* \mathbb{1}_a > 0)\mathbb{1}(\mathbb{P}_n \mathbb{1}_a = 0)}{\mathbb{P}_n^* \mathbb{1}_a} - \frac{\mathbb{P}_n f_{h,y}^a \mathbb{1}(\mathbb{P}_n \mathbb{1}_a > 0)\mathbb{1}(\mathbb{P}_n^* \mathbb{1}_a = 0)}{\mathbb{P}_n \mathbb{1}_a} \Bigg| dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \Bigg| \frac{\mathbb{1}((\widehat{\pi}_a)^* > 0)\mathbb{1}(\widehat{\pi}_a > 0)}{(\widehat{\pi}_a)^*} \left\{ \sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n) f_{h,y}^a - \hat{q}_h^a(y)\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)\mathbb{1}_a \right\} \\
&\qquad + \sqrt{n}\frac{\mathbb{P}_n^* f_{h,y}^a \mathbb{1}((\widehat{\pi}_a)^* > 0)\mathbb{1}(\widehat{\pi}_a = 0)}{(\widehat{\pi}_a)^*} - \sqrt{n}\frac{\mathbb{P}_n f_{h,y}^a \mathbb{1}(\widehat{\pi}_a > 0)\mathbb{1}((\widehat{\pi}_a)^* = 0)}{\widehat{\pi}_a} \Bigg| dy.
\end{aligned} \tag{B.17}$$

Now define a function $\Phi : \ell_\infty(\mathscr{F}_a) \times [0,1] \times [0,1] \times \ell_\infty(\mathbb{R}) \to \mathbb{R}$ by $\Phi(\mu, \theta, \theta^*, q) = \frac{\mathbb{1}(\theta>0)\mathbb{1}(\theta^*>0)}{\theta^*} \int |\mu f_y^a - q(y)\mu \mathbb{1}_a|dy$. Then Φ is continuous on $\ell_\infty(\mathscr{F}_a) \times [0,1] \times (0,1] \times \ell_\infty(\mathbb{R})$. Note that by the strong law of large numbers, $\widehat{\pi}_a \to \pi_a > 0$ a.s., $(\widehat{\pi}_a)^* \to \pi_a > 0$ a.s., and $\hat{q}_h^a \to q_h^a$ a.s.. Hence by the continuous mapping theorem [e.g., 78, Theorem 7.7]) together with $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n) \to \mathbb{G}$ as shown previously, we have

$$\Phi(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n),\, \widehat{\pi}_a,\, (\widehat{\pi}_a)^*,\, \hat{q}_{h_a}^a) \to \Phi(\mathbb{G}, \pi_a, \pi_a, q_h^a) = \frac{1}{\pi_a} \int \left| \mathbb{G} f_{h,y}^a - q_h^a(y)\mathbb{G}\mathbb{1}_a \right| dy \text{ weakly in } \mathbb{R}.$$

Next, by (B.17) it follows that

$$\begin{aligned}
&\left| \sqrt{n}D((\widehat{Q_h^a})^*, \widehat{Q_h^a}) - \Phi(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n),\, \widehat{\pi}_a,\, (\widehat{\pi}_a)^*,\, \hat{q}_h^a) \right| \\
&\le \sqrt{n}\mathbb{1}(\widehat{\pi}_a = 0) \int_{\mathscr{Y}_{h,R_K}} \frac{\mathbb{P}_n^* f_{h,y}^a \mathbb{1}((\widehat{\pi}_a)^* > 0)}{(\widehat{\pi}_a)^*} dy + \sqrt{n}\mathbb{1}((\widehat{\pi}_a)^* = 0) \int_{\mathscr{Y}_{h,R_K}} \frac{\mathbb{P}_n f_{h,y}^a \mathbb{1}(\widehat{\pi}_a > 0)}{\widehat{\pi}_a} dy.
\end{aligned}$$

For the first term in the last display, we have

$$\begin{aligned}\mathbb{E}\left\{\sqrt{n}\mathbb{1}(\widehat{\pi}_a = 0)\int_{\mathscr{Y}_{h,R_K}} \frac{\mathbb{P}_n^* f_{h,y}^a \mathbb{1}((\widehat{\pi}_a)^* > 0)}{(\widehat{\pi}_a)^*} dy\right\} &\leq \mathbb{E}\left\{\sqrt{n}\mathbb{1}(\widehat{\pi}_a = 0)\int_{\mathscr{Y}_{h,R_K}} h^{-d}\,\|K\|_2\, dy\right\} \\ &= \sqrt{n}(1-\pi_a)^n h^{-d}\,\|K\|_2 \operatorname{Vol}(\mathbb{B}_{R_K}(0)) \\ &\to 0,\end{aligned}$$

where the first inequality follows by assumption (A2). Then by Markov inequality,

$$\left|\sqrt{n}D((\widehat{Q_h^a})^*, \widehat{Q_h^a}) - \Phi(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n), \widehat{\pi}_a, (\widehat{\pi}_a)^*, \hat{q}_h^a)\right| \to 0 \text{ in probability.}$$

And hence,

$$\sqrt{n}D\left((\widehat{Q_h^a})^*, \widehat{Q_h^a}\right) \to \frac{1}{\pi_a}\int \left|\mathbb{G}f_{h,y}^a - q_h^a(y)\mathbb{G}\mathbb{1}_a\right| dy \text{ weakly in } \mathbb{R}.$$

□

B.2.9 Bootstrap validity of Theorem 3.4.3 for Multi-source randomized study

For this case, $\theta = \mathbb{E}_{\mathscr{P}}\left[D(Q_{h_1}^1, Q_{h_0}^0)\right]$ and $\hat{\theta} = \frac{1}{N}\sum_{i=1}^N D((\widehat{Q_{h_1}^1})_i, (\widehat{Q_{h_0}^0})_i)$. Then

$$\begin{aligned}|\hat{\theta} - \theta| &\leq \left|\frac{1}{N}\sum_{i=1}^N \left(D((\widehat{Q_{h_1}^1})_i, (\widehat{Q_{h_0}^0})_i) - D((Q_{h_1}^1)_i, (Q_{h_0}^0)_i)\right)\right| \\ &\quad + \left|\frac{1}{N}\sum_{i=1}^N D((Q_{h_1}^1)_i, (Q_{h_0}^0)_i) - \mathbb{E}_{\mathscr{P}}\left[D(Q_{h_1}^1, Q_{h_0}^0)\right]\right| \\ &\leq \frac{1}{N}\sum_{i=1}^N D((\widehat{Q_{h_1}^1})_i, (Q_{h_1}^1)_i) + \frac{1}{N}\sum_{i=1}^N D((\widehat{Q_{h_0}^0})_i, (Q_{h_0}^0)_i) \\ &\quad + \left|\frac{1}{N}\sum_{i=1}^N D((Q_{h_1}^1)_i, (Q_{h_0}^0)_i) - \mathbb{E}_{\mathscr{P}}\left[D(Q_{h_1}^1, Q_{h_0}^0)\right]\right|,\end{aligned}$$

hence one of the sufficient condition for the confidence interval $\hat{C}_\alpha$ to be valid is

$$\liminf_{n\to\infty}\mathbb{P}\Bigg(\frac{1}{N}\sum_{i=1}^{N}D(\widehat{(Q^1_{h_1})}_i,(Q^1_{h_1})_i)+\frac{1}{N}\sum_{i=1}^{N}D(\widehat{(Q^0_{h_0})}_i,(Q^0_{h_0})_i)$$
$$+\Bigg|\frac{1}{N}\sum_{i=1}^{N}D((Q^1_{h_1})_i,(Q^0_{h_0})_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\Bigg|\leq\frac{\bar{D}^1}{\sqrt{n}}+\frac{\bar{D}^0}{\sqrt{n}}+\frac{\hat{z}_\alpha}{\sqrt{N}}\Bigg)\geq 1-\alpha.$$

And this is implied from

$$\frac{1}{N}\sum_{i=1}^{N}\sqrt{n}D(\widehat{(Q^a_{h_a})}_i,(Q^a_{h_a})_i) \text{ and } \bar{D}^a \text{ converges to same limit,}$$
$$\liminf_{n\to\infty}\mathbb{P}\left(\sqrt{N}\Bigg|\frac{1}{N}\sum_{i=1}^{N}D((Q^1_{h_1})_i,(Q^0_{h_0})_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\Bigg|\leq\hat{z}_\alpha\right)\geq 1-\alpha.$$

And for the second one, it suffice to show that $\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}D((Q^1_{h_1})_i,(Q^0_{h_0})_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q^1_{h_1},Q^0_{h_0})\right]\right)$ and $\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}D((Q^1_{h_1})^*_i,(Q^0_{h_0})^*_i)-\frac{1}{N}\sum_{i=1}^{N}D((Q^1_{h_1})_i,(Q^0_{h_0})_i)\right)$ converges to same distribution, and then plugging in $\widehat{(Q^a_{h_a})}_i$ in place of $(Q^a_{h_a})_i$ when computing $\hat{z}_\alpha$.

Theorem B.2.5. *Under the assumptions (A1), (A2'),*

$$\frac{1}{N}\sum_{i=1}^{N}\sqrt{n}D(\widehat{(Q^1_{h_1})}_i,(Q^1_{h_1})_i)\to\mathbb{E}_{\mathscr{P}}\left[\frac{1}{\pi_a}\int\left|\mathbb{G}(f^a_{h,y})-q^a_h(y)\mathbb{G}(\mathbb{1}_a)\right|dy\right]\ \textit{a.s.},\tag{B.18}$$
$$\bar{D}^a\to\mathbb{E}_{\mathscr{P}}\left[\frac{1}{\pi_a}\int\left|\mathbb{G}(f^a_{h,y})-q^a_h(y)\mathbb{G}(\mathbb{1}_a)\right|dy\right]\ \textit{a.s.}.\tag{B.19}$$

Proof. For (B.18), from Theorem B.2.4 and stong law of large numbers,

$$\frac{1}{N}\sum_{i=1}^{N}\sqrt{n}D(\widehat{(Q^1_{h_1})}_i,(Q^1_{h_1})_i)\to\mathbb{E}_{\mathscr{P}}\left[\frac{1}{\pi_a}\int\left|\mathbb{G}(f^a_{h,y})-q^a_h(y)\mathbb{G}(\mathbb{1}_a)\right|dy\right]\text{ a.s..}$$

For (B.19), note that $\bar{D}^a=\frac{1}{N}\sum_{i=1}^{N}\sqrt{n}D(\widehat{(Q^a_{h_a})}^*_i,\widehat{(Q^a_{h_a})}_i)$. Then from Theorem B.2.4 and stong law of large numbers,

$$\bar{D}^a\to\mathbb{E}_{\mathscr{P}}\left[\frac{1}{\pi_a}\int\left|\mathbb{G}(f^a_{h,y})-q^a_h(y)\mathbb{G}(\mathbb{1}_a)\right|dy\right]\text{ a.s..}$$

□

Theorem B.2.6. *Under the assumptions (A1), (A2'),*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}D((Q_{h_1}^1)_i,(Q_{h_0}^0)_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q_{h_1}^1,Q_{h_0}^0)\right]\right)\to\mathscr{N}\left(0,Var_{\mathscr{P}}\left[D(Q_{h_1}^1,Q_{h_0}^0)\right]\right) \tag{B.20}$$

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}D((Q_{h_1}^1)_i^*,(Q_{h_0}^0)_i^*)-\frac{1}{N}\sum_{i=1}^{N}D((Q_{h_1}^1)_i,(Q_{h_0}^0)_i)\right)\to\mathscr{N}\left(0,Var_{\mathscr{P}}\left[D(Q_{h_1}^1,Q_{h_0}^0)\right]\right)\ a.s. \tag{B.21}$$

Proof. For (B.20), note that $\frac{1}{N}\sum_{i=1}^{N}D((Q_{h_1}^1)_i,(Q_{h_0}^0)_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q_{h_1}^1,Q_{h_0}^0)\right]=(\mathscr{P}_N-\mathscr{P})\Phi$, where $\Phi(\mathbb{P})=D(Q_{h_1}^1,Q_{h_0}^0)$. Hence from Central Limit Theorem,

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}D((Q_{h_1}^1)_i,(Q_{h_0}^0)_i)-\mathbb{E}_{\mathscr{P}}\left[D(Q_{h_1}^1,Q_{h_0}^0)\right]\right)\to\mathscr{N}\left(0,Var_{\mathscr{P}}\left[D(Q_{h_1}^1,Q_{h_0}^0)\right]\right).$$

For (B.21) note that $\frac{1}{N}\sum_{i=1}^{N}D((Q_{h_1}^1)_i^*,(Q_{h_0}^0)_i^*)-\frac{1}{N}\sum_{i=1}^{N}D((Q_{h_1}^1)_i,(Q_{h_0}^0)_i)=(\mathscr{P}_N^*-\mathscr{P}_N)\Phi$. Hence from (B.20) and Theorem (3.2.1),

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}D((Q_{h_1}^1)_i^*,(Q_{h_0}^0)_i^*)-\frac{1}{N}\sum_{i=1}^{N}D((Q_{h_1}^1)_i,(Q_{h_0}^0)_i)\right)\to\mathscr{N}\left(0,Var_{\mathscr{P}}\left[D(Q_{h_1}^1,Q_{h_0}^0)\right]\right)\ \text{a.s.}$$

□

B.2.10 Bootstrap validity of Theorem 3.4.3 for Observational study

For this case, $\theta=D(Q_h^1,Q_h^0)$ and $\hat{\theta}=D(\widehat{Q_h^1},\widehat{Q_h^0})$. Then $|\hat{\theta}-\theta|\le D(\widehat{Q_h^1},Q_h^1)+D(\widehat{Q_h^0},Q_h^0)$, hence one of the sufficient condition for the confidence interval $\hat{C}_\alpha$ to be valid is

$$\liminf_{n\to\infty}\mathbb{P}\big(D(\widehat{Q_h^1},Q_h^1)+D(\widehat{Q_h^0},Q_h^0)\le c_n\big)\ge 1-\alpha.$$

And this is implied from

$$\liminf_{n\to\infty}\mathbb{P}\left(\sqrt{n}D(\widehat{Q_h^1},Q_h^1)\le\hat{z}_{\alpha/2}^0\right)\ge 1-\frac{\alpha}{2},$$
$$\liminf_{n\to\infty}\mathbb{P}\left(\sqrt{n}D(\widehat{Q_h^0},Q_h^0)\le\hat{z}_{\alpha/2}^1\right)\ge 1-\frac{\alpha}{2}.$$

Hence it suffice to show that $\sqrt{n}D(\widehat{Q_h^a},Q_h^a)$ and $\sqrt{n}D(\widehat{Q_h^a}^*,\widehat{Q_h^a})$ converges to the same distribution.

For all $y \in \mathbb{R}^d$, let $T_{h,y} : \mathbb{R}^d \to \mathbb{R}$ be $T_{h,y}(y') = \frac{1}{h^d} K\left(\frac{\|y-y'\|_2}{h}\right)$, and let $\hat{f}^a_{h,y} : \mathbb{R}^k \times \{0,1\} \times \mathbb{R}^d \to \mathbb{R}$, $f^a_{h,y} : \mathbb{R}^k \times \{0,1\} \times \mathbb{R}^d \to \mathbb{R}$ be

$$\begin{aligned}
\hat{f}^a_{h,y}(x',a',y') &= \frac{\mathbb{1}_a(a')}{\hat{\pi}_a(x')}\left(T_{h,y}(y') - \hat{\mu}_{a'}(x')\right) + \hat{\mu}_a(x'), \\
f^a_{h,y}(x',a',y') &= \frac{\mathbb{1}_a(a')}{\bar{\pi}_a(x')}\left(T_{h,y}(y') - \bar{\mu}_{a'}(x')\right) + \bar{\mu}_a(x'),
\end{aligned}$$

and let $\mathscr{F}^a = \{f^a_{h,y} : y \in \mathbb{R}\}$.

Theorem B.2.7. *Under the assumptions (A1), (A2'), (B1), (B2), (B3),*

$$\sqrt{n}D(\hat{Q}^a_{h_a}, Q^a_{h_a}) \to \int |\mathbb{G}f^a_y| dy \text{ weakly in } \mathbb{R}, \tag{B.22}$$

$$\sqrt{n}D((\widehat{Q^a_h})^*, \widehat{Q^a_h}) \to \int |\mathbb{G}f^a_y| dy \text{ weakly in } \mathbb{R}, \tag{B.23}$$

where $\mathbb{G}$ is a centered Gaussian process with $Cov[\mathbb{G}(f), \mathbb{G}(g)] = \int fgd\mathbb{P} - \int fd\mathbb{P} \int gd\mathbb{P}$.

Proof. Note that we already have (B.22) is from Theorem 3.4.2, and we are to left show (B.23), which can be done by Theorem (3.2.1) and repetition of the proof of Theorem 3.4.2.

Combining Lemma B.2.7 and Theorem (3.2.1) implies that

$$\sqrt{n}(\mathbb{P}^*_n - \mathbb{P}_n) \to \mathbb{G} \text{ weakly in } \ell_\infty(\mathscr{F}^a).$$

Then as similar to proof of Theorem 3.4.2, $\sqrt{n}D((\widehat{Q^a_h})^*, \widehat{Q^a_h})$ can be expanded as

$$\begin{aligned}
\sqrt{n}D((\widehat{Q^a_h})^*, \widehat{Q^a_h}) &= \int \sqrt{n}\,|(\hat{\psi}^a_h)^*(y) - \hat{\psi}^a_h(y)|\, dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n}\,|(\hat{\psi}^a_h)^*(y) - \hat{\psi}^a_h(y)|\, dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n}\left|(\mathbb{P}^*_n - \mathbb{P}_n)\hat{f}^a_{h,y}\right| dy \\
&= \int_{\mathscr{Y}_{h,R_K}} \sqrt{n}\left|(\mathbb{P}^*_n - \mathbb{P}_n)f^a_{h,y} + (\mathbb{P}^*_n - \mathbb{P}_n)(\hat{f}^a_{h,y} - f^a_{h,y})\right| dy \\
&= \Phi(\sqrt{n}(\mathbb{P}^*_n - \mathbb{P}_n)) + r_n,
\end{aligned}$$

where we define $\Phi : \ell_\infty(\mathscr{F}_a) \to \mathbb{R}$ by $\Phi(\mu) = \int |\mu f^a_{h,y}| dy$. Now first note that

$$\begin{aligned} r_n &= \sqrt{n} D(\hat{Q}^a_{h_a}, Q^a_{h_a}) - \Phi(\sqrt{n}(\mathbb{P}_n - \mathbb{P})) \\ &\leq \int_{\mathscr{Y}_{h,R_K}} |\sqrt{n}(\mathbb{P}^*_n - \mathbb{P}_n)(\hat{f}^a_{h,y} - f^a_{h,y})| dy. \end{aligned}$$

Since Φ is continuous on $\ell_\infty(\mathscr{F}_a)$, by the continuous mapping theorem [e.g., 78, Theorem 7.7] and the previous result $\sqrt{n}(\mathbb{P}^*_n - \mathbb{P}_n) \to \mathbb{G}$, it follows

$$\Phi(\sqrt{n}(\mathbb{P}^*_n - \mathbb{P}_n)) \to \Phi(\mathbb{G}) = \int_{\mathscr{Y}_{h,R_K}} |\mathbb{G} f^a_{h,y}| dy \text{ a.s. weakly in } \mathbb{R}.$$

Moreover, under the condition (B1) and (B2'), by the previous result of (B.14) we have

$$|\sqrt{n}\mathbb{P}(\hat{f}^a_y - f^a_y)| \leq \sqrt{n} \left\| \frac{1}{\hat{\pi}_a} \right\|_\infty \|\widehat{\mu}_a - \overline{\mu}_a\|_2 \|\widehat{\pi}_a - \overline{\pi}_a\|_2 = o_{\mathbb{P}}(1).$$

Hence by Lemma B.2.4,

$$\begin{aligned} \left| \sqrt{n}(\mathbb{P}^*_n - \mathbb{P}_n)(\hat{f}^a_{h,y} - f^a_{h,y}) \right| &\leq \left| \sqrt{n}(\mathbb{P}^*_n - \mathbb{P})(\hat{f}^a_{h,y} - f^a_{h,y}) \right| + \left| \sqrt{n}(\mathbb{P}_n - \mathbb{P})(\hat{f}^a_{h,y} - f^a_{h,y}) \right| \\ &= o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1), \end{aligned}$$

which implies $\int_{\mathscr{Y}_{h,R_K}} |\sqrt{n}(\mathbb{P}^*_n - \mathbb{P}_n)(\hat{f}^a_{h,y} - f^a_{h,y})| dy = o_{\mathbb{P}}(1)$ and thus we obtain

$$r_n = o_{\mathbb{P}}(1).$$

Finally, putting these together we conclude

$$\sqrt{n} D((\widehat{Q^a_h})^*, \widehat{Q^a_h}) \to \int |\mathbb{G} f^a_y| dy \text{ a.s. weakly in } \mathbb{R}.$$

□

Appendix C

Supplementary Materials for Chapter 4

C.1 Additional Technical Details

C.1.1 Stability of the level set

This section supplements the concept of the level set stability that is used in Section 4.3.3.

Hausdorff distance is a common way of measuring difference between two sets that are embedded in the same space. Below we define the Hausdorff distance for any two subsets in Euclidean space.

Definition C.1.1 (Hausdorff distance)**.** *Let $A, B \subset \mathbb{R}^d$. Their Hausdorff distance $H(A,B)$ is defined as*

$$H(A,B) = \max\left\{\sup_{x\in A}\inf_{y\in B}\|x-y\|, \sup_{y\in B}\inf_{x\in A}\|x-y\|\right\}.$$

The Hausdorff distance can be equivalently defined as

$$H(A,B) = \inf\{\varepsilon \geq 0 : A \subset B_\varepsilon \text{ and } B \subset A_\varepsilon\},$$

where

$$A_\varepsilon := \{y \in \mathbb{R}^d : \text{there exists } x \in A \text{ with } \|x-y\| \leq \varepsilon\}.$$

When we estimate the target level set $L_{t,h} = \{p_h > t\}$ by the estimator $\hat{L}_t = \{\widehat{p_h} > t\}$, we rely on that the function difference $\|\widehat{p_h} - p_h\|_\infty$ is small. To transfer that to the set difference $H(L_{t,h}, \hat{L}_t)$, we need that the target level set $L_{t,h}$ doesn't change too much when the level t perturbs.

Definition C.1.2 (Level set stability)**.** *We say that the level set $L_{t,h} = \{w \in \mathbb{R}^2 : p_H(w) > t\}$ is stable if there exists $a > 0$ and $C > 0$ such that, for all $\delta < a$,*

$$H(L_{t-\delta,h}, L_{t+\delta,h}) \leq C\delta.$$

C.2 Proofs

C.2.1 k-means clustering: Theorem 4.3.1

Lemma C.2.1. *Suppose each $\widehat{\mu}_a$ is estimated in the separate sample set D^n with n samples.*

(a) The expectation of $\left\|\widehat{W}_i - W_i\right\|_2$ can be upper bounded as

$$\mathbb{P}\left[\left\|\widehat{W}_i - W_i\right\|_2\right] \leq \sum_a \|\widehat{\mu}_a - \mu_a\|_1 . \tag{C.1}$$

(b) Suppose Assumption (A4). For $\delta \in (0,1)$, $\frac{1}{n}\sum_{i=1}^n \left\|\widehat{W}_i - W_i\right\|_2$ can be bounded with probability at least $1-\delta$ as

$$\frac{1}{n}\sum_{i=1}^n \left\|\widehat{W}_i - W_i\right\|_2 \leq \sum_a \|\widehat{\mu}_a - \mu_a\|_1 + B\sqrt{\frac{\log(1/\delta)}{n}}. \tag{C.2}$$

Proof of Lemma C.2.1. (a)

$\mathbb{P}\left[\left\|\widehat{W}_i - W_i\right\|_2\right]$ can be upper bounded as

$$\begin{aligned}\mathbb{P}\left[\left\|\widehat{W}_i - W_i\right\|_2\right] &= \mathbb{P}\left[\sqrt{\sum_a (\widehat{\mu}_a(X) - \mu_a(X))^2}\right] \\ &\leq \sum_a \mathbb{P}\left[|\widehat{\mu}_a(X) - \mu_a(X)|\right] \\ &= \sum_a \|\widehat{\mu}_a - \mu_a\|_1 .\end{aligned}$$

(b)

We the high probability bound for $\frac{1}{n}\sum_{i=1}^n \left\|\widehat{W}_i - W_i\right\|_2$, note that Assumption (A4) implies $0 \leq \left\|\widehat{W}_i - W_i\right\|_2 \leq \sqrt{2}B$ a.s., and hence by Hoeffding's inequality,

$$P\left(\frac{1}{n}\sum_{i=1}^n \left\|\widehat{W}_i - W_i\right\|_2 - \mathbb{P}\left[\left\|\widehat{W}_i - W_i\right\|_2\right] > t\right) \leq \exp\left(-\frac{nt^2}{B^2}\right).$$

Hence for any $\delta > 0$, applying $t = B\sqrt{\frac{\log(1/\delta)}{n}}$ gives

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{W}_i - W_i\right\|_2 \leq \mathbb{P}\left[\left\|\widehat{W}_i - W_i\right\|_2\right] + B\sqrt{\frac{\log(1/\delta)}{n}}\right) \geq 1-\delta.$$

Then applying (C.1) gives

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\left\|\widehat{W}_i - W_i\right\|_2 \leq \sum_a \|\widehat{\mu}_a - \mu_a\|_1 + B\sqrt{\frac{\log(1/\delta)}{n}}\right) \geq 1-\delta.$$

□

Proof of Theorem 4.3.1

Proof of Theorem 4.3.1. We first bound $\left|R(\widehat{C}) - R(C^*)\right|$ by as

$$\begin{aligned}\left|R(\widehat{C}) - R(C^*)\right| &= R(\widehat{C}) - R(C^*)\\ &= R(\widehat{C}) - \widehat{R}_n(\widehat{C}) + \widehat{R}_n(\widehat{C}) - R(C^*)\\ &\leq R(\widehat{C}) - \widehat{R}_n(\widehat{C}) + \widehat{R}_n(C^*) - R(C^*)\\ &\leq 2\sup_{C\in\mathscr{C}_k}\left|R(C) - \widehat{R}_n(C)\right|.\end{aligned}$$

Then for any $C \in \mathscr{C}_k$, $\left|R(C) - \widehat{R}_n(C)\right|$ is further upper bounded as

$$\left|R(C) - \widehat{R}_n(C)\right| \leq |R(C) - R_n(C)| + \left|R_n(C) - \widehat{R}_n(C)\right|,$$

which yields

$$\mathbb{E}\left[\left|R(\widehat{C}) - R(C^*)\right|\right] \leq 2\mathbb{E}\left[\sup_{C\in\mathscr{C}_k}|R(C) - R_n(C)|\right] + 2\mathbb{E}\left[\sup_{C\in\mathscr{C}_k}\left|R_n(C) - \widehat{R}_n(C)\right|\right]. \quad \text{(C.3)}$$

For the first term of (C.3), we note that $||W_i||_2^2 \leq B^2$ by (A4), and hence Lemma 2, Theorem 1, and Remark of [89] altogether give its upper bound as

$$\begin{aligned}\mathbb{E}\left[2\sup_{C\in\mathscr{C}_k}|R(C)-R_n(C)|\right] &\leq 16B^2\sqrt{\frac{k(p+1)\log n}{n}}+o\left(\sqrt{\frac{\log n}{n}}\right)\\ &\leq 32B^2\sqrt{\frac{k(p+1)\log n}{n}},\end{aligned} \tag{C.4}$$

for large enough n.

For the second term of (C.3), note first that for any $C\in\mathscr{C}_k$, $\left|R_n(C)-\widehat{R}_n(C)\right|$ is upper bounded as

$$\begin{aligned}\left|R_n(C)-\widehat{R}_n(C)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}||W_i-\Pi_C[W_i]||_2^2-\frac{1}{n}\sum_{i=1}^{n}||\widehat{W}_i-\Pi_C[\widehat{W}_i]||_2^2\right|\\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left|||W_i-\Pi_C[W_i]||_2^2-||\widehat{W}_i-\Pi_C[\widehat{W}_i]||_2^2\right|\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(||W_i-\Pi_C[W_i]||_2+||\widehat{W}_i-\Pi_C[\widehat{W}_i]||_2\right)\left|||W_i-\Pi_C[W_i]||_2-||\widehat{W}_i-\Pi_C[\widehat{W}_i]||_2\right|\\ &\leq \frac{2\sqrt{2}B}{n}\sum_{i=1}^{n}\left|||W_i-\Pi_C[W_i]||_2-||\widehat{W}_i-\Pi_C[\widehat{W}_i]||_2\right|,\end{aligned} \tag{C.5}$$

where last line is from the boundedness assumption (A4). Now, note that for any $x,y\in\mathbb{R}^2$, $\Pi_C[x]=\min_{c\in C}||x-c||_2$ and the triangle inequality give

$$\begin{aligned}||x-\Pi_C[x]||_2-||y-\Pi_C[y]||_2 &\leq ||x-\Pi_C[y]||_2-||y-\Pi_C[y]||_2\\ &\leq ||(x-\Pi_C[y])-(y-\Pi_C[y])||_2\\ &= ||x-y||_2,\end{aligned}$$

and $||y-\Pi_C[y]||_2-||x-\Pi_C[x]||_2\leq||x-y||_2$ by symmetry as well, and hence

$$|||x-\Pi_C[x]||_2-||y-\Pi_C[y]||_2|\leq||x-y||_2.$$

Applying this result to (C.5) gives an upper bound for $\left|R_n(C)-\widehat{R}_n(C)\right|$ as

$$\left|R_n(C)-\widehat{R}_n(C)\right|\leq\frac{2\sqrt{2}B}{n}\sum_{i=1}^{n}\left\|\widehat{W}_i-W_i\right\|_2,$$

and the RHS bound is independent of C. Hence by applying (C.1) in Lemma C.2.1, the second term of (C.3) is further upper bounded as

$$\begin{aligned} 2\mathbb{E}\left[\sup_{C\in\mathscr{C}_k}\left|R_n(C)-\widehat{R}_n(C)\right|\right] &\leq 4\sqrt{2}B\mathbb{P}\left[\left\|\widehat{W}_i - W_i\right\|_2\right] \\ &\leq 4\sqrt{2}B\sum_a \|\widehat{\mu}_a-\mu_a\|_1 . \end{aligned} \tag{C.6}$$

Hence applying (C.4) and (C.6) to (C.3) gives the upper bound of $\mathbb{E}\left[\left|R(\widehat{C})-R(C^*)\right|\right]$ as

$$\mathbb{E}\left[\left|R(\widehat{C})-R(C^*)\right|\right] \leq 32B^2\sqrt{\frac{k(p+1)\log n}{n}} + 4\sqrt{2}B\sum_a \|\widehat{\mu}_a-\mu_a\|_1 .$$

□

C.2.2 hierarchical clustering: Lemma 4.3.1, Theorem 4.3.2

Proof of Lemma 4.3.1

Proof. We consider a pair of points $W_1 = (\mu_1(X_1),...,\mu_p(X_1))$, $W_2 = (\mu_1(X_2),...,\mu_p(X_2))$, and their estimates $\widehat{W_1} = \left(\widehat{\mu_1}(X_1),...,\widehat{\mu_p}(X_1)\right)$, $\widehat{W_2} = \left(\widehat{\mu_1}(X_2),...,\widehat{\mu_p}(X_2)\right)$ for $\forall X_1, X_2 \in \mathscr{X}$. To prove the theorem, first we upper bound the maximum discrepancy between $d(W_1,W_2)$ and $d(\widehat{W_1},\widehat{W_2})$ as below.

Lemma C.2.2. *For Euclidean distance d, we have*

$$\left|d(W_1,W_2)-d(\widehat{W_1},\widehat{W_2})\right| \leq 2p\,\|\widehat{\mu}-\mu\|_\infty .$$

Proof. We have $\|x\|_2 - \|y\|_2 \leq \|x-y\|_2$ for $\forall x,y$ in the same metric space. Hence by definition,

$$
\begin{aligned}
& d(W_1,W_2) - d(\widehat{W_1},\widehat{W_2}) \\
\leq & \sqrt{\sum_a \{\mu_a(X_1) - \mu_a(X_2) - (\widehat{\mu}_a(X_1) - \widehat{\mu}_a(X_2))\}^2} \\
= & \sqrt{\sum_a \{\mu_a(X_1) - \widehat{\mu}_a(X_1) - (\mu_a(X_2) - \widehat{\mu}_a(X_2))\}^2} \\
\leq & \sum_{j=1}^{2} \sum_{a\in\mathscr{A}} \left|\widehat{\mu}_a(X_j) - \mu_a(X_j)\right| \\
\leq & 2 \sum_{a\in\mathscr{A}} \|\widehat{\mu}_a - \mu_a\|_\infty .
\end{aligned}
$$

□

For the proof of Lemma 4.3.1, consider two sets A,B and their estimates $\widehat{A} = \{\widehat{W} : W \in A\}$, $\widehat{B} = \{\widehat{W} : W \in B\}$ respectively. Let $(a^*,b^*) = \underset{a\in A,b\in B}{\operatorname{argmin}} d(a,b)$ and $\widehat{a^*},\widehat{b^*}$ be their estimates. Then by definition of single linkage we have

$$
\begin{aligned}
\left|D(A,B) - D(\widehat{A},\widehat{B})\right| &= \left|\min_{\hat{a}\in\widehat{A},\hat{b}\in\widehat{B}} d(\hat{a},\hat{b}) - d(a^*,b^*)\right| \\
&\leq \left|d(\widehat{a^*},\widehat{b^*}) - d(a^*,b^*)\right| \\
&\leq 2\sum_{a\in\mathscr{A}} \|\widehat{\mu}_a - \mu_a\|_\infty .
\end{aligned}
$$

The exact same result follows for the case of complete linkage. The result for average linkage directly follows by Lemma C.2.2.

□

Proof of Theorem 4.3.2

As before, we will let $\boldsymbol{\mu}$ denote the conditional counterfactual mean vector space. Further by Assumption A5, we assume that every distribution satisfying the good neighborhood property in Definition 4.3.1 has a density bounded by $p_{\boldsymbol{\mu}} < \infty$. We begin with introducing some useful lemmas.

Lemma C.2.3. *Under Assumption A5, for any $W \in \boldsymbol{\mu}$*

$$\sup_{w \in \mathbb{R}^p, r>0} \mathbb{P}\left(W \in \mathbb{B}(w, r+s) \backslash \mathbb{B}(w,r)\right) \leq \mathsf{C}_1 s,$$

where C_1 is a constant that depends on $p_{\boldsymbol{\mu}}$, B, and p.

Proof. Let λ_p be the p-dimensional Lebesgue measure. By Assumption (A4), $\operatorname{supp}(W) \subset [-2B, 2B]^p$, and hence for any $w \in \mathbb{R}^p$ and $r, s > 0$,

$$\lambda_p\left(\{\mathbb{B}(w, r+s) \backslash \mathbb{B}(w,r)\} \cap \operatorname{supp}(W)\right) \leq \lambda_p\left(\{\mathbb{B}(w, r+s) \backslash \mathbb{B}(w,r)\} \cap [-2B, 2B]^p\right).$$

Now, we bound $\lambda_{p-1}(\partial \mathbb{B}(w,t) \cap [-2B, 2B]^p)$ for any $t \in \mathbb{R}$. First, note that for any $u \geq 0$, by considering that the map $\varphi : \partial \mathbb{B}(w,t) \cap [-2B, 2B]^p \to \partial \mathbb{B}(w, t+u) \cap [-2B-u, 2B+u]^p$ by $\varphi(w + tv) = w + (t+u)v$ for unit vector v satisfies $\|\varphi(x) - \varphi(y)\| \geq \|x - y\|$, we have

$$\lambda_{p-1}(\partial \mathbb{B}(w,t) \cap [-2B, 2B]^p) \leq \lambda_{p-1}(\partial \mathbb{B}(w, t+u) \cap [-2B-u, 2B+u]^p).$$

And hence

$$\begin{aligned}
\frac{2B}{p}\lambda_{p-1}(\partial \mathbb{B}(w,t) \cap [-2B, 2B]^p) &= \int_0^{\frac{2B}{p}} \lambda_{p-1}(\partial \mathbb{B}(w,t) \cap [-2B, 2B]^p) du \\
&\leq \int_0^{\frac{2B}{p}} \lambda_{p-1}\left(\partial \mathbb{B}(w, t+u) \cap [-2B-u, 2B+u]^p\right) du \\
&\leq \int_0^{\frac{2B}{p}} \lambda_{p-1}\left(\partial \mathbb{B}(w, t+u) \cap \left[-2(1+\frac{1}{p})B, 2(1+\frac{1}{p})B\right]^p\right) du \\
&= \lambda_p\left(\mathbb{B}(w, t+B) \backslash \mathbb{B}(w,t)) \cap \left[-2(1+\frac{1}{p})B, 2(1+\frac{1}{p})B\right]^p\right) \\
&\leq \lambda_p\left(\left[-2(1+\frac{1}{p})B, 2(1+\frac{1}{p})B\right]^p\right) \leq e4^p B^p,
\end{aligned}$$

and hence

$$\lambda_{p-1}(\partial \mathbb{B}(w,t) \cap [-2B, 2B]^p) \leq e2^{2p-1}B^{p-1}p.$$

Then $\lambda_p\left((\mathbb{B}(w, r+s) \backslash \mathbb{B}(w,r)) \cap [-2B, 2B]^p\right)$ is bounded as

$$\begin{aligned}
\lambda_p\left((\mathbb{B}(w, r+s) \backslash \mathbb{B}(w,r)) \cap [-2B, 2B]^p\right) &= \int_0^s \lambda_{p-1}(\partial \mathbb{B}(w, r+t) \cap [-2B, 2B]^p) dt \\
&\leq \int_0^s e2^{2p-1}B^{p-1}p dt = e2^{2p-1}B^{p-1}ps.
\end{aligned}$$

And hence for all $w \in \mathbb{R}^p$ and $r > 0$, Under Assumption A5,

$$\begin{aligned}\mathbb{P}(W \in \mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r)) &\leq p_{\boldsymbol{\mu}} \int_{(\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))\cap \mathrm{supp}(W)} \lambda_p(dw) \\ &\leq e p_{\boldsymbol{\mu}} 2^{2p-1} B^{p-1} ps.\end{aligned}$$

□

Lemma C.2.4. *With probability* $1-\delta_n$,

$$\begin{aligned}&\sup_{w\in\mathbb{R}^p, r>0}\left|\frac{|S\cap(\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))|}{n} - \mathbb{P}(W\in\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))\right| \\ &\leq \mathsf{C}_2\left(\frac{1}{n}\log(\frac{1}{\delta_n}) + \sqrt{\frac{s}{n}\log\left(\frac{1}{s}\right)} + \sqrt{\frac{s}{n}\log(\frac{1}{\delta_n})}\right),\end{aligned}$$

where C_2 *is a constant depending only on* p*,* B*,* $p_{\boldsymbol{\mu}}$*.*

Proof. For $w \in \mathbb{R}^p$ and $r,s > 0$, let $B_{w,r,s} := \mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r)$, and let $\mathscr{F}_s := \left\{\mathbb{1}_{B_{w,r,s}} : w \in \mathbb{R}^p, r > 0\right\}$. Then

$$\begin{aligned}&\sup_{w\in\mathbb{R}^p, r>0}\left|\frac{|S\cap(\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))|}{n} - \mathbb{P}(W\in\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))\right| \\ &= \sup_{f\in\mathscr{F}_s}\left|\frac{1}{n}\sum_{i=1}^n f(W_i) - \mathbb{E}[f(W_i)]\right|.\end{aligned}$$

Now, for $w \in \mathbb{R}^p$ and $r > 0$, let $B_{w,r} := \mathbb{B}(w,r)$ and $\tilde{B}_{w,r} := \mathbb{R}^p\backslash\mathbb{B}(w,r)$, and let $\mathscr{H} := \{B_{w,r} : w \in \mathbb{R}^p, r > 0\}$ and $\tilde{\mathscr{H}} := \left\{\tilde{B}_{w,r} : w \in \mathbb{R}^p, r > 0\right\}$. Then the VC dimension of $\mathscr{H}$ or $\tilde{\mathscr{H}}$ is no greater than $p+2$. Therefore, let $s(\mathscr{H},n)$ and $s(\tilde{\mathscr{H}},n)$ be shattering number of $\mathscr{H}$ and $\tilde{\mathscr{H}}$, respectively, then by Sauer's Lemma for $n \geq p+2$,

$$s(\mathscr{H},n) \leq \left(\frac{en}{p+2}\right)^{p+2} \qquad \text{and} \qquad s(\tilde{\mathscr{H}},n) \leq \left(\frac{en}{p+2}\right)^{p+2}.$$

Now, let $\mathscr{G}_s := \{B_{w,r,s} : w \in \mathbb{R}^p, r > 0\}$, then $\mathscr{G}_s \subset \left\{A\cap B : A \in \mathscr{H}, B \in \tilde{\mathscr{H}}\right\}$, and hence for $n \geq p+2$,

$$s(\mathscr{G}_s,n) \leq s(\mathscr{H},n)s(\tilde{\mathscr{H}},n) \leq \left(\frac{en}{p+2}\right)^{2p+4}.$$

Then, for $n = (2p+4)^2$,

$$\begin{aligned} s(\mathscr{G}_s, (2p+4)^2) &\leq (2e(2p+4))^{2p+4} \\ &\leq (2^{2p+4})^{2p+4} = 2^{(2p+4)^2}, \end{aligned}$$

so VC dimension of $\mathscr{G}_s$ is bounded by $(2p+4)^2$. Then from Theorem 2.6.4 in Van Der Vaart and Wellner [142],

$$\begin{aligned} \mathscr{N}(\mathscr{F}_s, \|\cdot\|, \varepsilon) &\leq K(2p+4)^2(4e)^{(2p+4)^2}\left(\frac{1}{\varepsilon}\right)^{2((2p+4)^2-1)} \\ &\leq \left(\frac{8K(p+2)e}{\varepsilon}\right)^{2((2p+4)^2-1)}, \end{aligned}$$

for some universal constant K. Now, for all $f \in \mathscr{F}_s$, $\mathbb{E}_P f^2 \leq C_{B,p_{\boldsymbol{\mu}}} ps$. Hence, by Theorem 30 in Kim et al. [76], with probability $1-\delta_n$,

$$\begin{aligned} &\sup_{f\in\mathscr{F}_s}\left|\frac{1}{n}\sum_{i=1}^{n} f(W_i) - \mathbb{E}\left[f(W_i)\right]\right| \\ &\leq C\left(\frac{\nu_p}{n}\log(2A_p) + \sqrt{\frac{\nu_p \mathsf{C}_3 s}{n}\log\left(\frac{2A_p}{\mathsf{C}_3 s}\right)} + \sqrt{\frac{\mathsf{C}_3 s \log(\frac{1}{\delta_n})}{n}} + \frac{\log(\frac{1}{\delta_n})}{n}\right), \end{aligned}$$

where $\nu_p = 2((2p+4)^2-1)$ and $A_p = 8K(p+2)e$. Hence, it can be simplified as

$$\sup_{f\in\mathscr{F}_s}\left|\frac{1}{n}\sum_{i=1}^{n} f(W_i) - \mathbb{E}\left[f(W_i)\right]\right| \leq \mathsf{C}_2\left(\frac{1}{n}\log(\frac{1}{\delta_n}) + \sqrt{\frac{s}{n}\log\left(\frac{1}{s}\right)} + \sqrt{\frac{s}{n}\log(\frac{1}{\delta_n})}\right),$$

where C_2 is a constant depending only on p, B, $p_{\boldsymbol{\mu}}$.

□

Corollary C.2.1. *Under Assumption A5, with probability* $1-\delta_n$*,*

$$\sup_{w\in\mathbb{R}^p, r>0}\frac{|S\cap(\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))|}{n} \leq \mathsf{C}_3\left(s + \frac{1}{n}\log(\frac{1}{\delta_n}) + \sqrt{\frac{s}{n}\log\left(\frac{1}{s}\right)}\right),$$

where C_3 *is a constant depending only on* p*,* B*,* $p_{\boldsymbol{\mu}}$*.*

Proof.

$$\begin{aligned}
&\sup_{w\in\mathbb{R}^p, r>0} |S\cap(\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))| \\
&\qquad \le \sup_{w\in\mathbb{R}^p, r>0} \mathbb{P}(W\in\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r)) \\
&\qquad\quad + \sup_{w\in\mathbb{R}^p, r>0} \left|\frac{|S\cap(\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))|}{n} - \mathbb{P}(W\in\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))\right|.
\end{aligned}$$

Then from Lemma C.2.3 and C.2.4,

$$\begin{aligned}
&\sup_{w\in\mathbb{R}^p, r>0} |S\cap(\mathbb{B}(w,r+s)\backslash\mathbb{B}(w,r))| \\
&\qquad \le \mathsf{C}_1' s + \mathsf{C}_2\left(\frac{1}{n}\log(\frac{1}{\delta_n}) + \sqrt{\frac{s}{n}\log\left(\frac{1}{s}\right)} + \sqrt{\frac{s}{n}\log(\frac{1}{\delta_n})}\right) \\
&\qquad \le \mathsf{C}_1' s + \mathsf{C}_2\left(\frac{1}{n}\log(\frac{1}{\delta_n}) + \sqrt{\frac{s}{n}\log\left(\frac{1}{s}\right)} + \frac{1}{2}\left(s + \frac{1}{n}\log(\frac{1}{\delta_n})\right)\right) \\
&\qquad \le \mathsf{C}_3\left(s + \frac{1}{n}\log(\frac{1}{\delta_n}) + \sqrt{\frac{s}{n}\log\left(\frac{1}{s}\right)}\right),
\end{aligned}$$

where $\mathsf{C}_3 = \max\left\{\mathsf{C}_1' + \frac{1}{2}\mathsf{C}_2, \frac{3}{2}\mathsf{C}_2\right\}$.

□

Lemma C.2.5. *Suppose $\mathsf{U}^N = \{W_1, ..., W_N\}$ are i.i.d samples from the mixture distribution $\mathbb{P}_{\alpha,\nu}$ defined in Definition 4.3.1. Then with probability $1-\delta_n$, the similarity function K constructed on U^N satisfies (α', ν')-good neighborhood property for the clustering problem (U^N, l), where*

$$\alpha' = \alpha + O\left(\sqrt{\frac{1}{N}\log\frac{1}{\delta_N}}\right) \quad \textit{and} \quad \nu' = \nu + O\left(\sqrt{\frac{1}{N}\log\frac{1}{\delta_N}}\right).$$

Proof. For any $\delta_N \in (0,1)$, by Hoeffding's inequality we have

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{W_i \sim \mathbb{P}_{\text{noise}}\} \ge \nu + \sqrt{\frac{B}{N}\log\frac{2}{\delta_N}}$$

with probability at most $\delta_N/2$. Again by Hoeffding's inequality, for all points $w \in \mathsf{U}^N$ we have

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{W_i \sim \mathbb{P}_\alpha \text{ and } W_i \in \mathbb{B}(w, r_w)\setminus\mathbb{C}(w)\} \ge \alpha + \sqrt{\frac{B}{N}\log\frac{2}{\delta_N}}$$

with probability at most $\delta_N/2$, as $\mathbb{P}_\alpha\{W \in \mathbb{B}(w,r_w) \setminus \mathbb{C}(w)\} \leq \alpha$ by the given condition. Therefore by definition, it follows that with probability at least $1-\delta_N$ the similarity function K satisfies $\left(\alpha + \sqrt{\frac{B}{N}\log\frac{2}{\delta_N}}, \nu + \sqrt{\frac{B}{N}\log\frac{2}{\delta_N}}\right)$-good neighborhood property. □

Proof of Theorem 4.3.2

Proof. Since the similarity function K satisfies (α,ν)-good property, there exists some subset $S' \subset S$ of size $(1-\nu)n$ such that for all points $w \in S'$ all but αn out of $n_{C(w)\cap S'}$ neighbors in S' belongs to the cluster $C(w)$. For each $w \in S'$, let $r_{S',w} := \inf\{r \geq 0 : \mid S' \cap B(w,r) \mid \geq n_{C(w)\cap S'}\}$ be the distance to the $n_{C(w)\cap S'}$-th nearest neighbor of w in S'. Then it follows $\mid S' \cap B(w, r_{S',w}) \setminus \mathbb{C}(w) \mid \leq \alpha n$.

Now we let $\gamma := \sum_{a\in\mathscr{A}} \|\widehat{\mu}_a - \mu_a\|_\infty$, and define β by

$$\beta := \sup_{w\in S'} \frac{\left|S' \cap (B(w, r_{S',w} + 4\gamma) \setminus B(w, r_{S',w}))\right|}{n}.$$

Then from Corollary C.2.1, under Assumption A5 with probability $1-\delta_n$,

$$\begin{aligned}\beta &\leq \sup_{w\in\mathbb{R}^p, r>0} \frac{|S \cap (B(w, r+4\gamma) \setminus \mathbb{B}(w,r))|}{n} \\ &\leq 4\mathsf{C}_3\left(\gamma + \frac{1}{n}\log(\frac{1}{\delta_n}) + \sqrt{\frac{\gamma}{n}\log\left(\frac{1}{\gamma}\right)}\right).\end{aligned}$$

Hence, $\beta = O(\gamma + \frac{1}{n}\log(\frac{1}{\delta_n}))$.

Let $\hat{w}$ be an estimate of w. Now, note that $d(w,w') \leq r_{S',w}$ implies $d(\hat{w},\hat{w}') \leq r_{S',w} + 2\gamma$, and hence $w' \in S' \cap B(w, r_{S',w})$ implies $\hat{w}' \in \hat{S}' \cap B(\hat{w}, r_{S',w} + 2\gamma)$. Hence

$$\left|\hat{S}' \cap B(\hat{w}, r_{S',w} + 2\gamma)\right| \geq \left|S' \cap B(w, r_{S',w})\right| \geq n_{C(w)\cap S'} = n_{C(\hat{w})\cap\hat{S}'}.$$

Therefore by definition,

$$r_{\hat{S}',\hat{w}} \leq r_{S',w} + 2\gamma.$$

Also, note that $d(\hat{w},\hat{w}') \leq r_{S',w} + 2\gamma$ implies $d(w,w') \leq r_{S',w} + 4\gamma$, and thereby $\hat{w}' \in \hat{S}' \cap B(\hat{w}, r_{S',w} + 2\gamma)$ implies $x' \in S' \cap B(w, r_{S',w} + 4\gamma)$. Thus we have

$$\begin{aligned}\left|\hat{S}' \cap B(\hat{w}, r_{S',w} + 2\gamma) \setminus \hat{C}(w)\right| &\leq \left|S' \cap B(w, r_{S',w} + 4\gamma) \setminus \mathbb{C}(w)\right| \\ &\leq \left|S' \cap B(w, r_{S',w}) \setminus \mathbb{C}(w)\right| + \left|S' \cap (B(w, r_{S',w} + 4\gamma) \setminus B(w, r_{S',w}))\right| \\ &\leq (\alpha + \beta)n,\end{aligned}$$

which leads to

$$\left|\hat{S}' \cap B(\hat{w}, r_{\hat{S}',\hat{w}}) \backslash \hat{C}(w)\right| \leq \left|\hat{S}' \cap B(\hat{w}, r_{S',w} + 2\gamma) \backslash \hat{C}(w)\right| \leq (\alpha+\beta)n.$$

Consequently, K satisfies $(\alpha+\beta, \nu)$-good property for the clustering problem $(\hat{S}, l)$. Then the result follows from Theorem 11 in Balcan et al. [7].

□

C.2.3 density clustering: Theorem 4.3.3

Theorem C.2.1. *Suppose that $L_{h,t}$ is stable and let $H(\cdot,\cdot)$ be the Hausdorff distance between two sets. Suppose each $\widehat{\mu}_a$ is estimated in the separate sample set D^n with size n, and suppose that assumptions (A1)-(A6) hold. Let $\delta \in (0,1)$ and $\{h_n\}_{n\in\mathbb{N}} \subset (0,h_0)$ be satisfying*

$$\limsup_n \frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d} < \infty.$$

Then, with probability at least $1-\delta$,

$$H(\widehat{L}_{h_n,t}, L_{h_n,t}) \leq \mathsf{C}_{P,K,B}\left(\sqrt{\frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d}}\right.$$
$$\left. + \frac{1}{h_n^{d+1}} \min\left\{\sum_a \|\widehat{\mu}_a - \mu_a\|_1 + \sqrt{\frac{\log(2/\delta)}{n}}, h_n\right\}\right)$$

In order to show Theorem C.2.1, we need the following Lemma.

Lemma C.2.6. *Suppose each $\widehat{\mu}_a$ is estimated in the separate sample set D^n, and suppose that assumptions (A1)-(A6) hold. Let $\delta \in (0,1)$ and $\{h_n\}_{n\in\mathbb{N}} \subset (0,h_0)$ be satisfying*

$$\limsup_n \frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d} < \infty.$$

Then, with probability at least $1-\delta$,

$$\|\widehat{p_{h_n}} - p_{h_n}\|_\infty \leq \mathsf{C}_{P,K,B}\left(\sqrt{\frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d}} + \frac{1}{h_n^{d+1}} \min\left\{\sum_a \|\widehat{\mu}_a - \mu_a\|_1 + \sqrt{\frac{\log(2/\delta)}{n}}, h_n\right\}\right).$$

for some constant $\mathsf{C}_{P,K,B}$ depending only on P, K, B.

For showing Lemma (C.2.6), we note that $\|\widehat{p}_h - p_h\|_\infty$ can be upper bounded as

$$\|\widehat{p}_{h_n} - p_{h_n}\|_\infty \leq \|\widetilde{p}_{h_n} - p_{h_n}\|_\infty + \|\widehat{p}_{h_n} - \widetilde{p}_{h_n}\|_\infty . \tag{C.7}$$

Therefore, in what follows we shall provide high probability bound for $\|\widetilde{p}_{h_n} - p_{h_n}\|_\infty$ in Lemma (C.2.7) and $\|\widehat{p}_{h_n} - \widetilde{p}_{h_n}\|_\infty$ in Lemma (C.2.8). Then applying these to (C.7) will conclude the proof.

The following is from applying Kim et al. [76, Corollary 13].

Lemma C.2.7. *Under Assumptions (A1)-(A6), if we let $\delta \in (0,1)$ and $\{h_n\}_{n\in\mathbb{N}} \subset (0,h_0)$ be satisfying*

$$\limsup_n \frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d} < \infty,$$

then with probability at least $1-\delta$ it follows

$$\|\widetilde{p}_{h_n} - p_{h_n}\|_\infty \leq \mathsf{C}_{P,K}\sqrt{\frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d}},$$

where C depends only on P and K.

Proof. Consider $\mathbb{X} = \mathbb{B}(0, B+h_0)$. Then by Assumption (A4) for $\forall w \in \mathbb{R}^d \backslash \mathbb{X}$ it follows

$$\frac{\|W_i - w\|_2}{h} > 1.$$

$\text{supp}(K) \subset \overline{B(0,1)}$ from Assumption (A6) implies that

$$\widetilde{p}_{h_n}(w) = \frac{1}{n}\sum_{i=1}^n K\left(\frac{W_i - w}{h}\right) = 0 \text{ a.s.},$$

and consequently that $p_{h_n}(w) = 0$ as well. Therefore,

$$\|\widetilde{p}_{h_n} - p_{h_n}\|_\infty = \sup_{w\in\mathbb{X}} |\widetilde{p}_{h_n}(w) - p_{h_n}(w)| . \tag{C.8}$$

Since under (A5), P has bounded density p, so by Kim et al. [76, Proposition 5] we have that

$$\limsup_{r\to 0} \sup_{x\in\mathbb{X}} \frac{\int_{\mathbb{B}(x,r)} p(w)dw}{r^d} < \infty.$$

Now note that under (A6), we have that $|K(x)-K(y)| \leq M_K \|x-y\|_2$ for any $x,y \in \mathbb{R}^d$ and $\mathrm{supp}(K) \subset \overline{\mathbb{B}(0,1)}$, which together implies that $\|K\|_\infty \leq M_K < \infty$. Hence,

$$\int_0^\infty t \sup_{\|x\| \geq t} K^2(x)dt \leq \int_0^1 t M_K^2 dt = \frac{1}{2} M_K^2 < \infty.$$

Then applying Kim et al. [76, Corollary 13] gives that with probability at least $1-\delta$,

$$\sup_{w \in \mathbb{X}} |\widetilde{p}_{h_n}(w) - p_{h_n}(w)| \leq \mathsf{C}_{P,K} \sqrt{\frac{(\log(1/h_n))_+ + \log(2/\delta)}{n h_n^d}}, \tag{C.9}$$

where $\mathsf{C}_{P,K}$ depends only on P and K. Finally (C.8) and (C.9) together imply that with probability at least $1-\delta$, we have

$$\|\widetilde{p}_{h_n} - p_{h_n}\|_\infty \leq \mathsf{C}_{P,K} \sqrt{\frac{(\log(1/h_n))_+ + \log(2/\delta)}{n h_n^d}}.$$

□

Lemma C.2.8. *Suppose Assumptions (A1)-(A4) and (A6). Then*

$$\|\widehat{p}_{h_n} - \widetilde{p}_{h_n}\|_\infty \leq \frac{\mathsf{C}_{M_K,B}}{h_n^{d+1}} \min\left\{ \sum_a \mathbb{E}\left[\|\widehat{\mu}_a - \mu_a\|_1\right] + \sqrt{\frac{\log(1/\delta)}{n}}, h_n \right\},$$

where $\mathsf{C}_{M_K,B}$ depends only on M_K and B.

Proof. By Assumption (A6) it follows that $|K(x)-K(y)| \leq M_K \|x-y\|_2$ for any $x,y \in \mathbb{R}^d$ and $\mathrm{supp}(K) \subset \overline{\mathbb{B}(0,1)}$, which together implies that $|K(x)-K(y)| \leq M_K$ and $\|K\|_\infty \leq M_K$. Thus it follows

$$|K(x)-K(y)| \leq \min\left\{M_K \|x-y\|_2, M_K\right\}.$$

Now for any $w \in \mathbb{R}^d$, $|\widehat{p}_{h_n}(w) - \widetilde{p}_{h_n}(w)|$ is upper bounded by

$$\begin{aligned}
|\widehat{p}_{h_n}(w) - \widetilde{p}_{h_n}(w)| &\leq \frac{1}{n h_n^d} \sum_{i=1}^n \left| K\left(\frac{\widehat{W}_i - w}{h_n}\right) - K\left(\frac{W_i - w}{h_n}\right) \right| \\
&\leq \frac{1}{n h_n^d} \sum_{i=1}^n \min\left\{ M_K \frac{\left\|\widehat{W}_i - W_i\right\|_2}{h_n}, M_K \right\} \\
&\leq \frac{M_K}{h_n^{d+1}} \min\left\{ \frac{1}{n} \sum_{i=1}^n \left\|\widehat{W}_i - W_i\right\|_2, h_n \right\}.
\end{aligned}$$

Since this holds for any $w \in \mathbb{R}^d$,

$$\|\widehat{p}_{h_n} - \widetilde{p}_{h_n}\|_\infty \leq \frac{M_K}{h_n^{d+1}} \min\left\{ \frac{1}{n}\sum_{i=1}^{n} \left\|\widehat{W}_i - W_i\right\|_2, h_n \right\}.$$

Then under (A4), applying (C.2) from Lemma C.2.1 gives that with probability $1-\delta$, $\|\widehat{p}_{h_n} - \widetilde{p}_{h_n}\|_\infty$ is upper bounded as

$$\begin{aligned}
\|\widehat{p}_{h_n} - \widetilde{p}_{h_n}\|_\infty &\leq \frac{M_K}{h_n^{d+1}} \min\left\{ \|\widehat{\mu}_a - \mu_a\|_1 + 2B\sqrt{\frac{\log(1/\delta)}{n}}, h_n \right\} \\
&\leq \frac{\mathsf{C}_{M_K,B}}{h_n^{d+1}} \min\left\{ \|\widehat{\mu}_a - \mu_a\|_1 + \sqrt{\frac{\log(1/\delta)}{n}}, h_n \right\},
\end{aligned}$$

where $\mathsf{C}_{M_K,B} = M_K \max\{1, 2B\}$.

□

Now we are ready to prove Lemma C.2.6.

Proof of Lemma C.2.6. As in (C.7), we upper bound $\|\widehat{p}_{h_n} - p_{h_n}\|_\infty$ as

$$\|\widehat{p}_{h_n} - p_{h_n}\|_\infty \leq \|\widehat{p}_{h_n} - \widetilde{p}_{h_n}\|_\infty + \|\widetilde{p}_{h_n} - p_{h_n}\|_\infty.$$

Then by Lemma C.2.7 and C.2.8, with probability $1-\delta$ it follows that

$$\begin{aligned}
\|\widehat{p}_{h_n} - p_{h_n}\|_\infty &\leq \mathsf{C}_{P,K}\sqrt{\frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d}} + \frac{\mathsf{C}_{M_K,B}}{h_n^{d+1}} \min\left\{ \sum_a \|\widehat{\mu}_a - \mu_a\|_1 + \sqrt{\frac{\log(2/\delta)}{n}}, h_n \right\} \\
&\leq \mathsf{C}_{P,K,B}\left(\sqrt{\frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d}} + \frac{1}{h_n^{d+1}} \min\left\{ \sum_a \|\widehat{\mu}_a - \mu_a\|_1 + \sqrt{\frac{\log(2/\delta)}{n}}, h_n \right\} \right),
\end{aligned}$$

where $\mathsf{C}_{P,K,B}$ depends only on P, K, B.

□

Proof of Theorem 4.3.3

Recall that $L_{h_n,t}$ is stable if there exist $a > 0$ and $C > 0$ such that, for all $0 < \zeta < a$, $H(L_{h_n,t-\zeta}, L_{h_n,t+\zeta}) \leq C\zeta$.

Proof. Let us define

$$r_n := \mathsf{C}_{P,K,B}\left(\sqrt{\frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d}} + \frac{1}{h_n^{d+1}}\min\left\{\sum_a \|\widehat{\mu}_a - \mu_a\|_1 + \sqrt{\frac{\log(2/\delta)}{n}}, h_n\right\}\right),$$

which is RHS of the inequality in Lemma C.2.6.

Suppose that we are given a sufficiently large n so that $\|\widehat{p}_{h_n} - p_{h_n}\|_\infty < r_n$ holds with probability at least $1-\delta$ where $r_n < a$ for some constant $a > 0$. We aim to show two things: (a) for every $x \in L_{h_n,t}$ there exists $y \in \widehat{L}_{h_n,t}$ with $\|x-y\|_2 \le Cr_n$, and (b) for every $x \in \widehat{L}_{h_n,t}$ there exists $y \in L_{h_n,t}$ with $\|x-y\|_2 \le Cr_n$.

To show (a), consider $x \in L_{h_n,t}$, Then by the stability property of $L_{h_n,t}$, there exists $y \in L_{h_n,t+r_n}$ such that $\|x-y\|_2 \le Cr_n$. Then $p_{h_n}(y) > t + r_n$ which implies that

$$\widehat{p}_{h_n}(y) \ge p_{h_n}(y) - \|\widehat{p}_{h_n} - p_{h_n}\|_\infty > p_{h_n}(y) - r_n > t.$$

Hence we conclude $y \in \widehat{L}_{h_n,t}$ with $\|x-y\|_2 \le Cr_n$.

Similarly, to show (b), consider $x \in \widehat{L}_{h_n,t}$ so that $\widehat{p}_{h_n}(x) > t$. Thus we have

$$p_{h_n}(x) \ge \widehat{p}_{h_n}(x) - \|\widehat{p}_{h_n} - p_{h_n}\|_\infty > t - r_n,$$

which leads to $x \in L_{h_n,t-r_n}$. Then again by the stability property of $L_{h_n,t}$, there exists $y \in L_{h_n,t}$ such that $\|x-y\|_2 \le Cr_n$.

Hence by definition, we upper bound the Hausdorff distance $H(\widehat{L}_t, L_{h,t})$ by

$$\begin{aligned} &Cr_n \\ &= C\mathsf{C}_{P,K,B}\left(\sqrt{\frac{(\log(1/h_n))_+ + \log(2/\delta)}{nh_n^d}} + \frac{1}{h_n^{d+1}}\min\left\{\sum_a \|\widehat{\mu}_a - \mu_a\|_1 + \sqrt{\frac{\log(2/\delta)}{n}}, h_n\right\}\right). \end{aligned}$$

□

C.3 Proofs for Section 4.4

C.3.1 Proof of Theorem 4.4.1 and the 2nd order remainder

The following lemma computes the efficient influence function (EIF) of ψ_C when our covariate space $\mathscr{X}$ is discrete. For the sake of simplicity, we consider the binary treatment case which is enough for our proof.

Lemma C.3.1 (Efficient influence function)**.** *Suppose that $\mathscr{X}$ is discrete. For ψ_C, the uncentered efficient influence function φ_C under a nonparametric model is as given by*

$$\varphi_C(Z) = \sum_{a\in\mathscr{A}} \left\{ 2\left[\sum_r f_r^a(\boldsymbol{\mu})\right] \sum_{a'\in\mathscr{A}} \left\{ \sum_r \left[\frac{\partial f_r^a}{\partial \mu_{a'}} \frac{\mathbb{1}(A=a')}{\pi_{a'}}(Y-\mu_{a'}) \right] \right\} + \left[\sum_r f_r^a(\boldsymbol{\mu})\right]^2 \right\}$$

where for $a,a' \in \mathscr{A}$

$$f_r^a(\boldsymbol{\mu};C,h) = \omega_r(\mu_a - c_{ra}),$$

$$\frac{\partial \omega_r}{\partial \mu_{a'}} = -\frac{\omega_r}{h}\left\{ \frac{\mu_{a'}-c_{ra'}}{\|\boldsymbol{\mu}-c_r\|_2} - \sum_j \frac{\mu_{a'}-c_{ja'}}{\|\boldsymbol{\mu}-c_j\|_2}\omega_j \right\}$$

, $c_r = [c_{r1},...,c_{rp}]^\top$. The weight term ω_r is defined in (4.7) *based on the Gaussian kernel.*

Proof. It suffices to prove for $p=2$ (binary treatments). By definition,

$$\begin{aligned}\psi_C &= \mathbb{E}\|\boldsymbol{\mu} - \widetilde{\Pi}_C(\boldsymbol{\mu};h)\|_2^2 \\ &= \mathbb{E}\left[\sum_{a\in\mathscr{A}}\left(\sum_r \omega_r(\mu_a - c_{ra})\right)^2\right].\end{aligned}$$

By letting $\psi_C^a \equiv \mathbb{E}\left[(\sum_r \omega_r(\mu_a - c_{ra}))^2\right]$, we can write $\psi_C = \sum_{a\in\mathscr{A}} \psi_C^a$.

Now define a function $f_r^a : \mathbb{R}^2 \to \mathbb{R}$ by $f_r^a(\boldsymbol{\mu}) = \omega_r(\mu_a - c_{ra})$ for $\forall a \in \mathscr{A}, k \in \{1,...,k\}$. Note thta since we use smooth Gaussian kernel, f_r^a is also smooth, differentiable in arbitrary order. Then we have $\psi_C^a = \mathbb{E}\left[(\sum_r f_r^a(\boldsymbol{\mu}))^2\right]$.

Let ϕ_C^a denote the EIF of ψ_C^a. In order to find an EIF of ψ_C^a we use derivative rule. First we suppose X is discrete. Then we have

$$\psi_C^a = \sum_{x\in\mathscr{X}}\left[\sum_r f_r^a(\boldsymbol{\mu}(x))\right]^2 p(x)$$

where $p(x) = \mathbb{P}(X=x)$. From this, it follows

$$\begin{aligned}\phi_C^a &= \sum_{x\in\mathscr{X}}\left\{ IF\left(\left[\sum_r f_r^a(\boldsymbol{\mu}(x))\right]^2\right)p(x) + \left[\sum_r f_r^a(\boldsymbol{\mu}(x))\right]^2 IF(p(x))\right\} \\ &= \sum_{x\in\mathscr{X}}\left\{2\left[\sum_r f_r^a(\boldsymbol{\mu}(x))\right]\sum_r\left[\frac{\partial f_r^a}{\partial \mu_0}(x)IF(\mu_0(x)) + \frac{\partial f_r^a}{\partial \mu_1}(x)IF(\mu_1(x))\right]\right\}p(x) \\ &\quad + \left[\sum_r f_r^a(\boldsymbol{\mu}(x))\right]^2 IF(p(x)).\end{aligned}$$

However we have

$$
\begin{aligned}
IF(\mu_0(x)) &= \frac{1-A}{1-\pi(x)}\frac{\mathbb{1}(X=x)}{p(x)}[Y-\mu_0(x)] \\
IF(\mu_1(x)) &= \frac{A}{\pi(x)}\frac{\mathbb{1}(X=x)}{p(x)}[Y-\mu_1(x)] \\
IF(p(x)) &= \mathbb{1}(X=x)-p(x).
\end{aligned}
$$

Plugging this into the last display yields

$$
\begin{aligned}
\phi_C^a &= 2\left[\sum_r f_r^a(\boldsymbol{\mu})\right]\sum_r\left[\frac{\partial f_r^a}{\partial\mu_0}\frac{1-A}{1-\pi}[Y-\mu_0]+\frac{\partial f_r^a}{\partial\mu_1}\frac{A}{\pi}[Y-\mu_1]\right] \\
&\quad+\left[\sum_r f_r^a(\boldsymbol{\mu})\right]^2-\sum_{x\in\mathscr{X}}\left[\sum_r f_r^a(\boldsymbol{\mu}(x))\right]^2 p(x) \\
&= 2\left[\sum_r f_r^a(\boldsymbol{\mu})\right]\sum_r\left[\frac{\partial f_r^a}{\partial\mu_0}\frac{1-A}{1-\pi}[Y-\mu_0]+\frac{\partial f_r^a}{\partial\mu_1}\frac{A}{\pi}[Y-\mu_1]\right] \\
&\quad+\left[\sum_r f_r^a(\boldsymbol{\mu})\right]^2-\psi_C^a. \qquad (C.10)
\end{aligned}
$$

Finally we obtain $\phi_C=\sum_{a\in\mathscr{A}}\phi_C^a$. Note that ϕ_C^a relies on a set of nuisance parameters $\eta=(\pi,\mu_0,\mu_1)$ for $\forall a\in\mathscr{A}$. By induction, the result for any p follows immediately.

□

In order to formally verify that (C.10) in Lemma C.3.1 is actually the EIF of ψ_C, we study the remainder term in the von-Mises expansion (1.4) and show that it is indeed a second-order term. Due to linearity it suffices to consider ϕ_C^a for $\forall a\in\mathscr{A}$. Let $\mathbb{P}$, $\overline{\mathbb{P}}$ be two arbitrary distributions. We use overbars to denote parameters or nuisance functions corresponding to $\overline{\mathbb{P}}$. Then we have the von Mises expansion

$$
\begin{aligned}
\psi_C^a(\overline{\mathbb{P}})-\psi_C^a(\mathbb{P}) &= -\int\phi_C^a(\overline{\mathbb{P}})d\mathbb{P}+R_2^a(\overline{\mathbb{P}},\mathbb{P}) \\
&= -\mathbb{E}_{Z\sim\mathbb{P}}[\overline{\phi}_C^a]+R_2^a(\overline{\mathbb{P}},\mathbb{P}).
\end{aligned} \qquad (C.11)
$$

The next lemma provides an explicit formula for $R_2^a(\overline{\mathbb{P}},\mathbb{P})$ in (C.11). For the notational brevity, we shall stick to the case $p=2$, as the extension to arbitrary p is straightforward.

Lemma C.3.2. *In (C.11), it follows that*

$$
\begin{aligned}
R_2^a(\overline{\mathbb{P}},\mathbb{P}) = {} & 2\mathbb{E}_{Z\sim\mathbb{P}}\left\{\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]\sum_r\left[\frac{\partial f_r^a}{\partial\mu_0}\frac{\overline{\pi}-\pi}{1-\overline{\pi}}[\mu_0-\overline{\mu_0}]+\frac{\partial f_r^a}{\partial\mu_1}\frac{\pi-\overline{\pi}}{\overline{\pi}}[\mu_1-\overline{\mu_1}]\right]\right\} \\
& -\mathbb{E}_{Z\sim\mathbb{P}}\left\{[\mu_0-\overline{\mu_0},\mu_1-\overline{\mu_1}]\boldsymbol{H}_g^*\begin{bmatrix}\mu_0-\overline{\mu_0}\\ \mu_1-\overline{\mu_1}\end{bmatrix}\right\}.
\end{aligned}
\tag{C.12}
$$

Proof. We compute

$$
\begin{aligned}
& \mathbb{E}_{Z\sim\mathbb{P}}[\overline{\phi}_C^a]+\psi_C^a(\overline{\mathbb{P}})-\psi_C^a(\mathbb{P}) \\
& = 2\mathbb{E}_{Z\sim\mathbb{P}}\left\{\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]\sum_r\left[\frac{\partial f_r^a}{\partial\mu_0}\frac{1-A}{1-\overline{\pi}}[Y-\overline{\mu_0}]+\frac{\partial f_r^a}{\partial\mu_1}\frac{A}{\overline{\pi}}[Y-\overline{\mu_1}]\right]\right\} \\
& \quad +\mathbb{E}_{Z\sim\mathbb{P}}\left\{\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]^2\right\}-\psi_C^a(\overline{\mathbb{P}})+\psi_C^a(\overline{\mathbb{P}})-\psi_C^a(\mathbb{P}) \\
& = 2\mathbb{E}_{Z\sim\mathbb{P}}\left\{\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]\sum_r\left[\frac{\partial f_r^a}{\partial\mu_0}\frac{1-\pi}{1-\overline{\pi}}[\mu_0-\overline{\mu_0}]+\frac{\partial f_r^a}{\partial\mu_1}\frac{\pi}{\overline{\pi}}[\mu_1-\overline{\mu_1}]\right]\right\} \\
& \quad +\mathbb{E}_{Z\sim\mathbb{P}}\left\{\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]^2\right\}-\psi_C^a(\mathbb{P}) \\
& = 2\mathbb{E}_{Z\sim\mathbb{P}}\left\{\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]\sum_r\left[\frac{\partial f_r^a}{\partial\mu_0}\frac{\overline{\pi}-\pi}{1-\overline{\pi}}[\mu_0-\overline{\mu_0}]+\frac{\partial f_r^a}{\partial\mu_1}\frac{\pi-\overline{\pi}}{\overline{\pi}}[\mu_1-\overline{\mu_1}]\right]\right\} \\
& \quad +\mathbb{E}_{Z\sim\mathbb{P}}\left\{2\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]\sum_r\left[\frac{\partial f_r^a}{\partial\mu_0}[\mu_0-\overline{\mu_0}]+\frac{\partial f_r^a}{\partial\mu_1}[\mu_1-\overline{\mu_1}]\right]\right\} \\
& \quad +\mathbb{E}_{Z\sim\mathbb{P}}\left\{\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]^2-\left[\sum_r f_r^a(\boldsymbol{\mu})\right]^2\right\}
\end{aligned}
$$

, where the second equality follows by the law of iterated expectations, the third by adding and subtracting the second term in the display.

Now let $g(\boldsymbol{\mu})=[\sum_r f_r^a(\boldsymbol{\mu})]^2$. Taylor's Theorem gives

$$
\begin{aligned}
& \left[\sum_r f_r^a(\boldsymbol{\mu})\right]^2-\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]^2-2\left[\sum_r f_r^a(\overline{\boldsymbol{\mu}})\right]\sum_r\left[\frac{\partial f_r^a}{\partial\mu_0}[\mu_0-\overline{\mu_0}]+\frac{\partial f_r^a}{\partial\mu_1}[\mu_1-\overline{\mu_1}]\right] \\
& = [\mu_0-\overline{\mu_0},\mu_1-\overline{\mu_1}]\boldsymbol{H}_g^*\begin{bmatrix}\mu_0-\overline{\mu_0}\\ \mu_1-\overline{\mu_1}\end{bmatrix}
\end{aligned}
$$

where $(\boldsymbol{H}_g^*)_{ij} = \frac{\partial^2 g}{\partial \mu_i \partial \mu_j}(\boldsymbol{\mu}^*)$ for some $\boldsymbol{\mu}^*$ on the line segment joining $\overline{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}$ for each $i, j \in \mathscr{A}$. Applying this to our last display in the above equation yields

$$\begin{aligned} R_2^a(\overline{\mathbb{P}}, \mathbb{P}) = 2\mathbb{E}_{Z\sim\mathbb{P}} & \left\{ \left[\sum_r f_r^a(\overline{\boldsymbol{\mu}}) \right] \sum_r \left[\frac{\partial f_r^a}{\partial \mu_0} \frac{\overline{\pi} - \pi}{1 - \overline{\pi}} [\mu_0 - \overline{\mu_0}] + \frac{\partial f_r^a}{\partial \mu_1} \frac{\pi - \overline{\pi}}{\overline{\pi}} [\mu_1 - \overline{\mu_1}] \right] \right\} \\ & - \mathbb{E}_{Z\sim\mathbb{P}} \left\{ [\mu_0 - \overline{\mu_0}, \mu_1 - \overline{\mu_1}] \boldsymbol{H}_g^* \begin{bmatrix} \mu_0 - \overline{\mu_0} \\ \mu_1 - \overline{\mu_1} \end{bmatrix} \right\} \end{aligned}$$

Since this remainder term depends only on the second-order products of differences between $\mathbb{P}$ and $\overline{\mathbb{P}}$, it is clear to see that the pathwise differentiability (1.4) holds.

□

C.3.2 Proof of Lemma 4.4.1

We need the following two auxiliary lemmas; for an optimal cluster codebook $C \in \mathscr{M}^*$, Lemma C.3.3 bounds an error from kernel-smoothing the original k-means risk function $R(C)$, and Lemma C.3.4 shows an asymptotic behavior of our estimator $\widehat{\psi}_C$.

Lemma C.3.3. *Under the margin condition, for an optimal codebook $C \in \mathscr{M}^*$ we have*

$$R_h(C) - R(C) = O\left(kh^{\alpha}\right).$$

Proof. For simplicity, we write $\widetilde{\Pi}_C \equiv \widetilde{\Pi}_C(\boldsymbol{\mu}; h)$, $\Pi_C \equiv \Pi_C(\boldsymbol{\mu}; h)$, $\omega_r \equiv \omega_r(\boldsymbol{\mu})$ in this proof. Now we have

$$\begin{aligned} R_h(C) - R(C) &= \mathbb{E}\|\boldsymbol{\mu} - \widetilde{\Pi}_C\|_2^2 - \mathbb{E}\|\boldsymbol{\mu} - \Pi_C\|_2^2 \\ &= \mathbb{E}\left\{ 2\left\langle \Pi_C - \widetilde{\Pi}_C, \boldsymbol{\mu} - \frac{\Pi_C + \widetilde{\Pi}_C}{2} \right\rangle \right\} \\ &\leq 4B\mathbb{E}\|\Pi_C - \widetilde{\Pi}_C\|_2 \end{aligned}$$

, where the last inequality follows by the Cauchy-Schwarz inequality and the boundedness assumption (A4).

By abuse of notation we let $k^* = \underset{j \in \{1,\ldots,k\}}{\operatorname{argmin}} \|\boldsymbol{\mu} - c_j\|_2$ and $k^{**} = \underset{j \in \{1,\ldots,k\}, j \notin k^*}{\operatorname{argmin}} \|\boldsymbol{\mu} - c_j\|_2$. Similarly, we let c_*, c_{**} denote cluster centers corresponding to k^* or k^{**} respectively (i.e.

$c^* = \Pi_C$). Finally let us write $\boldsymbol{K}^* = \boldsymbol{K}(\boldsymbol{\mu}, c_*)$, $\boldsymbol{K}^{**} = \boldsymbol{K}(\boldsymbol{\mu}, c_{**})$. Then we obtain

$$\begin{aligned}
\mathbb{E}\|\Pi_C - \widetilde{\Pi}_C\|_2 &= \mathbb{E}\left\|\sum_r c_r \left(\mathbb{1}\left\{r = k^*\right\} - \omega_r\right)\right\|_2 \\
&\leq \mathbb{E}\left\{\sum_{r \neq k^*} \|c_r\|_2 \omega_r + \|c_{k^*}\|_2 \left(1 - \omega_{k^*}\right)\right\} \\
&\lesssim B\mathbb{E}\left\{\sum_{r \neq k^*} \omega_r + \left(1 - \omega_{k^*}\right)\right\} \\
&\leq B\mathbb{E}\left\{2(k-1)\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^*}\right\}.
\end{aligned}$$

Next we note that

$$\mathbb{E}\left[\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^*}\right] \leq \mathbb{E}\left[\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^*}\middle|\boldsymbol{\mu} \notin N_C(\kappa)\right] + \mathbb{E}\left[\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^*}\middle|\boldsymbol{\mu} \in N_C(\kappa)\right].$$

The condition $\boldsymbol{\mu} \notin N_C(\kappa)$ implies $\kappa < \|\boldsymbol{\mu} - c_{**}\|_2 - \|\boldsymbol{\mu} - c_*\|_2$. Thus the first term is easily bounded by

$$\mathbb{E}\left[\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^*}\middle|\boldsymbol{\mu} \notin N_C(\kappa)\right] \lesssim \exp\left(-\frac{\kappa}{h}\right).$$

For the second term, we let $\xi_C(\boldsymbol{\mu}) := \|\boldsymbol{\mu} - c_{**}\|_2 - \|\boldsymbol{\mu} - c_*\|_2$. Then under the condition $\boldsymbol{\mu} \in N_C(\kappa)$, our margin condition implies that $\mathbb{P}\left(\xi_C(\boldsymbol{\mu}) \leq t\right) \leq C' \min\{t^\alpha, 1\}$ for some $C' \geq 1$ and all $0 \leq t \leq \kappa$. Then it follows

$$\begin{aligned}
\mathbb{E}\left[\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^*}\middle|\boldsymbol{\mu} \in N_C(\kappa)\right] &= \int\limits_{\boldsymbol{\mu} \in N_C(\kappa)} \exp\left(-\frac{\xi_C(\boldsymbol{\mu})}{h}\right) d\mathbb{P}(\boldsymbol{\mu}) \\
&= \int_0^\infty \mathbb{P}\left(\exp\left(-\frac{\xi_C(\boldsymbol{\mu})}{h}\right) \geq t\right) dt \\
&= \int_0^1 \mathbb{P}\left(\xi_C(\boldsymbol{\mu}) \leq -h \log t\right) dt \\
&\lesssim \int_0^1 \min\{(-h \log t)^\alpha, 1\} dt \\
&= \left\{\int_0^{\exp(-1/h)} dt + h^\alpha \int_{\exp(-1/h)}^1 \left(\log\frac{1}{t}\right)^\alpha dt\right\} \\
&\leq \left\{\exp\left(-\frac{1}{h}\right) + h^\alpha \left[\log\frac{1}{c'}\right]^\alpha\right\}
\end{aligned}$$

for some constant $c' \in (\exp(-1/h), 1)$, where we used the mean value theorem and the fact that $\exp\left(-\frac{1}{h}\right) > 0$ to obtain the last inequality. Therefore we conclude that

$$\mathbb{E}\left[\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^{*}}\middle|\boldsymbol{\mu} \in N_C(\kappa)\right] = O(h^{\alpha}).$$

Collecting two separate pieces, we have that

$$\mathbb{E}\left[\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^{*}}\right] = O\left(\exp\left(-\frac{\kappa}{h}\right) + h^{\alpha}\right) = O(h^{\alpha}).$$

Hence we finally conclude that

$$R_h(C) - R(C) = O(kh^{\alpha}).$$

□

Lemma C.3.4. *For $C \in \mathscr{M}^*$,*

$$\sqrt{n}\left(\widehat{\psi}_C - R(C)\right) = a_n + b_n + c_n$$

where $a_n = O_{\mathbb{P}}\left(\sum_{a\in\mathscr{A}} \|\widehat{\varphi}_C^a - \varphi_C^a\|\right)$, $b_n \rightsquigarrow N\left(0, var\left(\sum_{a\in\mathscr{A}} \bar{\phi}_{C^*}^a(Z)\right)\right)$, and

$$\begin{aligned} c_n \lesssim & \left(\frac{kh^{\frac{\alpha}{2}-1}}{2^{\frac{\alpha}{2}}} + 1\right) \sum_{a'\in\mathscr{A}} \|\pi_{a'} - \overline{\pi_{a'}}\|_{\mathbb{P},4} \|\mu_{a'} - \overline{\mu_{a'}}\|_{\mathbb{P},4} \\ & + \left(\frac{k^2}{4^{\frac{\alpha}{2}}} h^{\frac{\alpha}{2}-2} + \frac{k}{2^{\frac{\alpha}{2}}} h^{\frac{\alpha}{2}-1} + 1\right) \sum_{a',a''\in\mathscr{A}} \|\mu_{a'} - \overline{\mu_{a'}}\|_{\mathbb{P},4} \|\mu_{a''} - \overline{\mu_{a''}}\|_{\mathbb{P},4} \\ & + kh^{\alpha}. \end{aligned}$$

$\bar{\phi}_{C^*}^a(Z)$ is defined in (C.19) in the proof.

Proof. Here we show our proposed estimator $\widehat{\psi}_C$ is consistent and asymptotically normal estimator for $R(C)$. First, since $\psi_C = R_h(C)$ we have

$$\widehat{\psi}_C - R(C) = \widehat{\psi}_C - \psi_C + R_h(C) - R(C).$$

Recall the uncentered efficient influence function $\varphi_C^a = \phi_C^a + \psi_C^a$ for $a \in \mathscr{A}$ and let $\mathbb{G}_n^s$ denote the empirical process over group s by $\mathbb{G}_n^s = \sqrt{n}(\mathbb{P}_n^s - \mathbb{P})$. Then we have the following decomposition

$$\sqrt{n}\left(\widehat{\psi}_C - R(C)\right) = \underbrace{\frac{1}{S}\sum_{s=1}^{S}\sum_{a\in\mathscr{A}}\left[\mathbb{G}_n^s\left\{\varphi_C^a(\hat{\eta}_{-s}) - \varphi_C^a(\eta)\right\}\right]}_{\text{i}}$$
$$+ \underbrace{\sqrt{n}\left\{\frac{1}{S}\sum_{s=1}^{S}\sum_{a\in\mathscr{A}}\left[\mathbb{P}\left\{\varphi_C^a(\hat{\eta}_{-s}) - \varphi_C^a(\eta)\right\}\right] + R_h(C) - R(C)\right\}}_{\text{ii}} \quad \text{(C.13)}$$
$$+ \underbrace{\sum_{a\in\mathscr{A}}\mathbb{G}_n\left\{\varphi_C^a(\eta)\right\}}_{\text{iii}}$$

which follows by noting that $\psi_C^a = \mathbb{P}\left(\varphi_C^a\right)$ and $\sum_s \mathbb{P}_n^s\left(\varphi_C^a\right) = \sum_s \mathbb{P}_n\left(\varphi_C^a\right)$ and simple rearranging. In what follows, we analyze each term in the right-hand side of above display.

part i) Let us write $\overline{\varphi}_C^a = \varphi_C^a(\hat{\eta}_{-s})$ and $\overline{\phi}_C^a = \phi_C^a(\hat{\eta}_{-s})$. By the sample splitting lemma [72, Lemma 2] it immediately follows

$$\mathbb{G}_n^s\left(\overline{\varphi}_C^a - \varphi_C^a\right) = O_{\mathbb{P}}\left(\|\overline{\varphi}_C^a - \varphi_C^a\|\right)$$

and thereby the entire term is of order $O_{\mathbb{P}}\left(\sum_{a\in\mathscr{A}}\|\overline{\varphi}_C^a - \varphi_C^a\|\right)$.

part ii) From (C.11) and by $\mathbb{P}(\psi_C) = 0$, we first notice that for any a

$$\begin{aligned}\mathbb{P}\left\{\overline{\varphi}_C^a - \varphi_C^a\right\} &= \int \overline{\phi}_C^a d\mathbb{P} + \overline{\psi}_C^a - \psi_C^a \\ &= R_2^a(\overline{\mathbb{P}}, \mathbb{P})\end{aligned} \quad \text{(C.14)}$$

where $\overline{\mathbb{P}}$ is the probability distribution for units in all but group s that are used to estimate η. Hence we have that $\frac{1}{S}\sum_s\sum_{a\in\mathscr{A}}\left[\mathbb{P}\left\{\varphi_C^a(\hat{\eta}_{-s}) - \varphi_C^a(\eta)\right\}\right] \lesssim \sum_{a\in\mathscr{A}} R_2^a(\widehat{\mathbb{P}}, \mathbb{P})$. Now let us define

$$R_n := \frac{1}{S}\sum_s\sum_{a\in\mathscr{A}}\left[\mathbb{P}\left\{\varphi_C^a(\hat{\eta}_{-s}) - \varphi_C^a(\eta)\right\}\right] + R_h(C) - R(C),$$

which consists of the second-order remainders which we analyzed in (C.12) and the approximation error analyzed in Lemma C.3.3.

On the other hand, for $\forall a \neq a' \in \mathscr{A}$,

$$\sum_r \frac{\partial f_r^a}{\partial \mu_{a'}} = \sum_r\left\{(\mu_a - c_{ra})\frac{\partial \omega_r}{\partial \mu_{a'}}\right\}, \quad \sum_r \frac{\partial f_r^a}{\partial \mu_a} = \sum_r\left\{(\mu_a - c_{ra})\frac{\partial \omega_r}{\partial \mu_a}\right\} + \sum_r \omega_r.$$

As in Lemma C.3.3, we again let k^* denote an index corresponding to the projection $\Pi_C(\boldsymbol{\mu})$. Then it follows that

$$\begin{aligned}
\left|\sum_r \left\{(\mu_a - c_{ra})\frac{\partial \omega_r}{\partial \mu_{a'}}\right\}\right| &= \left|\sum_r \left\{(\mu_a - c_{ra})\frac{\omega_r}{h}\left(-\frac{\mu_{a'} - c_{ra'}}{\|\boldsymbol{\mu} - c_r\|_2} + \sum_j \frac{\mu_{a'} - c_{ja'}}{\|\boldsymbol{\mu} - c_j\|_2}\omega_j\right)\right\}\right| \\
&\leq \frac{1}{h}\left|(\mu_a - c_{k^*a})\omega_{k^*}\left\{\sum_j \frac{\mu_{a'} - c_{ja'}}{\|\boldsymbol{\mu} - c_j\|_2}\left(\omega_j - \mathbb{1}\{j = k^*\}\right)\right\}\right| \\
&\quad + \frac{1}{h}\left|\sum_{r \neq k^*}(\mu_a - c_{ra})\left(\omega_r - \mathbb{1}\{r = k^*\}\right)\left(-\frac{\mu_{a'} - c_{ra'}}{\|\boldsymbol{\mu} - c_r\|_2} + \sum_j \frac{\mu_{a'} - c_{ja'}}{\|\boldsymbol{\mu} - c_j\|_2}\omega_j\right)\right| \\
&\leq \frac{2B}{h}\left|\sum_j \frac{\mu_{a'} - c_{ja'}}{\|\boldsymbol{\mu} - c_j\|_2}\left(\omega_j - \mathbb{1}\{j = k^*\}\right)\right| + \frac{2}{h}\left|\sum_{r \neq k^*}(\mu_a - c_{ra})\left(\omega_r - \mathbb{1}\{r = k^*\}\right)\right| \\
&\lesssim \frac{1}{h}\left|\sum_r \Upsilon_r\left(\omega_r - \mathbb{1}\{r = k^*\}\right)\right|,
\end{aligned}$$

where $\Upsilon_r \in \mathbb{R}$ for all $k \in \mathbb{N}$ such that $|\Upsilon_r| \leq B'$ for $0 < B' < 4B$. With this result, by the similar logic used in Lemma C.3.3, we obtain that

$$\begin{aligned}
\left\|\sum_r \left\{(\mu_a - c_{ra})\frac{\partial \omega_r}{\partial \mu_{a'}}\right\}\right\| &= \left[\mathbb{E}\left\{\frac{1}{h^2}\left|\sum_r \Upsilon_r\left(\omega_r - \mathbb{1}\{r = k^*\}\right)\right|\left|\sum_r \Upsilon_r\left(\omega_r - \mathbb{1}\{r = k^*\}\right)\right|\right\}\right]^{1/2} \\
&\lesssim \frac{1}{h}\left[\mathbb{E}\left\{\left(k\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^*}\right)^2\right\}\right]^{1/2} \\
&\lesssim \frac{k}{2^{\frac{\alpha}{2}}}h^{\frac{\alpha}{2}-1}. \qquad \text{(C.15)}
\end{aligned}$$

Therefore for any $a, a' \in \mathscr{A}$,

$$\left\|\sum_r \frac{\partial f_r^a}{\partial \mu_{a'}}\right\| \lesssim \frac{k}{2^{\frac{\alpha}{2}}}h^{\frac{\alpha}{2}-1} + 1. \qquad \text{(C.16)}$$

On the other hand, given $g(\boldsymbol{\mu}) = [\sum_r f_r^a(\boldsymbol{\mu})]^2$, for $\forall a', a''$ we have

$$\frac{\partial^2 g}{\partial \mu_{a'}\partial \mu_{a''}} = 2\left(\sum_r \frac{\partial f_r^a}{\partial \mu_{a'}}\right)\left(\sum_r \frac{\partial f_r^a}{\partial \mu_{a''}}\right) + 2\left(\sum_r f_r^a\right)\left(\sum_r \frac{\partial^2 f_r^a}{\partial \mu_{a'}\partial \mu_{a''}}\right)$$

where $\sum_r \frac{\partial^2 f_r^a}{\partial \mu_{a'} \partial \mu_{a''}} = \sum_r \frac{\partial^2 \omega_r}{\partial \mu_{a'} \partial \mu_{a''}} + \sum_r \frac{\partial \omega_r}{\partial \mu_{a'}} + \sum_r \frac{\partial \omega_r}{\partial \mu_{a''}}$. Through the similar algebra to obtain (C.16) one can show that for $\forall a', a'' \in \mathscr{A}$,

$$\left\| \frac{\partial^2 g}{\partial \mu_{a'} \partial \mu_{a''}} \right\| \lesssim \frac{k^2}{4^{\frac{\alpha}{2}}} h^{\frac{\alpha}{2}-2} + \frac{k}{2^{\frac{\alpha}{2}}} h^{\frac{\alpha}{2}-1} + 1. \tag{C.17}$$

Now from (C.12), using (C.16), (C.17) and the fact that $\sum_r f_r^a(\boldsymbol{\mu})$ and all the other quantities are bounded, by the Cauchy-Schwarz and the triangle inequality we have

$$\begin{aligned} R_2^a(\overline{\mathbb{P}}, \mathbb{P}) &\lesssim \sum_{a' \in \mathscr{A}} \left\{ \left\| \sum_r \frac{\partial f_r^a}{\partial \mu_{a'}} \right\| \|(\pi_{a'} - \overline{\pi_{a'}})(\mu_{a'} - \overline{\mu_{a'}})\| \right\} + \sum_{a', a'' \in \mathscr{A}} \left\| \frac{\partial^2 g}{\partial \mu_{a'} \partial \mu_{a''}} \right\| \|(\mu_{a'} - \overline{\mu_{a'}})(\mu_{a''} - \overline{\mu_{a''}})\| \\ &\lesssim \left(\frac{k h^{\frac{\alpha}{2}-1}}{2^{\frac{\alpha}{2}}} + 1 \right) \sum_{a' \in \mathscr{A}} \|(\pi_{a'} - \overline{\pi_{a'}})(\mu_{a'} - \overline{\mu_{a'}})\| \\ &\quad + \left(\frac{k^2}{4^{\frac{\alpha}{2}}} h^{\frac{\alpha}{2}-2} + \frac{k}{2^{\frac{\alpha}{2}}} h^{\frac{\alpha}{2}-1} + 1 \right) \sum_{a', a'' \in \mathscr{A}} \|(\mu_{a'} - \overline{\mu_{a'}})(\mu_{a''} - \overline{\mu_{a''}})\| \\ &\lesssim \left(\frac{k h^{\frac{\alpha}{2}-1}}{2^{\frac{\alpha}{2}}} + 1 \right) \sum_{a' \in \mathscr{A}} \|\pi_{a'} - \overline{\pi_{a'}}\|_{\mathbb{P},4} \|\mu_{a'} - \overline{\mu_{a'}}\|_{\mathbb{P},4} \\ &\quad + \left(\frac{k^2}{4^{\frac{\alpha}{2}}} h^{\frac{\alpha}{2}-2} + \frac{k}{2^{\frac{\alpha}{2}}} h^{\frac{\alpha}{2}-1} + 1 \right) \sum_{a', a'' \in \mathscr{A}} \|\mu_{a'} - \overline{\mu_{a'}}\|_{\mathbb{P},4} \|\mu_{a''} - \overline{\mu_{a''}}\|_{\mathbb{P},4}. \end{aligned} \tag{C.18}$$

(C.18) together with the result of Lemma C.3.3 give the upper bound for R_n, thereby for *part ii*.

part iii) We fix a and define

$$\phi_{C,1}^a = 2 \left[\sum_j f_j^a(\boldsymbol{\mu}) \right] \sum_r \left\{ (\mu_a - c_{ra}) \sum_{a' \in \mathscr{A}} \frac{\partial \omega_r}{\partial \mu_{a'}} \frac{\mathbb{1}(A = a')}{\pi_{a'}} (Y - \mu_{a'}) \right\},$$

$$\phi_{C,2}^a = 2 \left[\sum_j f_j^a(\boldsymbol{\mu}) \right] m(A, X, Y) + \left[\sum_j f_j^a(\boldsymbol{\mu}) \right]^2 - \psi_C^a$$

where

$$m(A, X, Y) = \sum_{a' \in \mathscr{A}} \frac{\mathbb{1}(A = a')}{\pi_{a'}} (Y - \mu_{a'}).$$

Then the efficient influence function ϕ_C^a in (C.10) (for any $p \geq 2$) can be written by

$$\phi_C^a = \phi_{C,1}^a + \phi_{C,2}^a.$$

First for $\phi^a_{C,1}$, we note that $\phi^a_{C,1} \lesssim \sum_{a' \in \mathscr{A}} \sum_r \left\{ (\mu_a - c_{ra}) \frac{\partial \omega_r}{\partial \mu_{a'}} \right\}$ as all the other terms are bounded. Then, through the similar procedure to derive (C.16) we note that for any $a, a' \in \mathscr{A}$

$$\begin{aligned}\sum_r \left\{ (\mu_a - c_{ra}) \frac{\partial \omega_r}{\partial \mu_{a'}} \right\} &= \frac{1}{h} \sum_r \left\{ (\mu_a - c_{ra}) \omega_r \left(- \frac{\mu_{a'} - c_{ra'}}{\|\boldsymbol{\mu} - c_r\|_2} + \sum_j \frac{\mu_{a'} - c_{ja'}}{\|\boldsymbol{\mu} - c_j\|_2} \omega_j \right) \right\} \\ &\leq \frac{2}{h} \sum_r \Upsilon_r \left(\omega_r - \mathbb{1}\{r = k^*\} \right)\end{aligned}$$

for some bounded $\Upsilon_r \in \mathbb{R}$. Now for any $\gamma \in (0,1)$ consider $N_C(h^\gamma)$. Note that $N_C(h^\gamma)$ is shrinking toward ∂C as n grows. Then based on the same logic used in Lemma C.3.3, it follows that

$$\begin{aligned}\sum_r \Upsilon_r \left(\omega_r - \mathbb{1}\{r = k^*\} \right) &\lesssim k \frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^{*}} \\ &\leq \begin{cases} 1, & \text{if } \boldsymbol{\mu} \in N_C(h^\gamma) \\ \exp(-h^{\gamma - 1}), & \text{otherwise.} \end{cases}\end{aligned}$$

Since we only consider $C \in \mathscr{M}^*$, by the given margin condition (a) for all n,

$$\begin{aligned}\sum_j \mathbb{P}\left(\boldsymbol{\mu}_j \in N_C(h^\gamma) \right) &\leq \sum_j \mathbb{P}\left(\boldsymbol{\mu}_j \in N_C(h^\gamma) \mid \kappa \leq h^\gamma \right) + \sum_j \mathbb{P}\left(\boldsymbol{\mu}_j \in N_C(h^\gamma) \mid \kappa > h^\gamma \right) \\ &\leq M_0 + n h^{\alpha\gamma} \\ &< \infty,\end{aligned}$$

for some finite constant $M_0 > 0$, where the last inequality follows by Assumption (c). Hence by the Borel-Cantelli lemma, almost surely $\boldsymbol{\mu}_j \notin N_C(h^\gamma)$ for all but finitely many n.

Consequently we have

$$\begin{aligned}\sqrt{n}\mathbb{P}_n \left\{ \phi^a_{C,1} \right\} &= \frac{1}{\sqrt{n}} \left\{ \sum_{i : \boldsymbol{\mu}_i \in N_C(n^{-\gamma})} \phi^a_{C,1}(\boldsymbol{\mu}_i) + \sum_{i : \boldsymbol{\mu}_i \notin N_C(h^\gamma)} \phi^a_{C,1}(\boldsymbol{\mu}_i) \right\} \\ &\lesssim \frac{1}{h\sqrt{n}} \text{card}\left(\{ i : \boldsymbol{\mu}_i \in N_C(h^\gamma) \} \right) + \frac{1}{\sqrt{n}} \sum_{i : \boldsymbol{\mu}_i \notin N_C(h^\gamma)} \frac{1}{h} \exp(-h^{\gamma - 1}) \\ &= o(1) \quad \text{a.s.}\end{aligned}$$

where the last equality follows by the fact that $\text{card}\left(\{ i : \boldsymbol{\mu}_i \in N_C(h^\gamma) \} \right) < \infty$ for all n, and that $\frac{1}{h\sqrt{n}} = o(1)$ under Assumption (c).

On the other hand, if we let $\zeta^a_{a'}(X) = 2\left[\sum_j f^a_j(\boldsymbol{\mu})\right]\left[\sum_r (\mu_a - c_{ra})\frac{\partial \omega_r}{\partial \mu_{a'}}\right]$ then by the law of total expectation,

$$\begin{aligned}\mathbb{E}\left[\phi^a_{C,1}\right] &= \mathbb{E}\left[\sum_{a'} \zeta^a_{a'}(X)\mathbb{E}\left\{m(A,X,Y) \mid X, A = a'\right\}\mathbb{P}(a' \mid X)\right] \\ &= 0\end{aligned}$$

where we used $\mathbb{E}\{m_{a'}(A,X,Y) \mid X, A = a'\} = 0$ for $\forall a'$. Hence we conclude that $\mathbb{G}_n\left\{\phi^a_{C,1}\right\} = o(1)$ a.s., which implies that $\mathbb{G}_n\left\{\phi^a_{C,1}\right\} = o_{\mathbb{P}}(1)$.

Next for $\phi^a_{C,2}$, first note that $\sum_j f^a_j(\boldsymbol{\mu}) \to \mu_a - c_{k^*a}$ and that $\psi^a_C \to \mathbb{E}[(\mu_a - c_{k^*a})^2]$ by assumption (A3), (A4) and the dominated convergence theorem. Hence we define the limiting value of $\phi^a_{C,2}$ by

$$\bar{\phi}^a_{C^*}(A,X,Y) = 2(\mu_a - c_{k^*a})m(A,X,Y) + (\mu_a - c_{k^*a})^2 - \mathbb{E}[(\mu_a - c_{k^*a})^2] \tag{C.19}$$

which is a fixed function of Z, independent of n. As shown above, by the law of total expectation it is straightforward to show $\mathbb{E}\left[\phi^a_{C,2}\right] = 0$ and thus $\mathbb{E}\left[\bar{\phi}^a_{C^*}\right] = 0$, .

Furthermore, based on the similar algebra used to derive (C.16), with the additional fact that m_a is bounded, we have

$$\phi^a_{C,2} - \bar{\phi}^a_{C^*} \lesssim \sum_r \Upsilon'_r\left(\omega_r - \mathbb{1}\{r = k^*\}\right)$$

for some bounded $\Upsilon'_r \in \mathbb{R}$. Therefore, based on the exact same argument used to derive $\mathbb{G}_n\left\{\phi^a_{C,1}\right\} = o(1)$ a.s. as above, we conclude that

$$\begin{aligned}\mathbb{G}_n\left\{\phi^a_{C,2} - \bar{\phi}^a_{C^*}\right\} &= \sqrt{n}\mathbb{P}_n\left\{\phi^a_{C,2} - \bar{\phi}^a_{C^*}\right\} \\ &= o(1) \quad \text{a.s.}\end{aligned}$$

and thus $\mathbb{G}_n\left\{\phi^a_{C,2} - \bar{\phi}^a_{C^*}\right\} = o_{\mathbb{P}}(1)$.

Putting all the pieces together, finally we have

$$\begin{aligned}\mathbb{G}_n\left\{\varphi^a_C\right\} &= \mathbb{G}_n(\phi^a_C) \\ &= \mathbb{G}_n\left\{\phi^a_{C,1}\right\} + \mathbb{G}_n\left\{\phi^a_{C,2} - \bar{\phi}^a_{C^*}\right\} + \mathbb{G}_n\left\{\bar{\phi}^a_{C^*}\right\} \\ &= o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) + \mathbb{G}_n\left\{\bar{\phi}^a_{C^*}\right\}.\end{aligned}$$

Hence, by the central limit theorem and Slutsky theorem we obtain

$$\mathbb{G}_n\{\varphi_C^a\} \rightsquigarrow N\left(0, \operatorname{var}\left(\bar{\phi}_{C^*}^a\right)\right),$$

for any $a \in \mathscr{A}$. Therefore,

$$\mathbb{G}_n\left\{\sum_{a\in\mathscr{A}} \varphi_{C,\eta}^a\right\} \rightsquigarrow N\left(0, \operatorname{var}\left(\sum_{a\in\mathscr{A}} \bar{\phi}_{C^*}^a(Z)\right)\right).$$

□

Now we are in a position to prove Lemma 4.4.1.

Proof of Lemma 4.4.1

Proof. Using the same decomposition as in part ii) of the proof of Lemma C.3.4 first let us write $\phi_C^a = \phi_{C,1}^a + \phi_{C,2}^a - \bar{\phi}_{C^*}^a + \bar{\phi}_{C^*}^a$, where $\bar{\phi}_{C^*}^a$ is given in (C.19). Then from (C.16) and Assumption (d) we already know $\left\|\phi_{C,1}^a\right\| = o(1)$. Furthermore, by definition of f_j^a and the boundedness assumption (A4) one can easily show that

$$\left[\sum_j f_j^a(\boldsymbol{\mu})\right]^2 - (\mu_a - c_{k^*a})^2 \lesssim \sum_r c_{ra}\left(\mathbb{1}\{r = k^*\} - \omega_r\right).$$

Hence similarly as before, we obtain $\left\|\phi_{C,2}^a - \bar{\phi}_{C^*}^a\right\| = o(1)$ and $\psi_C^a - \psi_{C^*}^a = o(1)$ under Assumption (d), where $\psi_{C^*}^a \equiv \mathbb{E}[(\mu_a - c_{k^*a})^2]$.

Therefore we have

$$\begin{aligned}\|\overline{\varphi}_C^a - \varphi_C^a\| &\le \|\overline{\phi}_C^a - \phi_C^a\| + \|\overline{\psi}_C^a - \psi_C^a\| \\ &\le \|\overline{\bar{\phi}}_{C^*}^a - \bar{\phi}_{C^*}^a\| + o(1) + \|\overline{\psi}_{C^*}^a - \psi_{C^*}^a\| + o(1).\end{aligned}$$

Note that $\bar{\phi}_{C^*}^a$, $\psi_{C^*}^a$ are Lipschitz in $L_2(\mathbb{P})$ norm with respect to η and μ_a respectively, as they are everywhere differentiable and their first derivatives are all bounded in $L_2(\mathbb{P})$ norm. Hence, the last display is bounded by

$$\begin{aligned}&O(\|\overline{\eta} - \eta\|) + O(\|\overline{\mu} - \mu\|) + o(1) \\ &= O(o_{\mathbb{P}}(1)) + O(o_{\mathbb{P}}(1)) + o(1) = o_{\mathbb{P}}(1)\end{aligned}$$

where the equality follows by the given assumption (b). Consequently we have $O_{\mathbb{P}}\left(\sum_{a\in\mathscr{A}}\|\overline{\varphi}_C^a-\varphi_C^a\|\right)=o_{\mathbb{P}}(1)$.

Moreover, by Assumption (d) we have $\sqrt{n}R_n=o_{\mathbb{P}}(1)$. Finally, Lemma C.3.4 and the Slutzky theorem yield the result. □

C.3.3 Proof of Lemma 4.4.2

Before proceeding, we first introduce the following lemma which will be useful to prove Lemma 4.4.2.

Lemma C.3.5. *For any $a\in\mathscr{A}$ and sufficiently large n such that $h\wedge 1=h$, it follws that with probability at least* $1-\delta$

$$\sup_{C\in\mathscr{C}_k,\eta\in[\rho,1-\rho]^p\times\mathbb{R}^p}|(\mathbb{P}_n-\mathbb{P})\,\varphi_C^a(\eta)|\leq \mathsf{C}'\left(\sqrt{\frac{\log(1/\delta)}{nh^2}}+\frac{\log(1/\delta)}{nh}\right)$$

for any $\rho>0$ and global constant $\mathsf{C}'>0$ that does not depend on n,h,C,η.

Proof. Let us define a function class $\mathscr{G}_1{}^a\equiv\left\{g{1}_{C,\eta}^a:\mathscr{Z}\to\mathbb{R}\mid C\in\mathscr{C}_k,\eta\in[\rho,1-\rho]^p\times[-B,B]^p\right\}$ such that

$$g{1}_{C,\eta}^a(Z)=2\left[\sum_r f_r^a(W;C,h)\right]\sum_r\sum_{a'\in\mathscr{A}}\frac{\partial f_r^a(W;C,h)}{\partial\mu_{a'}}\frac{\mathbb{1}(A=a')}{\pi_{a'}}(Y-\mu_{a'}),$$

and a class $\mathscr{G}_2{}^a\equiv\left\{g{2}_{C,\mu}^a:\mathscr{Z}\to\mathbb{R}\mid C\in\mathscr{C}_k,\mu\in[-B,B]^p\right\}$ such that

$$g{2}_{C,\mu}^a(Z)=\left[\sum_r f_r^a(W;C,h)\right]^2,$$

where η is a set of all the nuisance parameters as before and $\mu=(\mu_1,...,\mu_p)$. Then, it follows that

$$\begin{aligned}&\sup_{C\in\mathscr{C}_k,\eta\in[\rho,1-\rho]^p\times\mathbb{R}^p}|(\mathbb{P}_n-\mathbb{P})\,\varphi_C^a(\eta)|\\&\qquad\leq\underbrace{\sup_{g_1\in\mathscr{G}_1^a}\left|\frac{1}{n}\sum_{i=1}^n g_1(Z_i)-\mathbb{E}\left[g_1(Z_i)\right]\right|}_{\text{(i)}}+\underbrace{\sup_{g_2\in\mathscr{G}_2^a}\left|\frac{1}{n}\sum_{i=1}^n g_2(Z_i)-\mathbb{E}\left[g_2(Z_i)\right]\right|}_{\text{(ii)}}.\end{aligned}$$

For the part (i) in RHS of the above inequality, we first note that from the proof of Lemma C.3.4 part (ii)

$$\|f_r^a\|_\infty \le 2B, \qquad \left\| \frac{\partial f_r^a}{\partial \mu_{a'}} \right\|_\infty \le \frac{2Bk}{h} + 1$$

for $\forall a, a' \in \mathscr{A}$, which leads to

$$\begin{aligned}\left\|g1^{a}_{C,\eta}\right\|_\infty &\le 2\left|\sum_r f_r^a\right| \left\{ \sum_r \sum_{a' \in \mathscr{A}} \left\| \frac{\partial f_r^a}{\partial \mu_{a'}} \frac{\mathbb{1}(A = a')}{\pi_{a'}} (Y - \mu_{a'}) \right\|_\infty \right\} \\ &\le 2k^2 B \left(\frac{2Bk}{h} + 1 \right) \left(\frac{|Y\|_\infty + B}{\rho} \right) \\ &\le (h \wedge 1)^{-1} \mathsf{C}_{k,B,\|Y\|_\infty,\rho}.\end{aligned}$$

with some finite constant $\mathsf{C}_{k,B,\|Y\|_\infty,\rho}$. Hence we conclude $\mathbb{E}_{\mathbb{P}}(g1^{a}_{C,\eta})^2 < \mathsf{C}^2_{k,B,\|Y\|_\infty,\rho}(h \wedge 1)^{-2}$.

Next, in order to consider the covering number of $\mathscr{G}^a$, suppose that for any indices r, a and some $\varepsilon > 0$

$$\left\|\mu_a - \mu_a'\right\|_\infty \le h\varepsilon, \quad \left\|C - C'\right\|_\infty \le h\varepsilon, \quad \left| \frac{1}{\pi_a} - \frac{1}{\pi_a'} \right| \le h\varepsilon, \quad \left|\omega_r - \omega_r'\right| \le h\varepsilon$$

where we use superscript $'$ to represent a different element in the same function/parameter class. Then, it is clear to see that

$$\left| \frac{\partial \omega_r}{\partial \mu_a} - \left(\frac{\partial \omega_r}{\partial \mu_a} \right)' \right| \le \frac{|\omega_r - \omega_r'|}{h} \le \varepsilon,$$

and

$$\begin{aligned}\left|f_r^a - (f_r^a)'\right| &\le \left|\omega_r - \omega_r'\right| \|\mu_a - c_{ra}\|_\infty + \|\omega_r\|_\infty \left\{ \left|\mu_a - \mu_a'\right| + \left|c_{ra} - c_{ra}'\right| \right\} \\ &\le h\varepsilon(2B + 1).\end{aligned}$$

Consequently it also follows that

$$\begin{aligned}\left| \frac{\partial f_r^a}{\partial \mu_{a'}} - \left(\frac{\partial f_r^a}{\partial \mu_{a'}} \right)' \right| &\le \left| \frac{\partial \omega_r}{\partial \mu_{a'}} - \left(\frac{\partial \omega_r}{\partial \mu_{a'}} \right)' \right| \|\mu_a - c_{ra}\|_\infty + \left|\omega_r - \omega_r'\right| \\ &\le \varepsilon(2B + h) \\ &\le \varepsilon \mathsf{C}_B'\end{aligned}$$

for some constant $\mathsf{C}'_B > 0$ as $h = o(1)$. Now we have

$$
\begin{aligned}
&\left\| g1^{a}_{C,\eta} - g1^{a}_{C',\eta'} \right\|_\infty \\
&\leq 2 \sum_r \| f_r^a - (f_r^a)' \|_\infty \left\| \sum_r \sum_{a' \in \mathscr{A}} \frac{\partial f_r^a}{\partial \mu_{a'}} \frac{\mathbb{1}(A = a')}{\pi_{a'}} (Y - \mu_{a'}) \right\|_\infty \\
&\quad + 2 \left\| \sum_r f_r^a \right\|_\infty \left\{ \sum_r \sum_{a' \in \mathscr{A}} \left\| \frac{\partial f_r^a}{\partial \mu_{a'}} \frac{\mathbb{1}(A = a')}{\pi_{a'}} (Y - \mu_{a'}) - \left(\frac{\partial f_r^a}{\partial \mu_{a'}} \right)' \frac{\mathbb{1}(A = a')}{\pi'_{a'}} (Y - \mu'_{a'}) \right\|_\infty \right\} \\
&\leq 2kh\varepsilon(2B+1)k \left\{ \frac{2Bk}{h} + 1 \right\} \frac{B + \|Y\|_\infty}{\rho} \\
&\quad + 4Bk \left\{ \left(\frac{2Bk}{h} + 1 \right) \left(\frac{h\varepsilon}{\rho} + (B + \|Y\|_\infty) h\varepsilon \right) + \varepsilon \mathsf{C}'_B \left(\frac{B + \|Y\|_\infty}{\rho} \right) \right\} \\
&\leq \varepsilon \mathsf{C}''_{k,B,\|Y\|_\infty,\rho}
\end{aligned}
$$

for some constant $\mathsf{C}''_{k,B,\|Y\|_\infty,\rho} > 0$, which follows by rearranging terms and applying triangle inequality with all the bounds we have discussed so far. Let $\varepsilon' = \varepsilon \mathsf{C}''_{k,B,\|Y\|_\infty,\rho}$. Then finally we have

$$
\begin{aligned}
&\mathscr{N}(\mathscr{G}^a, \|\cdot\|_\infty, \varepsilon') \\
&\quad \leq \mathscr{N}([-B,B]^p, \|\cdot\|_\infty, h\varepsilon) \mathscr{N}([-B,B]^{kp}, \|\cdot\|_\infty, h\varepsilon) \mathscr{N}([1, \frac{1}{\rho}]^p, \|\cdot\|_\infty, h\varepsilon)^p \mathscr{N}([0,1]^k, \|\cdot\|_\infty, h\varepsilon) \\
&\quad \leq \left(\frac{2B}{h\varepsilon} \right)^{p+kp} \left(\frac{1}{\rho h\varepsilon} \right)^p \left(\frac{1}{h\varepsilon} \right)^k \\
&\quad = (h\varepsilon / \mathsf{C}'''_{k,p,\rho,B})^{-2p-kp-k} \\
&\quad \leq \left(\frac{M \mathsf{C}_{k,B,\|Y\|_\infty,\rho} (h \wedge 1)^{-1}}{\varepsilon'} \right)^{2p+kp+k},
\end{aligned}
$$

where $M = \frac{\mathsf{C}''_{k,B,\|Y\|_\infty,\rho} \mathsf{C}'''_{k,p,\rho,B}}{\mathsf{C}_{k,B,\|Y\|_\infty,\rho}}$. Hence by Theorem 30 in Kim et al. [76], for some constants C'_1, C'_2, C'_3 , with probability at least $1 - \delta$

$$
\begin{aligned}
&\sup_{g_1 \in \mathscr{G}_1^a} \left| \frac{1}{n} \sum_{i=1}^n g_1(Z_i) - \mathbb{E}[g_1(Z_i)] \right| \\
&\leq \mathsf{C}'_1 \sqrt{\frac{(2p + kp + k) \mathsf{C}^2_{k,B,\|Y\|_\infty,\rho}}{n(h \wedge 1)^2} \log(2M)} + \mathsf{C}'_2 \sqrt{\frac{\mathsf{C}^2_{k,B,\|Y\|_\infty,\rho}}{n(h \wedge 1)^2} \log(1/\delta)} + \mathsf{C}'_3 \left(\frac{\mathsf{C}_{k,B,\|Y\|_\infty,\rho}}{n(h \wedge 1)} \log(1/\delta) \right).
\end{aligned}
$$

Hence provided that $h = h \wedge 1$, for sufficiently large constant $\mathsf{C}' > 0$ that only depends on $k, p, B, \rho, \|Y\|_\infty$ we have

$$\sup_{g_1 \in \mathscr{G}_1^a} \left| \frac{1}{n} \sum_{i=1}^{n} g_1(Z_i) - \mathbb{E}[g_1(Z_i)] \right| \leq \mathsf{C}' \left(\sqrt{\frac{\log(1/\delta)}{nh^2}} + \frac{\log(1/\delta)}{nh} \right)$$

with probability at least $1 - \delta$.

We omit the proof here for the sake of brevity, but we obtain the same upper bound for the part (ii) as well based on the similar procedure. Hence, the result follows.

□

Now we are back to the proof of Lemma 4.4.2.

Proof. First note that

$$\begin{aligned} R(\widehat{C}) - R(C^*) &= R(\widehat{C}) - \widehat{\psi}_{\widehat{C}} + \widehat{\psi}_{\widehat{C}} - R(C^*) \\ &\leq R(\widehat{C}) - \widehat{\psi}_{\widehat{C}} + \widehat{\psi}_{C^*} - R(C^*) \\ &\leq 2\|\widehat{\psi}_C - R(C)\|_{\mathscr{C}_k} \end{aligned}$$

where we adopt the notation $\|\widehat{\psi}_C - R(C)\|_{\mathscr{C}_k} \equiv \sup_{C \in \mathscr{C}_k} |\widehat{\psi}_C - R(C)|$. Then by the triangle inequality it follows that

$$\|\widehat{\psi}_C - R(C)\|_{\mathscr{C}_k} \leq \underbrace{\|R_h(C) - R(C)\|_{\mathscr{C}_k}}_{\text{i}} + \underbrace{\|\widehat{\psi}_C - R_h(C)\|_{\mathscr{C}_k}}_{\text{ii}}.$$

Let us analyze an asymptotic behavior of the right-hand side in the above display by term by term.

i) As we expand our scope to arbitrary $C \in \mathscr{C}_k$, we can no longer rely on the margin condition as in Lemma C.3.3 where we only considered an optimal codebook $C \in \mathscr{M}^*$. Nonetheless, for arbitrary C it still follows that

$$R_h(C) - R(C) \lesssim \mathbb{E}\left[\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^*}\right]$$

as seen in the proof of the Lemma C.3.3. Next for any ν such that $0 < \nu < 1$, we consider $N_C(h^\nu)$. Then, $\boldsymbol{\mu} \notin N_C(h^\nu) \Rightarrow h^\nu < \|\boldsymbol{\mu} - c_{k^{**}}\|_2 - \|\boldsymbol{\mu} - c_{k^*}\|_2$, and thereby

$$\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^{*}} \leq \begin{cases} \exp\left(-\frac{h^{\nu}}{h}\right) & \text{if } \boldsymbol{\mu} \notin N_C(h^{\nu}), \\ 1 & \text{elsewhere} \end{cases}.$$

Hence,

$$\begin{aligned}\mathbb{E}\left[\frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^{*}}\right] &\leq \int\limits_{\boldsymbol{\mu}\notin N_C(h^{\nu})} \frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^{*}} d\mathbb{P}(\boldsymbol{\mu}) + \int\limits_{\boldsymbol{\mu}\in N_C(h^{\nu})} \frac{\boldsymbol{K}^{**}}{\boldsymbol{K}^{*}} d\mathbb{P}(\boldsymbol{\mu}) \\ &\leq \exp\left(-h^{\nu-1}\right) + \int\limits_{\boldsymbol{\mu}\in N_C(h^{\nu})} d\mathbb{P}(\boldsymbol{\mu}) \\ &\lesssim \exp\left(-h^{\nu-1}\right) + h^{\nu} = O(h^{\nu}) = o(1), \end{aligned} \tag{C.20}$$

which leads to $R_h(C) - R(C) = o(1)$. Since this result is independent on C, we conclude that $\|R_h(C) - R(C)\|_{\mathscr{C}_k} = o(1)$.

ii) Using the following decomposition

$$\widehat{\psi}_C - R_h(C) = \sum_{a\in\mathscr{A}} \frac{1}{S}\sum_{s=1}^{S} \left(\mathbb{P}_n^s - \mathbb{P}^{-s}\right) \varphi_C^a(\hat{\eta}_{-s}) + \sum_{a\in\mathscr{A}} \frac{1}{S}\sum_{s=1}^{S} \mathbb{P}^{-s}\left\{\varphi_C^a(\hat{\eta}_{-s}) - \varphi_C^a(\eta)\right\}$$

where $\mathbb{P}^{-s}$ stands for the population distribution for samples not in group s, we have

$$\begin{aligned}\sup_{C\in\mathscr{C}_k} |\hat{\psi}_C - R_h(C)| \leq \frac{1}{S}\sum_{s=1}^{S}\sum_{a\in\mathscr{A}}\sup_{C\in\mathscr{C}_k} \left|\left(\mathbb{P}_n^s - \mathbb{P}^{-s}\right)\left\{\varphi_C^a(\hat{\eta}_{-s})\right\}\right| \\ + \frac{1}{S}\sum_{s=1}^{S}\sum_{a\in\mathscr{A}}\sup_{C\in\mathscr{C}_k} \left|\mathbb{P}^{-s}\left\{\varphi_C^a(\hat{\eta}_{-s}) - \varphi_C^a(\eta)\right\}\right|.\end{aligned}$$

For the first term in the right-hand side,

$$\begin{aligned}\sup_{C\in\mathscr{C}_k} \left|\left(\mathbb{P}_n^s - \mathbb{P}^{-s}\right)\left\{\varphi_C^a(\hat{\eta}_{-s})\right\}\right| &\leq \sup_{C\in\mathscr{C}_k, \eta\in[\rho,1-\rho]^p\times\mathbb{R}^p} \left|\left(\mathbb{P}_n^s - \mathbb{P}^{-s}\right)\varphi_C^a(\eta)\right| \\ &\cong \sup_{C\in\mathscr{C}_k, \eta\in[\rho,1-\rho]^p\times\mathbb{R}^p} \left|\left(\mathbb{P}_{n/S} - \mathbb{P}\right)\varphi_C^a(\eta)\right|,\end{aligned}$$

under random sample splitting, where $\cong$ means same in distribution. Hence by applying Lemma C.3.5, for each s we have that with probability at least $1-\delta$,

$$\sup_{C\in\mathscr{C}_k}\left|\left(\mathbb{P}_n^s-\mathbb{P}^{-s}\right)\{\varphi_C^a(\hat{\eta}_{-s})\}\right|\leq \mathsf{C}_S'\left(\sqrt{\frac{\log(1/\delta)}{nh^2}}+\frac{\log(1/\delta)}{nh}\right).$$

For the second term, from (C.12) we note that

$$\mathbb{P}^{-s}\{\varphi_C^a(\hat{\eta}_{-s})-\varphi_C^a(\eta)\}=R_2^a(\widehat{\mathbb{P}},\mathbb{P}),$$

where $\widehat{\mathbb{P}}$ is the distribution corresponding to $\hat{\eta}_{-s}$. Since our C is no longer guaranteed to be in $\mathscr{M}^*$ for the analysis of the above $R_2^a(\widehat{\mathbb{P}},\mathbb{P})$ we have to use (C.20) in the preceding part i). Based on the exact same algebra to deduce (C.18), we obtain

$$\begin{aligned}R_2^a(\widehat{\mathbb{P}},\mathbb{P})&\lesssim\sum_{a'\in\mathscr{A}}\left\{\left\|\sum_r\frac{\partial f_r^a}{\partial\mu_{a'}}\right\|\|\pi_{a'}-\widehat{\pi}_{a'}\|\|\mu_{a'}-\widehat{\mu}_{a'}\|\right\}+\sum_{a',a''\in\mathscr{A}}\left\|\frac{\partial^2 g}{\partial\mu_{a'}\partial\mu_{a''}}\right\|\|\mu_{a'}-\widehat{\mu}_{a'}\|\|\mu_{a''}-\widehat{\mu}_{a''}\|\\&\lesssim\left(kh^{\frac{\nu}{2}-1}+1\right)\sum_{a'\in\mathscr{A}}\|\pi_{a'}-\overline{\pi_{a'}}\|_{\mathbb{P},4}\|\mu_{a'}-\overline{\mu_{a'}}\|_{\mathbb{P},4}\\&\quad+\left(k^2h^{\frac{\nu}{2}-2}+kh^{\frac{\nu}{2}-1}+1\right)\sum_{a',a''\in\mathscr{A}}\|\mu_{a'}-\overline{\mu_{a'}}\|_{\mathbb{P},4}\|\mu_{a''}-\overline{\mu_{a''}}\|_{\mathbb{P},4}\end{aligned}\tag{C.21}$$

for any $C\in\mathscr{C}_k$.

From Assumption (d), we have $\|\pi_{a'}-\widehat{\pi}_{a'}\|\|\mu_{a'}-\widehat{\mu}_{a'}\|=o_{\mathbb{P}}(n^{-1/2})$, $\|\mu_{a'}-\widehat{\mu}_{a'}\|\|\mu_{a''}-\widehat{\mu}_{a''}\|=o_{\mathbb{P}}(n^{-1/2})$ for $\forall a',a''\in\mathscr{A}$, and $h^{\frac{\nu}{2}-2}=h^{\frac{\nu}{2}}h^{-2}=o(1)o_{\mathbb{P}}(n^{1/2})$. Hence $R_2^a(\widehat{\mathbb{P}},\mathbb{P})=o_{\mathbb{P}}(1)$, independent of C. Consequently

$$\sup_{C\in\mathscr{C}_k}\left|\mathbb{P}^{-k}\{\varphi_C^a(\hat{\eta}_{-s})-\varphi_C^a(\eta)\}\right|=o_{\mathbb{P}}(1).$$

Putting these together, we finally conclude

$$\|\hat{\psi}_C-R_h(C)\|_{\mathscr{C}_k}=O_{\mathbb{P}}\left(\frac{1}{\sqrt{nh^2}}\right)+o_{\mathbb{P}}(1).$$

Now, from above part i) and part ii) we have

$$\|\widehat{\psi}_C-R(C)\|_{\mathscr{C}_k}=o(1)+O_{\mathbb{P}}\left(\frac{1}{\sqrt{nh^2}}\right)+o_{\mathbb{P}}(1)=o_{\mathbb{P}}(1)+O_{\mathbb{P}}\left(o_{\mathbb{P}}(1)\right)=o_{\mathbb{P}}(1)$$

where the second last equality follows by $\frac{1}{\sqrt{nh^2}}=o_{\mathbb{P}}(1)$ from Assumption (d).

Finally, the desired consistency can be shown by validating Theorem 5.7 in [141] where we set $M(\cdot) = -R(\cdot), M_n(\cdot) = -\widehat{\psi}_{(\cdot)}$. We already have verified that $\| \widehat{\psi}_C - R(C) \|_{\mathscr{C}_k} = o_{\mathbb{P}}(1)$. Furthermore, since $\widehat{C}$ is a minimizer of $\widehat{\psi}$ it is clear that $-\widehat{\psi}_{C^*} + \widehat{\psi}_{\widehat{C}} \leq 0$.

Note that $R(\cdot)$ is a bounded, continuous function whose domain ($\mathscr{C}_k$) is compact. Hence due to the local uniqueness condition (d), each C^* is a locally well-separated minimizer of R. Consequently, by the same logic used in Theorem 5.7 in [141], one may show that for each $\delta > 0$ the probability of the event $\{d_{\text{codebook}}(\widehat{C}, C^*) < \delta, \text{ for some } C^* \in \mathscr{M}^*\}$ converges to 1. Thus we conclude that $\widehat{C}$ converges in probability to some $C^* \in \mathscr{M}^*$, which yields the result.

□

C.3.4 Proof of Theorem 4.4.2

The following lemma emphasizes the fact that if a codebook $C \in \mathscr{C}_k$ is sufficiently similar to $C^* \in \mathscr{M}^*$ (in terms of a valid distance function), then C satisfies the margin condition as well.

Lemma C.3.6. *Let C, C^* belong to $\mathscr{C}_k, \mathscr{M}^*$ respectively. If C is close enough to C^*, then C also satisfies the margin condition.*

Proof. Let W be in $N_C(\kappa')$ for some $\kappa' > 0$. Without loss of generality, let $c_i^* \in C^*$ denote the nearest optimal cluster center for W, i.e. $\boldsymbol{\mu} \in V_i(C^*)$, for a fixed $i \in \{1, ..., k\}$.

First consider the case $\boldsymbol{\mu} \in V_i(C^*) \bigcap V_i(C)$. Then by the triangle inequality for $\forall j \neq i$ such that $V_j(C^*)$ is adjacent to $V_i(C^*)$,

$$\begin{aligned} \|\boldsymbol{\mu} - c_j^*\|_2 - \|\boldsymbol{\mu} - c_i^*\|_2 &\leq \|\boldsymbol{\mu} - c_j\|_2 + \|c_j - c_j^*\|_2 - \|\boldsymbol{\mu} - c_i\|_2 + \|c_i - c_i^*\|_2 \\ &\lesssim \kappa' + d_{\text{codebook}}(C, C^*) \end{aligned}$$

for a valid distance d_{codebook} equipped with metric space $(\mathbb{R}^{|\mathscr{A}|}, L_2)$. Next we consider the case where $\boldsymbol{\mu} \in V_i(C^*) \bigcap V_j(C)$ for $\forall j \neq i$. In this case we have

$$\begin{aligned} \|\boldsymbol{\mu} - c_j^*\|_2 - \|\boldsymbol{\mu} - c_i^*\|_2 &\leq d\left(\partial V_i(C^*), c_j^*\right) + d\left(\partial V_i(C^*), W\right) - d\left(\partial V_i(C^*), c_i^*\right) + d\left(\partial V_i(C^*), W\right) \\ &\leq 2d\left(\partial V_i(C^*), W\right) \\ &\lesssim d_{\text{codebook}}(C, C^*) \end{aligned}$$

where the first inequality follows by the triangle inequality, the second by $d\left(\partial V_i(C^*), c_i\right) = d\left(\partial V_i(C^*), c_j\right)$ for all $V_j(C^*)$ adjacent to $V_i(C^*)$, and the last by Lemma 4.2 in [85].

Hence by setting κ' properly, for sufficiently small value of $d_{\text{codebook}}(C, C^*)$ we obtain

$$\min_{j \neq i} \left\{ \|\boldsymbol{\mu} - c_j^*\|_2 - \|\boldsymbol{\mu} - c_i^*\|_2 \right\} \leq \kappa$$

for $\boldsymbol{\mu} \in V_i(C^*)$. If we take a minimum of such κ' over all i, generalization to $\forall i$ can be done. Hence we conclude that $\boldsymbol{\mu} \in N_{C^*}(\kappa)$. □

Note that value of the margin gap does not affect our result in Lemma 4.4.1. Having Lemma C.3.6, it can be said that there exists a constant $\tau_\kappa > 0$ such that if $d_{\text{codebook}}(C, C^*) \leq \tau_\kappa$ then C satiesfies the margin condition as well.

We are now in a position to prove Theorem 4.4.2.

Proof of Theorem 4.4.2. Let C^* belong to $\mathscr{M}^*$. By Lemma 4.4.1, for each $\varepsilon > 0$ we can always find M_ε and n_ε such that $\mathbb{P}\left(\sqrt{n}|\widehat{\psi}_{C^*} - R(C^*)| > M_\varepsilon\right) < \varepsilon$ for all $n \geq n_\varepsilon$. Furthermore by the result of Lemma C.3.6, we can find the constant $\tau_\kappa > 0$ that makes a codeset C satiesfy the margin condition whenever $d_{\text{codebook}}(C, C^*) \leq \tau_\kappa$. Lastly note that by Lemma 4.4.2 we have $d_{\text{codebook}}(\widehat{C}, C^*) \xrightarrow{P} 0$, and thus for each $\varepsilon' > 0$ there exists $n_{\varepsilon'}$ such that $\mathbb{P}\left(d_{\text{codebook}}(\widehat{C}, C^*) > \tau_\kappa\right) < \varepsilon'$ for all $n \geq n_{\varepsilon'}$. Now by the law of total probability,

$$\begin{aligned}
\mathbb{P}\left(\sqrt{n}\left|\widehat{\psi}_{\widehat{C}} - R(\widehat{C})\right| > M_\varepsilon\right) &\leq \mathbb{P}\left(\sqrt{n}\left|\widehat{\psi}_{\widehat{C}} - R(\widehat{C})\right| > M_\varepsilon \mid d_{\text{codebook}}(\widehat{C}, C^*) \leq \tau_\kappa\right) + \mathbb{P}\left(d_{\text{codebook}}(\widehat{C}, C^*) > \tau_\kappa\right) \\
&< \mathbb{P}\left(\sqrt{n}|\widehat{\psi}_{C^*} - R(C^*)| > M_\varepsilon\right) + \varepsilon' \\
&< \varepsilon + \varepsilon'
\end{aligned}$$

for all $n \geq \max(n_\varepsilon, n_{\varepsilon'})$. Note that the second inequality follows by the local uniqueness of C^*, along with Lemma C.3.6. As ε and ε' are both arbitrary we conclude that

$$\sqrt{n}\left(\widehat{\psi}_{\widehat{C}} - R(\widehat{C})\right) = O_{\mathbb{P}}(1).$$

Finally,

$$\begin{aligned}
R(\widehat{C}) - R(C^*) &\leq R(\widehat{C}) - \widehat{\psi}_{\widehat{C}} + \widehat{\psi}_{C^*} - R(C^*) \\
&= O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

□

C.3.5 Proof of Theorem 4.4.3

To proceed, we define a function $\phi(Z;C,\eta') \equiv \phi_{C,\eta'} : \mathscr{Z} \to \mathbb{R}^{k\times|\mathscr{A}|}$ by $\phi_{C,\eta'} = \nabla_{C'=C}\{\varphi_{C'}(\eta')\}$, a vector of partial derivatives of the uncentered EIF φ with respect to each cluster center $c' \in C'$ evaluated at value $c \in C$, indexed by a cluster codebook C and a set of nuisance parameters η'. Then our estimate $\widehat{C}$ is zeros of the equation $\frac{1}{S}\sum_{s=1}^{S}\mathbb{P}_n^s\{\phi_{C,\hat{\eta}_{-s}}\} = 0$.

Let η denote true nuisance parameters. As before, we let $C^* \in \mathscr{M}^*$ denote a minimizer of the true risk function $R(C)$. In additiona to that, let us write $C_0 \equiv \underset{C\in\mathscr{C}_k}{\operatorname{argmin}}\,\psi_{C,\eta}$, a minimizer of the kernel-smoothed risk $\psi_{C,\eta} = R_h(C)$. In Lemma 4.4.2 we gives the result that $\widehat{C}$ is consistent to C^*. The following lemma provides the same consistency guarantee for C_0.

Lemma C.3.7. *Under the same condition as in Lemma 4.4.2, C_0 converges in probability to C^*.*

Proof. It suffices to show that $\widehat{C}$ converges to C_0 by recycling the proof structure used in Lemma 4.4.2: i.e., showing that assumptions of Theorem 5.7 in [141] are satisfied. We already showed that $\|\widehat{\psi}_{C,\eta} - \psi_{C,\eta}\|_{\mathscr{C}_k} = o_{\mathbb{P}}(1)$ in the second part of Lemma 4.4.2. Next by definition $\widehat{\psi}_{\widehat{C},\eta} - \psi_{C,\eta} \leq 0$.

Finally we claim that C_0 is a well-separated point of minimum of a function $R_h(\cdot) = \psi_{\cdot,\eta}$. Suppose that it is not and for any $\delta > 0$ there exists another minimizer C_1 such that $R_h(C_1) = R_h(C_0)$ and $d(C_1,C_0) \geq \delta$ for all n. Then since $\|R_h(C) - R(C)\|_{\mathscr{C}_k} = o(1)$ from the first part of Lemma 4.4.2, on one hand it follows that

$$\begin{aligned} R_h(C_0) - R(C_1) &= R_h(C_0) - R(C_0) + R(C_0) - R(C_1) \\ &= o(1) + R(C_0) - R(C_1) = o(1) \end{aligned}$$

which yields $R(C_0) - R(C_1) = o(1)$. On the other hand, $R_h(C_0) = \min_{C\in\mathscr{C}_k} R_h(C) = \min_{C\in\mathscr{C}_k}\{R(C) + o(1)\} = R(C^*) + o(1)$ for some $C^* \in \mathscr{M}^*$ and $R_h(C_0) = R(C_0) + o(1)$. Hence $R(C^*) = R(C_0) + o(1)$.

Consequently we have $R(C^*) - R(C_0) = o(1)$ and $R(C^*) - R(C_1) = o(1)$. Since each C^* is a locally well-separated point of minimum of R, there exists n_δ such that for $n \geq n_\delta$ $d(C^*,C_0) < \frac{\delta}{2}$ and $d(C^*,C_1) < \frac{\delta}{2}$, which yields $d(C_0,C_1) < \delta$, a contradiction. Hence we conclude that C_0 is a well-separated point of minimum of R_h, and thereby $\widehat{C} \xrightarrow{P} C_0$.

□

Note that Lemma C.3.7 does not require the strong margin condition in Definition **??**. With this lemma, we are ready to prove Theorem 4.4.3.

Proof of Theorem 4.4.3

Proof. First note that by Leibniz's rule

$$\begin{aligned}\mathbb{P}\left\{\phi_{C_0,\eta}\right\} &= \mathbb{P}\left\{\nabla_{C=C_0}\left(\phi_{C,\eta} + \psi_{C,\eta}\right)\right\} \\ &= \nabla_{C=C_0}\left\{\mathbb{P}\left(\phi_{C,\eta}\right)\right\} + \mathbb{P}\left\{\nabla_{C=C_0}\psi_{C,\eta}\right\} \\ &= 0\end{aligned}$$

as $\mathbb{P}\left(\phi_{C,\eta}\right) = 0$ for $\forall C$ and $\nabla_{C=C_0}\psi_{C,\eta} = 0$ by definition. Thus it follows,

$$\begin{aligned}0 &= \sqrt{n}\left[\frac{1}{S}\sum_{s=1}^{S}\mathbb{P}_n^s\left\{\phi_{\widehat{C},\hat{\eta}_{-s}}\right\} - \mathbb{P}\left\{\phi_{C_0,\eta}\right\}\right] \\ &= \frac{\sqrt{n}}{S}\sum_{s=1}^{S}\mathbb{P}_n^s\left\{\phi_{\widehat{C},\hat{\eta}_{-s}} - \phi_{C^*,\hat{\eta}_{-s}}\right\} + \sqrt{n}\left[\frac{1}{S}\sum_{s=1}^{S}\mathbb{P}_n^s\left\{\phi_{C^*,\hat{\eta}_{-s}}\right\} - \mathbb{P}\left\{\phi_{C_0,\eta}\right\}\right]. \qquad \text{(C.22)}\end{aligned}$$

Fix s. For any unit in group s, by Taylor's theorem we have

$$\phi_{\widehat{C},\hat{\eta}_{-s}} - \phi_{C^*,\hat{\eta}_{-s}} = \sum_{|\alpha|=1} D^{\alpha}\phi_{C^*,\hat{\eta}_{-s}}(\widehat{C} - C^*) + \frac{1}{2}\sum_{|\alpha|=2} D^{\alpha}\phi_{\widetilde{C^*},\hat{\eta}_{-s}}(\widehat{C} - C^*)^2$$

for some $\widetilde{C^*}$ between $\widehat{C}$ and C^* (in terms of linear interpolation between each pair of points), where D^{α} is the differential operator for multi-index $\alpha = (\alpha_1, ..., \alpha_{|\mathscr{A}|k})$ of length $|\mathscr{A}| \times k$ and $|\alpha| = \sum_{l=1}^{|\mathscr{A}|k}\alpha_l$.

On one hand, since ϕ_C consists of the first partial derivatives with respect to each coordinate of C, $\sum\limits_{|\alpha|=1} D^{\alpha}\phi_{C^*,\hat{\eta}_{-s}}$ is a summation of all the second order partial derivatives with respect to each coordinate of C^*. For a fixed $a \in \mathscr{A}$ and label indices j, j', l, due to their structural resemblance $\frac{\partial f_l^a}{\partial c_{ja}}$ can be analysed through the very similar algebra used for $\frac{\partial f_l^a}{\partial \mu_a}$ in part ii of Lemma C.3.4. Thus again we may write,

$$\sum_j\sum_l \frac{\partial f_l^a}{\partial c_{ja}}(\mu_a - c_{la}) \lesssim \sum_j \omega_j + \frac{1}{h}\sum_j \Upsilon_j\left(\omega_j - \mathbb{1}\{j = k^*\}\right)$$

for some bounded $\Upsilon_j \in \mathbb{R}$ for all $j \in \mathbb{N}$. Under the strong margin condition, we obtain faster rates of the approximation term

$$\sum_l \Upsilon_l\left(\omega_l - \mathbb{1}\{l = k^*\}\right) \lesssim \exp(-\frac{\kappa}{h}),$$

which leads to

$$\frac{\partial^2 \varphi_{C^*}}{\partial c_{ja} \partial c_{j'a}} \lesssim 1 + \frac{1}{h^2} \exp(-\frac{\kappa}{h}), \quad \text{and consequently} \quad \sum_{|\alpha|=1} D^\alpha \phi_{C^*,\hat{\eta}_{-s}} \lesssim 1 + \frac{1}{h^2} \exp(-\frac{\kappa}{h}).$$

This is an analogous result to (C.16) and (C.17) in Lemma C.3.4. On the other hand, when $|\alpha| \geq 2$ the addition of 1 (from $\sum_j \omega_j$) no longer exists as we start taking the additional derivative at the third or higher order, and it could be deduced that

$$\left| \frac{\partial^{|\alpha|}}{\partial^{\alpha_1} c_{1,a_1} \cdots \partial^{\alpha_{|\mathscr{A}|}k} c_{k,a_{|\mathscr{A}|}}} \varphi_{C^*,\eta} \right| \lesssim \sum_{j=1}^{|\alpha|} \frac{b_j}{h^j} \left\{ \exp\left(-\frac{\kappa}{h}\right) + \cdots + \exp\left(-\frac{j\kappa}{h}\right) \right\}$$

where b_j's are finite constants which do not vary with n. Hence $\sum_{|\alpha|=2} D^\alpha \phi_{C^*,\hat{\eta}_{-s}} \lesssim \frac{1}{h^2} \exp(-\frac{\kappa}{h})$
[1].

Next, we claim that for any codeset $\widetilde{C}$ such that $\widetilde{C} \xrightarrow{P} C^*$, $\|D^\alpha \phi_{\widetilde{C},\hat{\eta}_{-s}}\| = O_{\mathbb{P}}(1)$ for all α such that $|\alpha| \geq 0$. To this end, first we note that for each $\varepsilon > 0$ there exists n_ε such that $\mathbb{P}\left(d_{\text{codebook}}(\widetilde{C}, C^*) > \tau_\kappa\right) < \varepsilon$ for all $n \geq n_\varepsilon$, where we interpret the constant τ_κ in the exact same way as in Lemma C.3.6. Furthermore from before, it is clear that $\|D^\alpha \phi_{C^*,\hat{\eta}_{-s}}\| = O(1)$ for every α and $C^* \in \mathscr{M}^*$. Hence there exists a univeral constant M such that $\|D^\alpha \phi_{C^*,\hat{\eta}_{-s}}\| \leq M$. Now it follows

$$\begin{aligned}
\mathbb{P}\left(\|D^\alpha \phi_{\widetilde{C},\hat{\eta}_{-s}}\| \leq M\right) &\geq \mathbb{P}\left(\|D^\alpha \phi_{\widetilde{C},\hat{\eta}_{-s}}\| \leq M \mid d_{\text{codebook}}(\widehat{C}, C^*) \leq \tau_\kappa\right) \mathbb{P}\left(d_{\text{codebook}}(\widehat{C}, C^*) \leq \tau_\kappa\right) \\
&> \mathbb{P}\left(\|D^\alpha \phi_{C^*,\hat{\eta}_{-s}}\| \leq M\right)(1-\varepsilon) \\
&= (1-\varepsilon).
\end{aligned}$$

Since ε is arbitrary, we get the desired result. Now we notice that

$$\text{var}\left(\frac{\sqrt{n}}{S} \mathbb{P}_n^s \left\{ D^\alpha \phi_{\widetilde{C^*},\hat{\eta}_{-s}} \right\}\right) = \frac{n}{S^2} \frac{S}{n} \text{var}\left(D^\alpha \phi_{\widetilde{C^*},\hat{\eta}_{-s}}\right) \leq \frac{1}{S} \|D^\alpha \phi_{\widetilde{C^*},\hat{\eta}_{-s}}\|^2.$$

Hence using Chebyshev's inequality, it follows

$$\mathbb{P}\left(\frac{\left| \frac{\sqrt{n}}{S} \mathbb{P}_n^s \left\{ D^\alpha \phi_{\widetilde{C^*},\hat{\eta}_{-s}} \right\} \right|}{\|D^\alpha \phi_{\widetilde{C^*},\hat{\eta}_{-s}}\| / S} \geq t \right) \leq \frac{1}{t^2}$$

[1] We omit the detailed algebra here since they are mostly very analogous to those used in the part ii of the proof of Lemma C.3.4 and thereby not particularly illuminating here.

for any $t > 0$. By letting $t = 1/\sqrt{\varepsilon}$ and noting $\|D^{\alpha}\phi_{\widetilde{C^*},\hat{\eta}_{-s}}\| = O_{\mathbb{P}}(1)$, we obtain that $\frac{\sqrt{n}}{S}\mathbb{P}_n^s\left\{D^{\alpha}\phi_{\widetilde{C^*},\hat{\eta}_{-s}}\right\} = \frac{1}{S}O_{\mathbb{P}}(1)$.

Finally we are ready to analyze the first term in (C.22). Putting all things together,

$$
\begin{aligned}
\frac{\sqrt{n}}{S}\sum_{s=1}^{S}\mathbb{P}_n^s\left\{\phi_{\widehat{C},\hat{\eta}_{-s}} - \phi_{C^*,\hat{\eta}_{-s}}\right\} &= \sqrt{n}(\widehat{C}-C^*)\frac{1}{S}\sum_{s=1}^{S}\left\{\mathbb{P}_n\left(1+\frac{1}{h^2}\exp(-\frac{\kappa}{h})\right)\right\} \\
&\quad + \frac{1}{2}\sum_{s=1}^{S}\frac{\sqrt{n}}{S}\mathbb{P}_n^s\left\{\sum_{|\alpha|=2}D^{\alpha}\phi_{\widetilde{C^*},\hat{\eta}_{-s}}\right\}(\widehat{C}-C^*)^2 \\
&= \sqrt{n}(\widehat{C}-C^*) + o_{\mathbb{P}}(1)o(1) + \frac{1}{S}\sum_{s=1}^{S}O_{\mathbb{P}}(1)\,o_{\mathbb{P}}(1) \\
&= \sqrt{n}(\widehat{C}-C^*) + o_{\mathbb{P}}(1). \qquad \text{(C.23)}
\end{aligned}
$$

Now for the second term in (C.22) we have,

$$
\begin{aligned}
&\sqrt{n}\left[\frac{1}{S}\sum_{s=1}^{S}\mathbb{P}_n^s\left\{\phi_{C^*,\hat{\eta}_{-s}}\right\} - \mathbb{P}\left\{\phi_{C_0,\eta}\right\}\right] \\
&= \frac{1}{S}\sum_{s=1}^{S}\sqrt{n}\left(\mathbb{P}_n^s-\mathbb{P}\right)\left\{\phi_{C^*,\hat{\eta}_{-s}} - \phi_{C^*,\eta}\right\} + \frac{\sqrt{n}}{S}\sum_{s=1}^{S}\left(\mathbb{P}_n^s-\mathbb{P}\right)\phi_{C_0,\eta} \\
&\quad + \frac{1}{S}\sum_{s=1}^{S}\sqrt{n}\mathbb{P}\left\{\phi_{C^*,\hat{\eta}_{-s}} - \phi_{C^*,\eta}\right\} + \frac{\sqrt{n}}{S}\sum_{s=1}^{S}\mathbb{P}_n^s\left\{\phi_{C^*,\eta} - \phi_{C_0,\eta}\right\} \\
&= \underbrace{\frac{1}{S}\sum_{s=1}^{S}\mathbb{G}_n^s\left\{\phi_{C^*,\hat{\eta}_{-s}} - \phi_{C^*,\eta}\right\}}_{\text{i}} + \underbrace{\mathbb{G}_n\phi_{C_0,\eta}}_{\text{ii}} \\
&\quad + \underbrace{\frac{1}{S}\sum_{s=1}^{S}\sqrt{n}\mathbb{P}\left\{\phi_{C^*,\hat{\eta}_{-s}} - \phi_{C^*,\eta}\right\}}_{\text{iii}} + \underbrace{\sqrt{n}\mathbb{P}_n\left\{\phi_{C^*,\eta} - \phi_{C_0,\eta}\right\}}_{\text{iv}} \qquad \text{(C.24)}
\end{aligned}
$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ and $\mathbb{G}_n^s = \sqrt{n}(\mathbb{P}_n^s - \mathbb{P})$ as before. In (C.24), we shall proceed term by term as below.

i) Without loss of generality let us fix $c \in C^*$, $a \in \mathscr{A}$ and by abuse of notation let $\varphi'_{C,\eta}$ denote a particular coordinate of $\phi_{C,\eta}$ corresponding to c and a, i.e. $\varphi'_{C,\eta} = \frac{\partial}{\partial c_a}\varphi(\cdot;C,\eta)$. Then we define a function class $\mathscr{F}^* = \left\{\varphi'_{C,\eta} : C \in \mathscr{M}^*\right\}$, where η is still defined as the set of true nuisance parameters.

As shown in above $\|D^{\alpha}\phi_{C^*,\hat{\eta}_{-s}}\| = O(1)$ for arbitrary α such that $|\alpha| \geq 0$ and $\mathscr{M}^* \subset \mathscr{C}_k$ is compact, $\mathscr{F}^*$ is Lipschitz of order $\lfloor k/2 \rfloor + 1$ with respect to each coordinate of C. Hence, by Theorem 2.7.2 in [142] we have

$$\log N_{[]}\left(\varepsilon, \mathscr{F}^*, L_2(\mathbb{P})\right) \lesssim \left(\frac{1}{\varepsilon}\right)^{\frac{k|\mathscr{A}|}{\lfloor k|\mathscr{A}|/2\rfloor+1}}.$$

Therefore by Theorem 19.5 in [141], $\mathscr{F}^*$ is Donsker and due to the consistency of $\hat{\eta}_{-s}$ we have

$$\mathbb{G}_n^s\left\{\varphi'_{C^*,\hat{\eta}_{-s}} - \varphi'_{C^*,\eta}\right\} = o_{\mathbb{P}}(1).$$

This claim holds for every pair of c, a. Thus we conclude that $\mathbb{G}_n^s\left\{\phi_{C^*,\hat{\eta}_{-s}} - \phi_{C^*,\eta}\right\} = o_{\mathbb{P}}(1)$.

ii) We may write

$$\mathbb{G}_n \phi_{C_0,\eta} = \mathbb{G}_n\left\{\phi_{C_0,\eta} - \phi_{C^*,\eta}\right\} + \mathbb{G}_n \phi_{C^*,\eta}.$$

As shown previously in part i, a function class for each coordinate of $\phi_{C^*,\eta}$ is Donsker. Furthermore $C_0 \xrightarrow{P} C^*$ by Lemma C.3.7. Hence basically by recycling the argument used in part i, we obtain $\mathbb{G}_n\left\{\phi_{C_0,\eta} - \phi_{C^*,\eta}\right\} = o_{\mathbb{P}}(1)$.

Next, to analyze $\mathbb{G}_n \phi_{C^*,\eta}$ we can follow the same logic used in the part iii in Lemma C.3.4. Again let $\varphi'_{C,\eta}$ denote a particular coordinate of $\phi_{C,\eta}$ corresponding to a pair c, a; so we need to take an additional derative of the EIF $\varphi_{C^*,\eta}$ with respect to $c_a^* \in C^*$. After a course of simple algebra as in the part iii in Lemma C.3.4, we may write

$$\begin{aligned}
\varphi'_{C^*,\eta}(Z) &= m_a(Z) + (\mu_a - c^*_{k^*a})\, m_a(Z) + \widetilde{\varphi}'_{C^*,\eta}(Z) \\
&\equiv \bar{\varphi}'_{C^*,\eta}(Z) + \widetilde{\varphi}'_{C^*,\eta}(Z)
\end{aligned}$$

with the same definition of m_a and k^* as before, where all the terms in $\widetilde{\varphi}'_{C^*,\eta}$ decay to zero at exponential rates this time due to the strong margin condition (unlike before we do not need any probablistic arguments). Namely, the function $\varphi'_{C^*,\eta}$ consists of the fixed function $\bar{\varphi}'_{C^*,\eta}$ and the shrinking function $\widetilde{\varphi}'_{C^*,\eta}$ whose terms go to zero at exponential rates. Finally, the central limit theorem and Slutsky theorem yield

$$\mathbb{G}_n \phi_{C^*,\eta} \rightsquigarrow N\left(0, \Sigma'_{C^*,\eta}\right) \tag{C.25}$$

where $\Sigma'_{C^*,\eta}$ is a covariance matrix for due to $\bar{\varphi}'_{C^*,\eta}$ part at each coordinate.

iii) As shown in (C.14) in the proof of Lemma C.3.4, we have

$$\mathbb{P}\left\{\phi_{C^*,\hat{\eta}_{-s}} - \phi_{C^*,\eta}\right\} \lesssim R_2^a(\widehat{\mathbb{P}}, \mathbb{P})$$

and under the condition in Theorem 4.4.2 it follows $R_2^a(\widehat{\mathbb{P}}, \mathbb{P}) = o_{\mathbb{P}}(n^{-1/2})$. Consequently,

$$\sqrt{n}\mathbb{P}\left\{\phi_{C^*,\hat{\eta}_{-s}} - \phi_{C^*,\eta}\right\} = o_{\mathbb{P}}(1).$$

iv) By Lemma C.3.7, we have $C_0 \xrightarrow{P} C^*$. Thus we can proceed with the same argument we used for analyzing the first term in (C.22). Using Taylor's theorem and noting that $\|D^\alpha \phi_{C^*,\eta}\| = O(1)$ we obtain $\|\phi_{C_0,\eta} - \phi_{C^*,\eta}\| = o_{\mathbb{P}}(1)$. Then by the Chebyshev's inequality argument as before, we conclude that $\sqrt{n}\mathbb{P}_n\left\{\phi_{C^*,\eta} - \phi_{C_0,\eta}\right\} = o_{\mathbb{P}}(1)$.

Putting all the pieces in part i - iv together we are guaranteed an asymptotic normaility in (C.24). Therefore, plugging all the results back into (C.22), again by Slutsky theorem we reach to the desired result.

□