Topics in causal inférence and policy learning with applications to precision medicine

Pan Zhao

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistique

École doctorale : Information, Structures, Systèmes

Unité de recherche : Institut Desbrest d'Épidémiologie et de Santé Publique (IDESP), France

TOPICS IN CAUSAL INFERENCE AND POLICY LEARNING WITH APPLICATIONS TO PRECISION MEDICINE

Présentée par Pan Zhao
Le 4 septembre 2024

Sous la direction de Julie JOSSE
et Antoine CHAMBAZ

Devant le jury composé de

Agathe GUILLOUX, Directrice de recherche, Inria	Rapporteure
Alex LUEDTKE, Associate Professor, University of Washington	Rapporteur
Fabrizia MEALLI, Professor, European University Institute	Examinatrice (Présidente)
Stijn VANSTEELANDT, Professor, Universiteit Gent	Examinateur
Stefan WAGER, Associate Professor, Stanford Graduate School of Business	Examinateur
Antoine CHAMBAZ, Professeur, Université Paris Cité	Directeur
Julie JOSSE, Directrice de recherche, Inria	Directrice
Shu YANG, Associate Professor, North Carolina State University	Invitée

UNIVERSITÉ DE MONTPELLIER

ABSTRACT

Causality is a fundamental concept in science and philosophy, and with the increasing complexity of data collection and structure, statistics plays a pivotal role in inferring causes and effects. This thesis delves into advanced causal inference methods, with a focus on policy learning, instrumental variables (IV), and difference-in-differences (DiD) approaches.

The IV and DiD methods are critical tools widely used by researchers in fields like epidemiology, medicine, biostatistics, econometrics, and quantitative social sciences. However, these methods often face challenges due to restrictive assumptions, such as the IV's requirement to have no direct effect on the outcome other than through the treatment, and the parallel trends assumption in DiD, which may be violated in the presence of unmeasured confounding.

In that context, this thesis introduces an innovative instrumented DiD approach to policy learning, which combines these two natural experiments to relax some of the key assumptions of conventional IV and DiD methods. To the best of our knowledge, the thesis presents the first comprehensive study of policy learning under the DiD setting. The direct policy search approach is proposed to learn optimal policies, based on the conditional average treatment effect estimators using instrumented DiD. Novel identification results for optimal policies under unmeasured confounding are established. Moreover, a range of estimators, including a Wald estimator, inverse probability weighting estimators, and semiparametric efficient and multiply robust estimators, are introduced. Theoretical guarantees for these multiply robust policy learning approaches are provided, including the cubic rate of convergence for parametric policies and valid statistical inference with flexible machine learning algorithms for nuisance parameter estimation. These methods are further extended to the panel data setup.

The majority of causal inference methods in the literature heavily depend on three standard causal assumptions to identify causal effects and optimal policies. While there has been progress in relaxing the consistency and unconfoundedness assumptions, addressing the violations of the positivity assumption has seen limited advancements.

In that context, this thesis presents a novel policy learning framework that does not rely on the positivity assumption, instead focusing on dynamic and stochastic policies that are practical for real-world applications. Incremental propensity score policies, which adjust propensity scores by individualized parameters, are proposed, requiring only the consistency and unconfoundedness assumptions. This approach enhances the concept of incremental intervention effects, adapting it to individualized treatment policy contexts, and employs semiparametric theory to develop efficient influence functions and debiased machine learning estimators. Methods to optimize policy by

maximizing the value function under specific constraints are also introduced.

Additionally, the optimal individualized treatment regime (ITR) learned from a source population may not generalize well to a target population due to covariate shifts. A transfer learning framework is proposed for ITR estimation in heterogeneous populations with right-censored survival data, which is common in clinical studies and motivated by medical applications. This framework characterizes the efficient influence function and proposes a doubly robust estimator for the targeted value function, accommodating a broad class of survival distribution functionals. For a pre-specified class of ITRs, a cubic rate of convergence for the estimated parameter indexing the optimal ITR is established. The use of cross-fitting procedures ensures the consistency and asymptotic normality of the proposed optimal value estimator, even with flexible machine learning methods for nuisance parameter estimation.

Résumé en français

La causalité est un concept fondamental en science et en philosophie. Dans un contexte où la collecte massive de données de grande complexité s'impose dans tous les domaines, les statistiques jouent un rôle crucial dans l'inférence des causes et des effets. Cette thèse explore des méthodes avancées d'inférence causale. Elle met l'accent sur l'apprentissage de politiques d'action ("politiques" dans la suite), les variables instrumentales (IV), et les approches de différences en différences (DiD).

Les méthodes IV et DiD sont utilisées par les chercheurs en épidémiologie, médecine, biostatistique, économétrie et sciences sociales quantitatives. Elles reposent sur des hypothèses restrictives, telles que, d'une part, l'exigence que l'IV n'ait aucun effet direct sur le résultat autre qu'à travers le traitement et, d'autre part, l'hypothèse de tendances parallèles en DiD, qui peut être violée en présence de confusion non mesurée.

Dans ce contexte, cette thèse propose une approche innovante de DiD instrumentalisée pour l'apprentissage de politiques. Cette combinaison permet de relâcher certaines des hypothèses clés des méthodes IV et DiD conventionnelles. Des résultats d'identification novateurs pour les politiques optimales en présence de confusion non mesurée sont établis, et une gamme d'estimateurs (de Wald ; par pondération inverse des probabilités ; semi-paramétriques efficaces et multiplement robustes) sont introduits. Des garanties théoriques multiplement robustes sont fournies, incluant le taux cubique de convergence pour les politiques paramétriques et une inférence statistique valide avec des algorithmes de machine learning (ML) flexibles pour l'estimation des paramètres de nuisance. Ces méthodes sont en outre étendues à la configuration de données de panel.

La majorité des méthodes d'inférence causale dans la littérature dépendent fortement de trois hypothèses causales standard pour identifier les effets causaux et les politiques optimales. Bien que des progrès aient été réalisés pour relâcher les hypothèses de consistance et de non-confusion, les avancées pour traiter les violations de l'hypothèse de positivité sont restées limitées.

Dans ce contexte, cette thèse présente un cadre novateur d'apprentissage des politiques qui ne repose pas sur l'hypothèse de positivité, se concentrant plutôt sur des politiques dynamiques et stochastiques pratiques pour des applications réelles. Des politiques de score de propension incrémentale, ajustant les scores de propension par des paramètres individualisés, sont proposées. Leur analyse ne met en jeu que les hypothèses de consistance et de non-confusion. Ce cadre améliore le concept d'effets d'intervention incrémentale, l'adaptant aux contextes de politique de traitement individualisée, et utilise la théorie semi-paramétrique pour développer des fonctions d'influence efficaces et des estimateurs ML dédiés. Des méthodes pour optimiser les politiques en maximisant la fonction de valeur sous des contraintes spécifiques sont également introduites.

De plus, le régime de traitement individualisé optimal (ITR) appris d'une population source peut ne pas se généraliser bien à une population cible en raison des décalages de covariables. Un cadre d'apprentissage par transfert est proposé pour l'estimation de l'ITR dans des populations hétérogènes avec des données de survie

censurées à droite, que l'on rencontre fréquemment dans les études cliniques. Un estimateur doublement robuste pour la fonction de valeur ciblée est proposé, qui accommode une large classe de fonctionnelles de distributions de survie. Pour une classe pré-spécifiée d'ITRs, un taux cubique de convergence pour le paramètre estimé indexant l'ITR optimal est établi. L'utilisation de procédures de cross-fitting (ajustement croisé) assure la consistance et la normalité asymptotique de l'estimateur de valeur optimal proposé, y compris lorsque l'on a recours à des méthodes ML flexibles pour estimer des paramètres de nuisance.

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my two advisors, Julie Josse and Antoine Chambaz. Julie guided me from being a good student to becoming an independent and passionate researcher. She has always been incredibly generous with her time, discussing issues and providing advice. Her enthusiasm and support have been unwavering in everything I do. Julie has invested tremendous energy and passion into our PreMeDICaL team at Inria, with broad interests and deep insights ranging from traditional statistics to modern machine learning research. She has consistently maintained a positive attitude toward collaborating with researchers from different fields, creating a comprehensive, dynamic, and unique research environment that has been invaluable to my development.

Antoine's love and passion have made my three years as a doctoral student truly enjoyable. During times of sadness and difficulty, his unconditional support reassured me. Despite the distance between Montpellier and Paris, and the fact that most of our interactions were via Zoom, his meticulous guidance, especially in statistical theory, has been immensely beneficial. Every time I visited Paris, his thoughtful care deeply touched me. These wonderful memories will stay with me forever.

I would also like to thank Alex Luedtke, Agathe Guilloux, Stefan Wager, Stijn Vansteelandt, Fabrizia Mealli, and Shu Yang for serving on my doctoral defense committee.

My heartfelt thanks also go to the wonderful faculty and friendly staff of the Université de Montpellier and Inria. Special thanks to Sylvie Barthelemy (IDESP, Université de Montpellier) and Marie-Hélène Gbaguidi (MAP5, Université Paris Cité) for their efficient and meticulous administrative support. Their work is incredibly important, yet often overlooked.

I am very fortunate to have collaborated with Shu Yang and Yifan Cui on research projects. Their professionalism and efficiency have been a great help, and they have provided valuable guidance and advice for my career development.

I am immensely grateful to Stijn Vansteelandt and Oliver Dukes for the opportunity to visit Ghent University for three months. They generously spent a lot of time discussing problems with me and provided invaluable advice on my career development and research direction. The causal inference research group is exceptionally strong, and I am sure that this experience will have a lasting impact on my future research. I also want to thank my many friendly friends: Kelly Van Lancker, Georgi Baklicharov, Edoardo Gervasoni, Zehao Su, Johan Steen, Muluneh Alene, Eline Anslot, and Wout Waterschoot. We have shared many wonderful memories together.

The past three years would not have been as fulfilling and exciting without my friends, colleagues and collaborators: Imke Mayer, Bénédicte Colnet, Paul Roussel, Margaux Zaffran, Maxime Fosset, Charlotte Voinot, Rémi Khellaf, Ahmed Boughdiri, Jeffrey Näf, Marie Felicia Beclin, Laura Fuentes Vicente, Claire-Marine Parodi, Tanguy Lefort, Camille Garcin, Aurélien Bellet, Pascal Demoly, Nicolas Molinari, Joseph Salmon, Alexander Reisach, Thi Thanh Yen Nguyen, Nicolas Gatulle, Bo Zhang, Jiwei Zhao.

I would also like to thank Nintendo and my Switch gaming console for bringing me pure and endless joy. Satoru Iwata, the fourth president of Nintendo and its most genius and affable leader, once said, "In my life, there is no such thing as wasted experiences." Such words are truly admirable. His thoughts and pursuits regarding games, management, and creativity have been a source of inspiration for me.

感谢我的母亲柳孟娟、父亲赵水贵的爱与支持。

Lastly, this PhD experience has brought many changes to my attitude towards life and way of thinking, while the world around us is also undergoing radical and profound transformations. I do not have the answers, so I will end with a quote from one of my favorite French philosophers, Louis Pierre Althusser [1]:

> Oui, nous sommes d'abord unis par cette institution qu'est le spectacle, mais plus profondément unis par les mêmes mythes, par les mêmes thèmes, qui nous gouvernent sans notre aveu, par la même idéologie spontanément vécue. Oui, bien qu'il soit par excellence celui des pauvres, comme en El Nost Milan, nous mangeons le même pain, avons les mêmes colères, les mêmes révoltes, les mêmes délires (au moins dans la mémoire où rôde sans cesse ce possible imminent), sinon le même accablement devant un temps que nulle Histoire ne meut. Oui, comme Mère Courage, nous avons la même guerre à la porte, et à deux doigts de nous, sinon en nous, le même horrible aveuglement, la même cendre dans les yeux, la même terre dans la bouche. Nous avons la même aube et la même nuit, nous frôlons les mêmes abîmes : notre inconscience. Nous partageons bien la même histoire, – et c'est par là que tout commence.

1. Quatrième partie. Le « piccolo », Bertolazzi et Brecht (Notes sur un théâtre matérialiste). Louis Althusser, dans *Pour Marx* (2005), pages 153 à 177.

TABLE OF CONTENTS

List of Figures

List of Tables

Part I

INTRODUCTION

CHAPTER 1

MODICUM OF CAUSAL INFERENCE AND CONTRIBUTIONS

As human beings, we intuitively grasp fundamental concepts of causal inference. We understand what a causal effect is, distinguish between association and causation, and use this knowledge to make decisions in our daily lives. These concepts are so ingrained that we often apply them unconsciously.

Aristotle stated in the *Posterior Analytics*, "We think we have knowledge of a thing only when we have grasped its cause." Later philosophers, such as Hume and Mill, also laid foundational work in the study of causality.

This thesis, however, is not rooted in philosophy but focuses on the formal mathematical language and statistical tools used in scientific studies and data analysis. Causal inference, in a formal sense, involves the assumptions, study designs, and estimation strategies that enable researchers to draw causal conclusions from data [Pearl, 2009, Imbens and Rubin, 2015].

The literature on causal inference suggests we distinguish between three types of questions:

Associational: For example, "How many people take paracetamol when they have a headache in France?"

Interventional (effects of causes): For instance, "If I have a headache, will taking paracetamol help?"

Counterfactual (causes of effects): Such as, "My headache has gone away. Is it because I took paracetamol or because I got enough rest?"

Classical statistics and modern machine learning algorithms primarily address the first type of question, focusing on associations and predictions. The recent advancements in machine learning and artificial intelligence have elevated the sophistication of associational inference.

However, association does not imply causation. To address the latter two questions, randomized controlled trials (RCTs) are the gold standard in statistical causal inference. Fisher's seminal 1935 book, *The Design of Experiments*, underscored the importance of randomization in experiments.

Even when RCTs are not feasible for ethical or practical reasons, the quality of observational studies is often evaluated based on how closely they approximate an RCT [Hernan and Robins, 2020]. Given the proven capability of many observational studies to infer causation and the increasing availability of big data (primarily observational), the study of causal effects using observational data and statistical methods is invaluable [Rosenbaum, 2002, Small, 2024].

In presenting our studies, we must address two key questions:

— What mathematical language or model should we use to study causality?
— Can we use observational studies to learn causal relationships? Specifically, what causal parameters can we identify and interpret?

The dominant perspective on causal inference in statistics is grounded in counterfactual states. This approach considers the potential outcomes that could manifest under different treatment conditions. Causal effects are defined as comparisons between these potential outcomes. For instance, the causal effect of a drug on systolic blood pressure one month after starting the drug (versus no exposure) is the comparison of systolic blood pressure measured under drug exposure with that measured without drug exposure. The challenge lies in the fact that we cannot observe both states simultaneously for the same individual [Gelman et al., 2021].

Another perspective on causal modeling involves a more fundamental structure: a causal structure that includes a probability model with additional information. Researchers refer to this as structure learning or causal discovery. To understand causal structures from observational data, we must grasp how causal and statistical models relate [Peters et al., 2017]. Reichenbach's common cause principle states that if two variables, X and Y, are statistically dependent, there exists a third variable, Z, that causally influences both and renders X and Y independent when conditioned on Z.

The rest of this chapter describes the two major mathematical languages designed to answer the causal questions.

1.1 Potential Outcomes

To introduce the fundamental concepts, we begin with the simplest setup where we observe the data (X, A, Y). Let $A \in \{0, 1\}$ denote the treatment assignment, X the pretreatment covariates, $Y(0)$ and $Y(1)$ the potential outcomes, and Y the observed outcome. The fundamental problem of causal inference is that $Y(1)$ and $Y(0)$ can never be observed at the same time, so the individual treatment effect $Y(1) - Y(0)$ is never known.

In many cases, the quantity of interest (causal estimand) is the average treatment effect (ATE),

$$E[Y(1) - Y(0)],$$

and its variants, such as the average treatment effect on the treated, or the conditional average treatment effect,

$$E[Y(1) - Y(0) \mid X].$$

To identify the ATE, three standard causal assumptions are required.

Assumption 1.1.1 (consistency)**.** It holds that $Y = Y(A)$.

Assumption 1.1.2 (positivity)**.** It holds almost surely that $0 < P(A \mid X) < 1$.

Assumption 1.1.3 (unconfoundedness)**.** For both $a \in \{0, 1\}$, $A \perp Y(a) \mid X$.

Assumption 1.1.1 is also known as the stable unit treatment value assumption, which requires there should be no multiple versions of the treatment and no interference between units. Assumption 1.1.2 says that each unit has a positive probability of receiving either treatment level. Assumption 1.1.3 states that there are no unmeasured confounders so that treatment assignment is as good as random conditional on the covariates X.

Under Assumptions 1.1.1 - 1.1.3, the ATE is identified by three common methods:

Outcome regression

$$E[Y(1) - Y(0)] = E\left[\mu(1, X) - \mu(0, X)\right],$$

where $\mu(A, X) = E\left[Y \mid A, X\right]$ almost surely.

Inverse probability weighting

$$E[Y(1) - Y(0)] = E\left[\frac{AY}{e(X)} - \frac{(1 - A)Y}{1 - e(X)}\right],$$

where $e(X) = P\left(A = 1 \mid X\right)$ almost surely.

Double robustness

$$E[Y(1) - Y(0)] = E\left[\mu(1, X) - \mu(0, X) + A\frac{Y - \mu(1, X)}{e(X)} - (1 - A)\frac{Y - \mu(0, X)}{1 - e(X)}\right].$$

The effectiveness of the outcome regression and inverse probability weighting methods hinges on the correct specification of the outcome and propensity score models respectively, while the doubly robust method combines both models and is still consistent even if one of the two models is misspecified. The three methods represent the main approaches to identifying causal estimands, and serves as the foundation of the advanced methods for more complex data structures in the subsequent parts.

1.2 Graphical Models

The formal graphical model was initially developed to describe associative relationship rather than causal relationship. However, humans naturally interpret causality using graphs, where an arrow from X to Y signifies that X causes Y. Before delving into causal interpretations, it is essential to briefly review how conditional dependence is represented in a graph, specifically through a Bayesian network.

Moreover, even if we are not directly interested in causal learning or discovery tasks, graphical models are still highly valuable for visually representing causal relationships in a straightforward manner. Additionally, conditional independence relationships can

be easily inferred from causal DAGs. With minimal assumptions, it is also possible to link graphs and counterfactuals using single-world intervention graphs [Richardson and Robins, 2013].

Let P denote a joint distribution over some set of random variables $X = (X_1, \ldots, X_p)$. The essence of the Bayesian network representation lies in a directed acyclic graph (DAG) $\mathcal{G}$, where the nodes represent the random variables in X. This graph $\mathcal{G}$ can be interpreted in two distinct but equivalent ways. Firstly, $\mathcal{G}$ serves as the framework for factorizing the joint distribution. Secondly, it encapsulates a set of conditional independence assumptions.

A Bayesian network is a pair $(\mathcal{G}, P)$ such that P factorizes over $\mathcal{G}$:

$$P(X_1, \ldots, X_p) = \prod_{i=1}^{p} P(X_i \mid Pa_{\mathcal{G}}(X_i)),$$

where $Pa_{\mathcal{G}}(X_i)$ denote the parents of node X_i in graph $\mathcal{G}$.

Next we consider the conditional independence assumptions. Let $\mathcal{I}(P)$ denote the set of conditional independence relationships of the form $X \perp Y \mid Z$ in P, where X, Y and Z can be multivariate. Let $\mathcal{I}_l(\mathcal{G})$ denote the local independence relationships in $\mathcal{G}$: $\mathcal{I}_l(\mathcal{G}) = \{X \perp \text{non-descendants of } X \mid Pa_{\mathcal{G}}(X)\}$. Then P factorizes over $\mathcal{G}$ if and only if $\mathcal{I}_l(\mathcal{G}) \subseteq \mathcal{I}(P)$ [Peters et al., 2017].

The equivalence can be understood through the concepts of an active path and an active vertex on a path:

— $X \to Z \to Y$ is called a chain. It is active if and only if Z is not included.
— $X \leftarrow Z \to Y$ is called a fork (common cause). It is active if and only if Z is not included.
— $X \to Z \leftarrow Y$ is called a collider (common effect). It is active if and only if Z is included.

Intuitively, a path is considered active if it transmits information or indicates dependence. Two variables, X and Y, might be connected by numerous paths within a graph $\mathcal{G}$, where these paths can be all active, some active, or none active. The variables X and Y are said to be d-separated by a set of variables Z if every path that connects X and Y is blocked by at least one variable in Z. Define $\mathcal{I}(\mathcal{G}) = \{X \perp Y \mid Z : X \text{ and } Y \text{ is d-separated by } Z\}$. Whenever P factorizes over $\mathcal{G}$, we have $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$. But note that the converse is not always true.

A causal Bayesian network aims to represent stable and autonomous physical mechanisms, enabling us to predict the effects of interventions that disrupt the natural course of events. This approach is more informative than purely probabilistic models because it incorporates causal knowledge, allowing us to answer more complex questions about causality [Peters et al., 2017].

1.3 Policy Learning

In previous sections, we have concentrated on methods for estimating causal effects in various statistical contexts. However, in many application areas, the primary goal of

causal analysis is not just to estimate treatment effects, but to inform decision-making. We aim to understand treatment effects so we can effectively prescribe treatments and allocate limited resources.

The task of learning optimal treatment assignment policies is closely related to—but distinct from—the task of estimating treatment heterogeneity. On the one hand, policy learning might seem simpler: we only need to determine whether to assign individuals to treatment or control groups, without requiring precise estimates of treatment effects. On the other hand, policy learning involves additional considerations that are not present when merely estimating treatment effects. Any policy we implement must be straightforward enough to be practically deployable, avoid discrimination based on protected characteristics, and not rely on manipulable features. We will discuss how to learn treatment assignment policies by directly optimizing a relevant welfare criterion.

A treatment assignment policy d is a mapping

$$d : \mathcal{X} \to \{0, 1\},$$

such that individuals with covariates x get treated if and only if $d(x) = 1$. Let $\mathcal{D}$ denote a pre-specified class of policies of interest, where each policy $d \in \mathcal{D}$ induces the value function defined by

$$V(d) = E[Y(d)] = E[Y(1)d(X) + Y(0)(1 - d(X))],$$

where $Y(d)$ is the potential outcome under the policy d. The optimal policy can be defined as

$$d = \arg\max_{d \in \mathcal{D}} V(d).$$

1.4 Contributions

Subsequently, this manuscript is divided into five parts. Part 2 combines two widely used natural experiments, Instrumental Variables (IV) and Difference-in-Differences (DiD), to learn optimal treatment assignment policies. Part 3 introduces incremental propensity score policies to handle positivity violations. Part 4 explores combining randomized trials and observational data to learn optimal treatment regimes that generalize well to a target population. All additional results and proofs are provided in Part 5 (Supplementary Material).

The contributions of each part, which are summarized below, have led to three articles:

— A Semiparametric Instrumented Difference-in-Differences Approach to Policy Learning, currently undergoing a major revision at *Biometrika*,

— Positivity-free Policy Learning with Observational Data, published in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR 238:1918-1926, 2024*, and selected as an *oral presentation*.

— Efficient and robust transfer learning of optimal individualized treatment regimes with right-censored survival data, rejected and resubmitted to *Journal of Machine Learning Research*.

Part 2 The instrumental variable (IV) and difference-in-differences (DiD) methods are both important tools widely used by empirical researchers in epidemiology, medicine, biostatistics, econometrics and quantitative social sciences. However, concerns often arise regarding the restrictive assumptions, for instance, requiring that the IV cannot have a direct causal effect on the outcome other than through the treatment, and the parallel trends assumption, which may be violated in the presence of unmeasured confounding.

Policy learning is pivotal across various domains, with the objective of learning the optimal treatment assignment policy. In this work, we combine the two natural experiments and propose an instrumented DiD approach to policy learning, relaxing some key assumptions of the conventional IV and DiD methods. To our knowledge, this is the first work to systematically study policy learning under the DiD setting.

— First, we propose the direct policy search approach to learn optimal policies, based on the conditional average treatment effect estimators using instrumented DiD.

— Second, we establish novel identification results of optimal policies for the instrumented DiD design subject to unmeasured confounding, without necessarily identifying the value function. A Wald estimator, novel inverse probability weighting (IPW) estimators, and a class of semiparametric efficient and multiply robust estimators are proposed.

— Third, we prove theoretical guarantees for the proposed multiply robust policy learning approaches. Specifically, $n^{-1/3}$ rate of convergence is established for the Euclidean parameter indexing parametric policies. And valid statistical inference results are achieved even when relying on flexible machine learning algorithms for nuisance parameters estimation.

— Fourth, we extend our proposed methods to the panel data setup.

Part 3 Most causal inference methods in the literature heavily rely on three standard causal assumptions 1.1.1 - 1.1.3 to identify causal effects and optimal policies. While there has been significant progress in relaxing the consistency and unconfoundedness assumptions, advancements addressing violations of the positivity assumption remain limited.

This work introduces a novel policy learning framework that does not depend on the positivity assumption, focusing instead on dynamic and stochastic policies that are practical for many real-world applications. We propose incremental propensity score policies that adjust propensity scores by an individualized parameter, requiring only the consistency and unconfoundedness assumptions.

Our approach enhances the concept of incremental intervention effects, adapting it to individualized treatment policy contexts. We employ semiparametric theory to characterize the efficient influence function and propose debiased machine learning estimators. Building on these efficient off-policy evaluation results, we introduce methods to learn the optimal policy by maximizing the value function, potentially under application-specific constraints.

Part 4 The optimal individualized treatment regime (ITR) learned from a source population, due to covariate shift, may not generalize well to the target population that we aim to apply the ITR on. We propose a transfer learning framework, where covariate information from the target population is available, for ITR estimation with heterogeneous populations and right-censored survival data, which is common in clinical studies and motivated by a medical application.

We characterize the efficient influence function (EIF) and propose a doubly robust estimator of the targeted value function, which accommodates a broad class of functionals of survival distributions. For a pre-specified class of ITRs, we establish the cubic rate of convergence for the estimated parameter indexing the optimal ITR. Based on the Neyman orthogonality of the EIF, we also propose a cross-fitting procedure and show that the proposed optimal value estimator is consistent and asymptotically normal with flexible machine learning methods for nuisance parameter estimation.

1.5 Other Contributions

Beyond the aforementioned methodological contributions, some applied research has also been conducted, resulting in the following studies:

Learning, Evaluating and Analysing An Individualized Treatment Rule This project is driven by an application focused on early intervention in intensive care units (ICU).

We present a statistical framework for the learning, evaluation, and analysis of individualized decision rules, inspired by the application of early interventions (e.g., lifesaving blood products) in the ICU, as demonstrated in our TraumaBase® data analysis. Severe trauma remains a leading cause of mortality. When faced with hemorrhage, the primary goal in managing ICU patients with severe trauma is to swiftly and effectively control bleeding, thereby improving survival rates.

We propose a super learning method that recovers an optimal treatment assignment rule from a set of possible options for each patient based on their individual characteristics, even in the presence of missing covariate information. We discuss the causal interpretation and underlying assumptions of our approach. To facilitate and inform medical practice for clinicians using flexible machine learning algorithms, we evaluate the learned rule using a novel algorithm-agnostic variable importance measure and introduce a new restricted score test for cases with degenerate efficient influence functions.

The proposed methods are validated through extensive simulations. Additionally, we have developed an `R` package, `missSuperLearner`, which implements the super learning algorithm that handles missing data.

CRAN Task View: Causal Inference This review (https://cran.r-project.org/view=CausalInference) aims to provide guidance on the `R` packages that are relevant for causal inference tasks.

TRADUCTION EN FRANÇAIS

En tant qu'êtres humains, nous comprenons intuitivement les concepts fondamentaux de l'inférence causale. Nous comprenons ce qu'est un effet causal, distinguons entre association et causalité, et utilisons ces connaissances pour prendre des décisions dans notre vie quotidienne. Ces concepts sont si ancrés que nous les appliquons souvent de manière inconsciente.

Aristote a déclaré dans les *Seconds Analytiques*, "Nous pensons avoir connaissance d'une chose seulement lorsque nous en avons saisi la cause." Des philosophes ultérieurs, tels que Hume et Mill, ont également jeté les bases de l'étude de la causalité.

Cette thèse, cependant, n'est pas enracinée dans la philosophie mais se concentre sur le langage mathématique formel et les outils statistiques utilisés dans les études scientifiques et l'analyse de données. L'inférence causale, au sens formel, implique les hypothèses, les conceptions d'études et les stratégies d'estimation qui permettent aux chercheurs de tirer des conclusions causales à partir des données.

La littérature sur l'inférence causale suggère de distinguer trois types de questions :

Associatives : Par exemple, "Combien de personnes prennent du paracétamol lorsqu'elles ont un mal de tête en France ?"

Interventionnelles (effets des causes) : Par exemple, "Si j'ai un mal de tête, prendre du paracétamol m'aidera-t-il ?"

Contrefactuelles (causes des effets) : Par exemple, "Mon mal de tête a disparu. Est-ce parce que j'ai pris du paracétamol ou parce que je me suis suffisamment reposé ?"

Les statistiques classiques et les algorithmes modernes d'apprentissage automatique abordent principalement le premier type de question, se concentrant sur les associations et les prédictions. Les récentes avancées en apprentissage automatique et en intelligence artificielle ont élevé la sophistication de l'inférence associative.

Cependant, l'association n'implique pas la causalité. Pour aborder les deux dernières questions, les essais contrôlés randomisés (ECR) sont la norme en matière d'inférence causale statistique. Le livre séminal de Fisher de 1935, *The Design of Experiments*, a souligné l'importance de la randomisation dans les expériences.

Même lorsque les ECR ne sont pas réalisables pour des raisons éthiques ou pratiques, la qualité des études observationnelles est souvent évaluée en fonction de leur proximité avec un ECR. Étant donné la capacité prouvée de nombreuses études observationnelles à inférer la causalité et la disponibilité croissante de jeux massifs de données (principalement observationnelles), l'étude des effets causaux à l'aide de données observationnelles et de méthodes statistiques est inestimable.

En présentant nos études, nous devons aborder deux questions clés :

— Quel langage mathématique ou modèle devons-nous utiliser pour étudier la causalité ?
— Pouvons-nous utiliser des études observationnelles pour apprendre les relations causales ? Plus précisément, quels paramètres causaux pouvons-nous identifier et interpréter ?

La perspective dominante sur l'inférence causale en statistiques repose sur les états contrefactuels. Cette approche considère les résultats potentiels qui pourraient se manifester dans différentes conditions de traitement. Les effets causaux sont définis comme des comparaisons entre ces résultats potentiels. Par exemple, l'effet causal d'un médicament sur la pression artérielle systolique un mois après le début du traitement (par rapport à aucune exposition) est la comparaison de la pression artérielle systolique mesurée sous exposition au médicament avec celle mesurée sans exposition au médicament. Le défi réside dans le fait que nous ne pouvons pas observer les deux états simultanément pour un même individu.

Une autre perspective sur la modélisation causale implique une structure plus fondamentale : une structure causale qui inclut un modèle probabiliste avec des informations supplémentaires. Les chercheurs se réfèrent à cela comme apprentissage de structure ou découverte causale. Pour comprendre les structures causales à partir de données observationnelles, nous devons saisir comment les modèles causaux et statistiques sont liés. Le principe de la cause commune de Reichenbach stipule que si deux variables, X et Y, sont statistiquement dépendantes, il existe une troisième variable, Z, qui influence causalement les deux et rend X et Y indépendants lorsqu'on conditionne sur Z.

Contributions

Ce manuscrit est divisé en cinq parties. La Partie 2 combine deux expériences naturelles largement utilisées, les Variables Instrumentales (IV) et les Différences en Différences (DiD), pour apprendre des politiques d'attribution de traitements optimales. La Partie 3 introduit des politiques de score de propension incrémentales pour gérer les violations de positivité. La Partie 4 explore la combinaison d'essais randomisés et de données observationnelles pour apprendre des régimes de traitement optimaux qui se généralisent bien à une population cible. Tous les résultats supplémentaires et les preuves sont fournis dans la Partie 5 (Matériel Supplémentaire).

Les contributions de chaque partie, résumées ci-dessous, ont conduit à trois articles :

— A Semiparametric Instrumented Difference-in-Differences Approach to Policy Learning, actuellement en révision majeure chez *Biometrika*,
— Positivity-free Policy Learning with Observational Data, publié dans les *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR 238 :1918-1926, 2024*, et sélectionné pour une *présentation orale*,
— Efficient and robust transfer learning of optimal individualized treatment regimes with right-censored survival data, rejeté et resoumis au *Journal of Machine Learning Research*.

Partie 2 Les méthodes de variable instrumentale (IV) et de différences en différences (DiD) sont toutes deux des outils importants largement utilisés par les chercheurs empiriques en épidémiologie, médecine, biostatistique, économétrie et sciences sociales quantitatives. Cependant, des préoccupations surgissent souvent concernant les hypothèses restrictives, par exemple, celle requérant que l'IV ne puisse avoir d'effet causal direct sur le résultat autre que par le biais du traitement, et l'hypothèse de tendances parallèles peut être violée en présence de confusion non mesurée.

L'apprentissage de politiques est essentiel dans divers domaines, avec l'objectif d'apprendre la politique d'attribution de traitements optimale. Dans ce travail, nous combinons les deux expériences naturelles et proposons une approche instrumentée DiD pour l'apprentissage de politiques, en assouplissant certaines hypothèses clés des méthodes IV et DiD conventionnelles. À notre connaissance, c'est la première étude systématique de l'apprentissage de politiques dans le cadre DiD.

— Nous proposons d'abord une approche de recherche directe de politiques pour apprendre des politiques optimales, fondée sur les estimateurs de l'effet moyen conditionnel du traitement utilisant DiD instrumenté.
— Ensuite, nous établissons des résultats d'identification novateurs des politiques optimales pour la conception instrumentée DiD sous confusion non mesurée, sans nécessairement identifier la fonction de valeur. Un estimateur de Wald, des estimateurs IPW (Inverse Probability Weighting) novateurs, et une classe d'estimateurs semi-paramétriques efficaces et multiplement robustes sont proposés.
— Troisièmement, nous prouvons des garanties théoriques pour les approches d'apprentissage de politiques multiplement robustes proposées. Plus précisément, un taux de convergence de $n^{-1/3}$ est établi pour le paramètre euclidien indexant les politiques paramétriques. Des résultats d'inférence statistique valides sont aussi obtenus, qui sont valables y compris lorsque des algorithmes d'apprentissage automatique flexibles sont mis en œuvre pour l'estimation des paramètres de nuisance.
— Enfin, nous étendons nos méthodes proposées à la configuration de données de panel.

Partie 3 La plupart des méthodes d'inférence causale dans la littérature reposent fortement sur trois hypothèses causales standard pour identifier les effets causaux et les politiques optimales. Bien que des progrès significatifs aient été réalisés pour assouplir les hypothèses de consistance et d'absence de confusion, les avancées concernant les violations de l'hypothèse de positivité restent limitées.

Ce travail introduit un nouveau cadre d'apprentissage de politiques qui ne dépend pas de l'hypothèse de positivité, en se concentrant plutôt sur des politiques dynamiques et stochastiques qui sont pratiques pour de nombreuses applications réelles. Nous proposons des politiques de score de propension incrémentales qui ajustent les scores de propension par un paramètre individualisé, ne nécessitant que les hypothèses de consistance et d'absence de confusion.

Notre approche améliore le concept des effets d'intervention incrémentale, en l'adaptant au contexte de politique de traitement individualisée. Nous utilisons la théorie semi-paramétrique pour caractériser la fonction d'influence efficace et proposons des estimateurs d'apprentissage automatique débiaisés. En nous appuyant sur ces résultats efficaces d'évaluation hors politique, nous introduisons des méthodes pour apprendre la politique optimale en maximisant la fonction de valeur, potentiellement sous des contraintes spécifiques à l'application.

Partie 4 Le régime de traitement individualisé optimal (ITR) appris à partir d'une population source peut, en raison du changement de la loi des covariables, ne pas bien se généraliser à une population cible à laquelle nous visons à appliquer l'ITR. Nous proposons un cadre d'apprentissage par transfert, où les informations de covariables de la population cible sont disponibles, pour l'estimation de l'ITR avec des populations hétérogènes et des données de survie censurées à droite, ce qui est courant dans les études cliniques et motivé par notre application médicale.

Nous caractérisons la fonction d'influence efficace (EIF) et proposons un estimateur doublement robuste de la fonction de valeur ciblée, qui accueille une large classe de fonctionnels de distributions de survie. Pour une classe présélectionnée d'ITR, nous établissons un taux cubique de convergence pour le paramètre estimé indexant l'ITR optimal. Basé sur l'orthogonalité de Neyman de l'EIF, nous proposons également une procédure de cross-fitting (apprentissage croisé) et montrons que l'estimateur de valeur optimale proposé est cohérent et asymptotiquement normal quand bien même des méthodes d'apprentissage automatique flexibles sont mises en œuvre pour estimer des paramètres de nuisance.

Part II

A SEMIPARAMETRIC INSTRUMENTED DIFFERENCE-IN-DIFFERENCES APPROACH TO POLICY LEARNING

Recently, there has been a surge in methodological development for the difference-in-differences (DiD) approach to evaluate causal effects. Standard methods in the literature rely on the parallel trends assumption to identify the average treatment effect on the treated. However, the parallel trends assumption may be violated in the presence of unmeasured confounding, and the average treatment effect on the treated may not be useful in learning a treatment assignment policy for the entire population. In this article, we propose a general instrumented DiD approach for learning the optimal treatment policy. Specifically, we establish identification results using a binary instrumental variable (IV) when the parallel trends assumption fails to hold. Additionally, we construct a Wald estimator, novel inverse probability weighting (IPW) estimators, and a class of semiparametric efficient and multiply robust estimators, with theoretical guarantees on consistency and asymptotic normality, even when relying on flexible machine learning algorithms for nuisance parameters estimation. Furthermore, we extend the instrumented DiD to the panel data setting. We evaluate our methods in extensive simulations and a real data application. [1]

1. co-authored with Yifan Cui (Zhejiang University), currently undergoing a major revision at *Biometrika*.

CHAPTER 2

INSTRUMENTAL VARIABLE AND DIFFERENCE-IN-DIFFERENCES

2.1 Introduction

Data-driven individualized decision making has received increasing interests in many fields, such as precision medicine [Luedtke and van der Laan, 2016b, Tsiatis et al., 2019], econometrics and quantitative social sciences [Imai and van Dyk, 2004, Athey and Wager, 2021], computer science and operations research [Shi et al., 2022, Kallus et al., 2022]. The common goal is to learn optimal treatment assignment policies (also known as regimes, rules or plans) which map individual characteristics to treatment assignments so as to optimize some functional of the counterfactual outcome distributions, leveraging observational data where causal effects can be identified under various strategies and assumptions.

Popular existing methods in the statistical and machine learning literature include model-based approaches such as Q-learning [Watkins and Dayan, 1992, Murphy, 2003, Linn et al., 2017], A-learning [Robins et al., 2000, Shi et al., 2018], and direct model-free policy search approaches [Zhang et al., 2012a, Zhao et al., 2012]. Recent advances of policy learning have also considered a variety of data structures, optimization objectives, criteria or constraints, such as survival and longitudinal data [Goldberg and Kosorok, 2012, Ertefaie and Strawderman, 2018, Zhao et al., 2023], networks [Viviano, 2019, Sherman et al., 2020], distributional robustness [Mo et al., 2021, Sahoo et al., 2022], budget, fairness, or interpretability constraints [Luedtke and van der Laan, 2016a, Fang et al., 2022], among others [Luedtke and Chambaz, 2020, Hadad et al., 2021, Nie et al., 2021, Hu et al., 2022, Jin et al., 2023].

With few exceptions, most methods in prior work rely on the pivotal assumption that there is no unmeasured confounding. This is a key threat to credible causal inference in observational studies, and may lead to suboptimal policies, because this assumption is impossible to verify or test in practice. An ad hoc work-around commonly adopted by practitioners is to collect and appropriately adjust for a large number of covariates, which still lacks theoretical guarantee and seems likely to be error-prone. To address this limitation, there has been recent progress made in several directions. Kallus and Zhou [2018] propose to minimize the worst-case regret of a policy under a marginal sensitivity model for the unmeasured confounding. Zhang

et al. [2021] utilize a randomization test to rank by a partial order and select treatment rules within a given finite collection. While partial identification results provide certain improvement, the performance of such a learned policy may still be suboptimal. Qi et al. [2023] build on the semiparametric proximal causal inference framework introduced by Cui et al. [2023b] to establish point identification results on different policy classes and accordingly propose several classification-based approaches; but this framework requires the analyst to correctly classify the measured covariates into three types of proxies, and it may be difficult to estimate the confounding bridge functions.

Instrumental variable methods are widely used to handle unmeasured confounding in observational studies or randomized trials with non-compliance. The core requirements for a pretreatment variable to be a valid IV are: (i) it is associated with the treatment; (ii) it is independent of all unmeasured confounders; (iii) it does not have a direct causal effect on the outcome other than through the treatment. Along with the seminal work of Imbens and Angrist [1994], Angrist et al. [1996], extensive development has been made in using the IV to estimate the local average treatment effect [Tan, 2006, Ogburn et al., 2015], defined as the average treatment effect for the complier subgroup who would always comply with their treatment assignments. Since the complier subgroup is unknown and may have systematically different characteristics from the population, the population (conditional) average treatment effect is arguably the causal parameter of primary interest in most studies [Hernán and Robins, 2006, Aronow and Carnegie, 2013], especially for policy learning. More recently, Pu and Zhang [2021] consider a partial identification approach to optimal treatment rule estimation; and Wang and Tchetgen Tchetgen [2018] formally establish point identification of the population average treatment effect under alternative no-interaction assumptions, upon which Cui and Tchetgen Tchetgen [2021] propose various IV methods for estimating optimal treatment regimes. It is notable that all of these IV methods in the literature only consider the setting with a single time point, with the only exception of Xu et al. [2023], where the authors propose an IV approach to off-policy evaluation in confounded Markov decision processes with infinite horizons.

There has always been interest in exploiting the longitudinal structure common in datasets such as electronic health records and medical claims in epidemiology and biomedicine [Robins et al., 2000], as well as cross-sectional or panel data in program evaluations, economic censuses, and surveys [Athey and Imbens, 2017]. DiD methods have been an important tool widely used by empirical researchers [Card and Krueger, 1994]. The key identification assumption of DiD is that the trend in outcome of the control group over time is informative about what the trend would have been for the treatment group in the absence of the treatment. Specifically, under the standard (conditional) parallel trends assumption, which states that the (conditional) expected trends in the potential outcomes of the two groups in the absence of the treatment are identical, the average treatment effect on the treated can be identified [Abadie, 2005, Sant'Anna and Zhao, 2020]; we refer interested readers to Lechner et al. [2011] and Roth et al. [2023] for detailed reviews. However, concerns often arise that the parallel trends assumption may be violated due to unmeasured confounding. Athey and Imbens [2006] develop a new changes-in-changes model that relates outcomes to an individual's group, time, and unobservable characteristics; and various recent

extensions for DiD include partial identification [Ye et al., 2020], sensitivity analysis [Keele et al., 2019] and negative control [Sofer et al., 2016], among others [Dukes et al., 2022, Park and Tchetgen, 2023]. Moreover, DiD methods focus on the identification and estimation of the average treatment effect on the treated, which limits its application in policy learning since the treated cannot represent the population. To the best of our knowledge, this is the first work to systematically study policy learning under the DiD setting.

In this article, we combine the two natural experiments and propose an instrumented DiD approach to policy learning when the parallel trends assumption fails to hold in the presence of unmeasured confounding. Specifically, we adapt and extend the recent progress in Ye et al. [2022] and Vo et al. [2022], relaxing some key assumptions of the conventional IV and DiD methods. We allow for the violation of the parallel trends assumption by leveraging an IV which has no direct effect on the the trend in outcome, and does not modify the average treatment effect. Notably, this exogenous variable is not necessarily a valid instrument for the conventional treatment-outcome association, since we allow it to have a direct effect on the outcome not just through the treatment at each time point.

The contributions of this article are summarized as follows. First, we propose the direct policy search approach to learn optimal treatment assignment policies, based on the conditional average treatment effect estimators using instrumented DiD. This approach essentially allows us to learn the optimal policy that maximizes the estimated value within a restricted policy class. Second, we establish novel identification results of optimal policies for the instrumented DiD design subject to unmeasured confounding. The new results give rise to new inverse probability weighting estimators of optimal policies without necessarily identifying the value function for a given policy. Another interesting progress is also made towards identifying optimal policies without necessarily using the subjects' realized treatment values. In summary, we construct a Wald estimator and novel inverse probability weighting estimators. A class of semiparametric efficient and multiply robust estimators is also proposed, which is consistent provided that a subset of several posited models indexing the observed data distribution is correctly specified. Third, we prove theoretical guarantees for the proposed multiply robust policy learning approaches. Specifically, we consider both parametric models and flexible data-adaptive machine learning algorithms with the cross-fitting procedure to estimate the nuisance parameters, to draw valid inferences under mild regularity conditions and certain rate of convergence conditions. In particular, we consider a restricted policy class indexed by an Euclidean parameter η and establish the $n^{-1/3}$ convergence rate of $\hat{\eta}$, even though its resultant limiting distribution is not standard. Fourth, we extend our proposed methods to the panel data setup. We establish identification of the conditional average treatment effect under alternative assumptions and provide the direct policy search approaches for panel data. The theoretical results for panel data can be similarly derived.

The rest of this article is organized as follows. In Section 2.2, we introduce the statistical framework of instrumental variable, DiD and policy learning. Section 2.3 develops our main methodology of learning the optimal policy using the instrumented DiD. Semiparametric efficiency results and multiply robust estimators are presented in

Section 3.1. Section 3.2 establishes the asymptotic properties of the proposed estimators. Extensive simulations are reported in Section 4.1 to demonstrate the proposed methods, followed by a real data application in Section 4.2. Next, we consider the extension of our methods to panel data in Section 5.1. The article concludes in Section 5.2 with a discussion of some remarks and future work. All proofs and additional results are provided in the Supplementary Material.

2.2 Statistical framework

We first introduce some notation. Let X denote the p-dimensional vector of covariates that belongs to a covariate space $\mathcal{X} \subset \mathbb{R}^p$, $A \in \mathcal{A} = \{0, 1\}$ denote the binary treatment, $Y \in \mathbb{R}$ denote the outcome of interest, and $T \in \mathcal{T} = \{0, 1\}$ denote the time period. Suppose that $U = (U_0, U_1)$ is an unmeasured confounder of the effect of A on Y, and $Z \in \{0, 1\}$ is a binary instrumental variable; the observed data are $O = (X, A, Y, T, Z)$. We assume that the random samples $(O_1, \ldots, O_n)$ collected at the two time periods are independent and identically distributed (i.i.d.) observations of $O \sim P_0$, and there is no overlap between individuals in these two time periods. This setup is commonly known as the repeated cross-sectional data. Extension to panel data setting is studied in Section 5.1.

We use the potential outcomes framework [Neyman, 1923, Rubin, 1974] to define causal effects. Let $A_t(z)$ denote the potential exposure at time t if the instrument were set to level z, $Y_t(a)$ denote the potential outcome at time t if the exposure were set to level a and the instrument would take the same value it actually had, and $Y_t(z, a)$ denote the potential outcome at time t had the instrument and exposure been set to z, a respectively.

Without loss of generality, we assume that larger values of Y are more desirable. Our aim is to identify and estimate an policy $d : \mathcal{X} \to \mathcal{A}$, that maximizes the expected potential outcome in a counterfactual world had this policy been implemented on the population. The optimal policy at time t is given by $d_{\text{opt},t}(x) = I\{\tau_t(x) > 0\}$, where $\tau_t(x) = E[Y_t(1) - Y_t(0) \mid X = x]$ is the conditional average treatment effect (CATE) at time t.

Let $Y_t(d) = d(X)Y_t(1) + (1 - d(X))Y_t(0)$ denote the potential outcome under a hypothetical intervention that assigns treatment according to policy d. The value function of a policy d at time t is defined as $V_t(d) = E[Y_t(d)]$. Let $\mathcal{D}$ be the class of candidate policies of primary interest. The optimal policy can be obtained by directly maximizing the value function:

$$d_{\text{opt},t} = \arg\max_{d \in \mathcal{D}} V_t(d) = \arg\max_{d \in \mathcal{D}} E[\tau_t(X)d(X)]. \tag{2.1}$$

Throughout this article, we assume that the stable treatment effect over time assumption holds, which says that the CATE does not vary over time, and thus ensures that the optimal policy remains the same between the two time periods. The subscript t is omitted when it is clear from the context.

Remark 2.2.1. Our proposed instrumented DiD methodology can also be readily formulated in the weighted classification perspective. Pioneered by Zhang et al. [2012a], this perspective has been widely used in the biostatistics and precision medicine literature, and enjoys certain robustness empirically. Specifically, the above maximization problem (2.1) can be transformed into the following equivalent weighted classification problem:

$$d_{\text{opt}}(x) = \arg\max_{d\in\mathcal{D}} E[WI\{A = d(X)\}], \tag{2.2}$$

where W is regarded as a weight that is motivated by standard outcome regression, inverse probability weighting and doubly robust methods. Many robust classification methods and off-the-shelf implementations can be utilized.

2.3 Instrumented difference-in-differences

In this section, we introduce a general instrumented DiD framework for policy learning under endogeneity, and provide novel identification results. Let $\pi(t, z, x) = Pr(T = t, Z = z \mid X = x)$, and for any random variable $C \in \{A, Y\}$, we define $\mu_C(t, z, x) = E[C \mid T = t, Z = z, X = x]$, $\delta_C(x) = \mu_C(1, 1, x) - \mu_C(0, 1, x) - \mu_C(1, 0, x) + \mu_C(0, 0, x)$. We make the following identification assumptions.

Assumption 2.3.1 (Consistency). $A = A_T(Z)$ and $Y = Y_T(A)$.

Assumption 2.3.2 (Positivity). $c_1 < \pi(t, z, x) < 1 - c_1$ for some $0 < c_1 < 1/2$.

Assumption 2.3.3 (Random sampling). $T \perp \{A_t(z), Y_t(a) : t = 0, 1, z = 0, 1, a = 0, 1\} \mid X, Z$.

Assumption 2.3.4 (Stable treatment effect over time). $E[Y_0(1) - Y_0(0) \mid X] = E[Y_1(1) - Y_1(0) \mid X]$.

Assumption 2.3.1 is also known as the stable unit treatment value assumption, which states that there is no interference between subjects and no multiple versions of the instrument and treatment. Assumption 2.3.2 ensures the same support of X for each (T, Z) level. Assumption 2.3.3 is commonly assumed for repeated cross-sectional data [Abadie, 2005]. Assumption 2.3.4 requires that the CATE $\tau(x)$ does not vary over time, and thus ensures that the optimal policy remains the same between the two time periods.

Assumption 2.3.5 (Trend relevance). $E[A_1(1) - A_0(1) \mid Z = 1, X] \neq E[A_1(0) - A_0(0) \mid Z = 0, X]$.

Assumption 2.3.6 (Independence & exclusion restriction). $Z \perp \{A_t(1), A_t(0), Y_t(1) - Y_t(0), Y_1(0) - Y_0(0) : t = 0, 1\} \mid X$.

Assumption 2.3.7 (No unmeasured common effect modifier). $Cov\{A_t(1) - A_t(0), Y_t(1) - Y_t(0) \mid X\} = 0$ for $t = 0, 1$.

Assumption 2.3.5 and 2.3.6 are parallel to the core assumptions in the standard IV literature. Directed acyclic graphs illustrating the causal structure are provided in Section A.1 of the Supplementary Material. Assumption 2.3.5 states that the IV affects the trend in treatment. Assumption 2.3.6 requires that the IV is unconfounded, has

no direct effect on the trend in outcome, and does not modify the treatment effect. This exogenous variable is not necessarily a valid instrument for the conventional treatment-outcome association, since we allow it to have a direct effect on the outcome not just through the treatment at each time point. Assumption 2.3.7 essentially states that there is no common effect modifier by an unmeasured confounder, of the additive effect of treatment on the outcome, and the additive effect of the IV on treatment. It has been studied in Cui and Tchetgen Tchetgen [2021], and relax certain no additive interaction assumptions in Wang and Tchetgen Tchetgen [2018]. We refer interested readers to Ye et al. [2022] for detailed discussion and concrete examples of an IV for DiD. Now we present our first identification result under the above assumptions.

Theorem 2.3.8. *Under Assumptions 2.3.1-2.3.7, the optimal policy is nonparametrically identified by*

$$\arg\max_{d\in\mathcal{D}} E\left[\frac{\delta_Y(X)}{\delta_A(X)}d(X)\right]. \tag{2.3}$$

Theorem 2.3.8 combines the Wald estimator for CATE and the direct policy search approach in Equation (2.1). Similarly, the IPW estimator proposed by Ye et al. [2022] can also be used to learn the optimal policy. Semiparametric efficient and multiply robust estimators are presented in Section 3.1. Next we propose our novel identification results, which also serves as basis for the estimators proposed in Section 3.1.

Theorem 2.3.9. *Under Assumptions 2.3.1-2.3.7, the optimal policy is nonparametrically identified by*

$$\arg\max_{d\in\mathcal{D}} E\left[\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right]. \tag{2.4}$$

Theorem 2.3.9 extends prior identification of CATE, and proposes a novel IPW estimator of the optimal policy without necessarily identifying the value function. Semiparametric efficiency results based on (2.4) are given in Section A.6 and A.7 of the Supplementary Material.

Theorem 2.3.10. *Under Assumptions 2.3.1-2.3.7, the optimal policy is nonparametrically identified by*

$$\arg\max_{d\in\mathcal{D}} E\left[\frac{(2T-1)YI\{Z=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right]. \tag{2.5}$$

Theorem 2.3.10 essentially proves that we can identify the optimal policy without necessarily using the subjects' realized treatment values, for instance when $\delta_A(X)$ is known a priori, or when a separate sample with data on (A,X,T,Z) is available to estimate $\delta_A(X)$. To conclude this section, we propose the following estimators for

optimal policies:

$$\hat{d}_{\text{Wald}} = \arg\max_{d\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{\delta}_Y(X_i)}{\hat{\delta}_A(X_i)} d(X_i),$$
$$\hat{d}_{\text{IPW1}} = \arg\max_{d\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \frac{(2Z_i-1)(2T_i-1)(2A_i-1)Y_i I\{A_i = d(X_i)\}}{\hat{\pi}(T_i,Z_i,X_i)\hat{\delta}_A(X_i)},$$
$$\hat{d}_{\text{IPW2}} = \arg\max_{d\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} \frac{(2T_i-1)Y_i I\{Z_i = d(X_i)\}}{\hat{\pi}(T_i,Z_i,X_i)\hat{\delta}_A(X_i)},$$

where $\hat{\delta}_Y$, $\hat{\delta}_A$ and $\hat{\pi}$ are estimated by parametric models or machine learning algorithms. Our simulation studies in Section 4.1 empirically shows comparable performance of the IPW estimators (2.4) and (2.5).

Remark 2.3.11. Similarly, classification-based estimators based on Theorem 2.3.9 and 2.3.10 can be proposed:

$$\arg\max_{d\in\mathcal{D}} E[\tilde{W}_1 I\{A = d(X)\}], \quad \arg\max_{d\in\mathcal{D}} E[\tilde{W}_2 I\{Z = d(X)\}], \tag{2.6}$$

respectively, where the weights are given by

$$\tilde{W}_1 = \frac{(2Z-1)(2T-1)(2A-1)Y}{\pi(T,Z,X)\delta_A(X)}, \quad \tilde{W}_2 = \frac{(2T-1)Y}{\pi(T,Z,X)\delta_A(X)}.$$

The Fisher consistency, excess risk bound and universal consistency of the estimated policy can also be established [Zhao et al., 2012].

CHAPTER 3

SEMIPARAMETRIC EFFICIENCY AND INFERENCE

3.1 Semiparametric efficiency and multiply robust estimators

In this section, we use semiparametric theory and propose multiply robust estimators. The Wald and the IPW approaches require the corresponding models to be correctly specified. Hence, methods that are robust against model misspecification are highly desired, where consistency is guaranteed when a subset of several posited models indexing the observed data distribution is correctly specified.

We consider the (uncentered) efficient influence function:

$$\Delta(O) = \frac{\delta_Y(X)}{\delta_A(X)} + \frac{(2Z-1)(2T-1)}{\pi(T,Z,X)\delta_A(X)}\left\{Y - \mu_Y(T,Z,X) - \frac{\delta_Y(X)}{\delta_A(X)}(A - \mu_A(T,Z,X))\right\},$$

which has been proposed in Ye et al. [2022]. Therefore, the optimal policy is identified by $\arg\max_{\mathcal{D}} E\left[\Delta(X)d(X)\right]$. Moreover, in light of the optimization tasks formulated in (2.6), we propose the following two choices of statistic:

$$W_1 = \frac{(2A-1)\delta_Y(X)}{\delta_A(X)} + \frac{(2A-1)(2Z-1)(2T-1)}{\pi(T,Z,X)\delta_A(X)}\left\{Y - \mu_Y(T,Z,X) - \frac{\delta_Y(X)}{\delta_A(X)}(A - \mu_A(T,Z,X))\right\},$$

and

$$W_2 = \frac{(2Z-1)\delta_Y(X)}{\delta_A(X)} + \frac{2T-1}{\pi(T,Z,X)\delta_A(X)}\left\{Y - \mu_Y(T,Z,X) - \frac{\delta_Y(X)}{\delta_A(X)}(A - \mu_A(T,Z,X))\right\},$$

which also enjoy the multiply robustness property.

First, we consider positing parametric models. Let $\mu_A(t,z,x;\alpha)$, $\mu_Y(t,z,x;\beta)$ and $\pi(t,z,x;\theta)$ denote the posited models. $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\theta}$ can be estimated by maximum likelihood estimation. In Theorem 3.1.1, we show the multiple robustness in the sense of maximizing the objective function (or minimizing the weighted classification error) in the union model of the following models:
$\mathcal{M}_1$: models for $\pi(t,z,x)$ and $\delta_A(x)$ are correct;

$\mathcal{M}_2$: models for $\pi(t,z,x)$ and $\delta_Y(x)/\delta_A(x)$ are correct;
$\mathcal{M}_3$: models for $\delta_Y(x)/\delta_A(x)$ and $\mu_C(0,0,x)$, $\mu_C(1,0,x)$, $\mu_C(0,1,x)$ for $C \in \{A, Y\}$ are correct.

Theorem 3.1.1. *Under Assumptions 2.3.1-2.3.7, the optimal policy is identified by*

$$\arg\max_{\mathcal{D}} E\left[W_1 I\{A = d(X)\}\right] = \arg\max_{\mathcal{D}} E\left[W_2 I\{Z = d(X)\}\right] = \arg\max_{\mathcal{D}} E\left[\Delta(X) d(X)\right], \tag{3.1}$$

under the union model $\mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$.

We also consider using modern machine learning methods to estimate these nuisance parameters. In practice, we apply the cross-fitting technique [Schick, 1986, Zheng and van der Laan, 2010, Chernozhukov et al., 2018], which is easy to implement. The cross-fitting procedure goes as follows. We randomly split data into K folds; the cross-fitted estimator is given by

$$\hat{M}_{CF} = \frac{1}{K}\sum_{k=1}^{K} P_{n,k}\{\Delta(O; \hat{\mu}_{A,-k}, \hat{\mu}_{Y,-k}, \hat{\pi}_{-k}) d(X)\},$$

where $P_{n,k}$ denote empirical averages only over the k-th fold, and $\hat{\mu}_{A,-k}$, $\hat{\mu}_{Y,-k}$ and $\hat{\pi}_{-k}$ denote the nuisance estimators constructed excluding the k-th fold. Similar cross-fitted estimators for $E\left[W_1 I\{A = d(X)\}\right]$ and $E\left[W_2 I\{Z = d(X)\}\right]$ can also be constructed in the same way.

3.2 Asymptotic analysis of policy learning

In this section, we study theoretical guarantees for our proposed policy learning approaches. While researchers have suggested applying machine learning algorithms to estimate the optimal policies from large classes which cannot be described by a finite dimensional parameter [Luedtke and van der Laan, 2016b, Künzel et al., 2019], it is also important to consider certain classes of policies for better interpretability and transparency, especially in clinical medicine and policy research [Zhang et al., 2015, Athey and Wager, 2021]. Specifically, here we focus on a class of feasible policies $\mathcal{D} = \left\{I\{\eta^\top X > 0\} : \eta \in \mathbb{H}\right\}$, where η indexes different policies and $\mathbb{H}$ is a compact subset of $\mathbb{R}^p$. That is, we analyze the following estimator:

$$\hat{\eta} = \arg\max_{\eta \in \mathbb{H}} \hat{M}(\eta) = \arg\max_{\eta \in \mathbb{H}} \frac{1}{n}\sum_{i=1}^{n} \hat{\Delta}(O_i) d(X_i; \eta),$$

where $\hat{M}(\eta)$ is estimated by posited parametric models, or the cross-fitted estimator. Let $\eta^* = \arg\max_{\eta \in \mathbb{H}} E[\Delta(X) d(X; \eta)]$ denote the Euclidean parameter that indexes the optimal policy. We detail the main large sample property of our proposed estimator, that $\hat{\eta}$ converges to η^* at $n^{1/3}$ rate, and that $\hat{M}(\hat{\eta})$ is $n^{1/2}$-consistent and asymptotically normal under weak conditions (mostly requiring standard regularity conditions [White, 1982], or only that the nuisance parameters are estimated at faster than $n^{1/4}$ rates).

Remark 3.2.1. In order to obtain certain rates of convergence or regret bounds, it is necessary to require some control over the complexity of the class $\mathcal{D}$; see Athey and Wager [2021, Section 2.2] for examples of the VC-dimension of classes of linear rules, decision trees and monotone rules. Here we apply the empirical process techniques to establish theoretical guarantees for linear rules, which also hold on any other $\mathcal{D}$ indexed by finite-dimensional parameters. Also note that all identification and semiparametric efficiency results hold for any class of policies, and other optimization methods can be readily utilized.

We assume the following regularity conditions.

Condition 1. (i) *The supports of X and Y are bounded.* (ii) *The functions $\mu_Y(t,z,x)$, $\mu_A(t,z,x)$ and $\pi(t,z,x)$ are smooth and bounded for all (t,z,x).* (iii) *The function $M(\eta)$ is twice continuously differentiable in a neighborhood of η^*;* (iv) *For all $\delta > 0$, we have that $Pr(|X^T\eta^*| \leq \delta) \leq c_2\delta$, for some constant $c_2 > 0$ such that $c_2\delta \leq 1$.*

Condition 2. (i) $\sqrt{n}(\hat{\alpha}-\alpha^*) = O_p(1)$; (ii) $\sqrt{n}(\hat{\beta}-\beta^*) = O_p(1)$; (iii) $\sqrt{n}(\hat{\theta}-\theta^*) = O_p(1)$.

Theorem 3.2.2. *Under Assumptions 2.3.1-2.3.7, if Conditions 1 and 2 hold, we have* (i) $\|\hat{\eta}-\eta^*\|_2 = O_p(n^{-1/3})$; (ii) $\sqrt{n}\{M(\hat{\eta})-M(\eta^*)\} = o_p(1)$; (iii) $\sqrt{n}\{\hat{M}(\hat{\eta})-M(\eta^*)\} \to \mathcal{N}(0,\sigma_1^2)$, *where σ_1^2 is given in the Supplementary Material.*

Condition 1 (i), (ii) and (iii) are standard regularity conditions to establish uniform convergence. Condition 1 (iv), also known as the margin condition, is often assumed in the literature of classification [Tsybakov, 2004], reinforcement learning [Hu et al., 2022] and treatment assignment policies [Luedtke and Chambaz, 2020], to guarantee fast convergence rates. Condition 2 requires $\sqrt{n}$ convergence rates of parameter estimates of the posited models, which holds under mild conditions.

We assume the following conditions for the machine learning algorithms used to construct cross-fitted estimators.

Condition 3. $\|\hat{\mu}_A(t,z,X)-\mu_A(t,z,X)\|_{L_2} = o_p(n^{-1/4})$, $\|\hat{\mu}_Y(t,z,X)-\mu_Y(t,z,X)\|_{L_2} = o_p(n^{-1/4})$ *and* $\|\hat{\pi}(t,z,X)-\pi(t,z,X)\|_{L_2} = o_p(n^{-1/4})$, *for* $t,z=0,1$.

Theorem 3.2.3. *Under Assumptions 2.3.1-2.3.7, if Conditions 1 and 3 hold, we have* (i) $\|\hat{\eta}-\eta^*\|_2 = O_p(n^{-1/3})$; (ii) $\sqrt{n}\{M(\hat{\eta})-M(\eta^*)\} = o_p(1)$; (iii) $\sqrt{n}\{\hat{M}(\hat{\eta})-M(\eta^*)\} \to \mathcal{N}(0,\sigma_2^2)$, *where σ_2^2 is given in the Supplementary Material.*

Condition 3 says the nuisance estimators must be consistent and converge at a fast enough rate (essentially $n^{1/4}$ in L_2 norm). This is quite general and can be achieved by many existing algorithms under nonparametric smoothness, sparsity, or other structural constraints. According to Theorems 3.2.2 and 3.2.3 (ii), the regret of our estimated regime vanishes as the sample size increases. Theorems 3.2.2 and 3.2.3 (iii) imply that $\hat{M}(\hat{\eta})$ is a regular and asymptotic normal estimator of $M(\eta^*)$.

CHAPTER 4

NUMERICAL EXPERIMENTS

4.1 Simulations

In this section, we conduct extensive simulations to evaluate the finite-sample performance of the proposed estimators. Specifically, we compare them to the instrumental variable approach proposed by Cui and Tchetgen Tchetgen [2021], which is in principle valid only for a single time point. Replication code is available at GitHub.

We first describe the complete data generation process as follows. Baseline covariates $X = (X_1, X_2)^\top$ are generated from independent standard normal distributions. The time period indicator T is generated from a Bernoulli distribution with probability 0.5. The unmeasured confounders $U = (U_0, U_1)^\top$ are generated from independent bridge distributions with parameter 0.5[1]. The instrumental variable Z is generated from a Bernoulli distribution with probability 0.5. The potential treatments and outcomes at time points $t = 0, 1$ are generated from the models:

$$
\begin{aligned}
Pr(A_0 = 1 \mid Z, U, X) &= \text{expit}(2 - 7Z + 0.2U_0 + 2X_1), \\
Pr(A_1 = 1 \mid Z, U, X) &= \text{expit}(-1.5 + 5Z - 0.15U_1 + 1.5X_2), \\
(Y_0 \mid Z, U, X, A_0) \sim \mathcal{N}(\mu_0, 1), &\quad (Y_1 \mid Z, U, X, A_1) \sim \mathcal{N}(\mu_1, 1),
\end{aligned}
$$

where $\mu_0 = 200 + 10(A_0(1.5X_1 + 2X_2 - 0.5) + 0.5U_0 + 2Z + 1.5X_1 + 2X_2)$, and $\mu_1 = 240 + 10(A_1(1.5X_1 + 2X_2 - 0.5) + 0.5U_1 + 2Z + 2X_1 + 1.5X_2)$. Therefore, the optimal policy is $d_{\text{opt}}(x) = I\{3x_1 + 4x_2 - 1 > 0\}$. Let $A = TA_1 + (1-T)A_0, Y = TY_1 + (1-T)Y_0$; thus the observed cross-sectional data are (X, A, Y, T, Z).

A large test dataset of size $N = 1 \times 10^6$ is generated independently to evaluate the performance of different estimators. The percentage of correct decisions (PCD) of an estimated policy $\hat{d}(x)$ is computed by $1 - N^{-1} \sum_{i=1}^{N} |\hat{d}(X_i) - d_{\text{opt}}(X_i)|$.

We compare 7 estimators in our study: the two IPW estimators, the Wald estimator, and the two multiply robust estimators, along with the below IV estimators proposed

1. The bridge density function is $p(u) = 1/(2\pi \cosh(u/2))$. We use the bridge distribution because by Wang and Louis [2003], the data generation process ensures that upon marginalizing over U, the model for $Pr(A_t = 1 \mid Z, X)$ remains a logistic regression.

by Cui and Tchetgen Tchetgen [2021]:

$$d_{\text{IV.t0}} = \arg\max_{d\in\mathcal{D}} \frac{1}{n_{t0}} \sum_{i=1}^{n} \frac{Z_i A_i Y_i I\{A_i = d(X_i)\} I\{T_i = 0\}}{\hat{\delta}_{t0}(X_i)\hat{\pi}_{t0}(Z_i, X_i)},$$
$$d_{\text{IV.t1}} = \arg\max_{d\in\mathcal{D}} \frac{1}{n_{t1}} \sum_{i=1}^{n} \frac{Z_i A_i Y_i I\{A_i = d(X_i)\} I\{T_i = 1\}}{\hat{\delta}_{t1}(X_i)\hat{\pi}_{t1}(Z_i, X_i)},$$

where n_{t0} and n_{t1} are the sample sizes at time point $0, 1$, respectively; $\delta_{t0}(x) = \mu_A(0,1,x) - \mu_A(0,0,x)$, $\delta_{t1}(x) = \mu_A(1,1,x) - \mu_A(1,0,x)$, $\pi_{t0}(z,x) = Pr(Z = z \mid X = x, T = 0)$, $\pi_{t1}(z,x) = Pr(Z = z \mid X = x, T = 1)$ are the nuisance parameters, and $\hat{\delta}_{t0}, \hat{\delta}_{t1}, \hat{\pi}_{t0}, \hat{\pi}_{t1}$ can be estimated using parametric models or machine learning algorithms. We utilize the genetic algorithm implemented in the R package `rgenoud` [Mebane Jr and Sekhon, 2011] to solve the optimization tasks.

First, we posit parametric models for the nuisance parameters. The linear/logistic regression models for $\mu_A(t,z,x;\alpha)$, $\mu_Y(t,z,x;\beta)$ and $\pi(t,z,x;\theta)$ are correctly specified. The sample size is $n = 5000$.

We also consider flexible machine learning algorithms for nuisance parameter estimation. Specifically, we apply the generalized random forests [Athey et al., 2019] implemented in the R package `grf` with default tuning parameters. For the cross-fitting procedure, we use $K = 4$ folds. The sample size is $n = 10^4$.

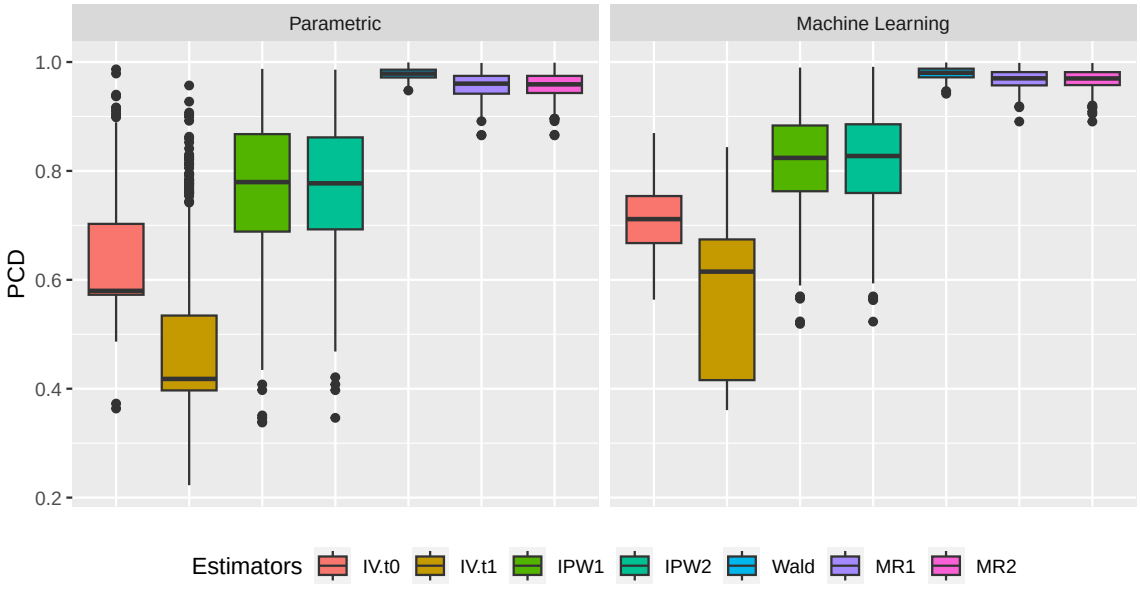


Figure 4.1 – *The percentage of correct decisions (PCD) results of the estimated optimal policies, using parametric models (left) or machine learning (right).*

Figure 4.1 reports the main simulation results from 500 Monte Carlo replications. In both scenarios, the two standard IV estimators fail to learn the optimal policy, due to the direct effects of the treatment A on the outcomes Y_0, Y_1. The two IPW estimators perform much better, but the variability can be large due to possibly extreme weights. The Wald and multiply robust estimators generally lead to lower variability, and attain

superior performance. Additional simulation results are reported in Section A.11 of the Supplementary Material to illustrate how different sample sizes and the strength of the IV affect the performance of the estimated policies. We observe that a stronger strength of IV generally leads to lower variability and better accuracy, and also as sample size increases, our proposed methods have better performance.

4.2 Data application

In this section, we illustrate the use of the instrumented DiD approach for policy learning with a analysis of the Australian Longitudinal Survey (ALS) data. Researchers in labor economics have a longstanding interest in investigating the causal effect of education on earnings in the labor market. Card [2001] suggests that the endogeneity of education might partially explain the continuing interest "in this very difficult task of uncovering the causal effect of education in labor market outcomes", and argues that the effects of education are heterogeneous since the economic benefits are individual-specific. Besides the well acknowledged benefits of personal growth and social good from education, we aim to provide a personalized recommendation on whether an individual should pursuit more education or not, in order to gain higher earnings.

The Australian Longitudinal Survey was conducted annually since 1984. Specifically, we include the 1984 and 1985 waves as cross sectional data in our analysis. The 1984 wave surveyed a sample of 3000 people aged $15-24$, and the 1985 wave consisted of 9000 interviews with people aged $16-25$. The surveys aim mainly at providing data on the dynamics of the youth labour market, and include basic demographic variables, labour market variables, background variables and topics related to the main labour market theme. We follow the guidelines from Su et al. [2013], Cai et al. [2006] and Vella [1994], who was among the first researchers extensively working with the ALS data. Finally, our data include 2401 subjects from the 1984 wave, and 8997 subjects from the 1985 wave. We consider the following baseline covariates: whether a person is born in Australia, marital status, union membership, government employment, age and work experience. The treatment is the education level, and the outcome is the hourly wage. We use an index of labor market attitudes as the instrumental variable [Su et al., 2013]. The details of our analysis are provided in Section A.12 of the Supplementary Material.

The nuisance parameters are estimated by posited linear/logistic regression models, and we apply our proposed methods with the same configurations as Section 4.1. The policy coefficient estimates of all covariates are reported in Table 4.1.

The coefficients should be interpreted cautiously. We also find that there exists some discrepancies among the treatment recommendations by our proposed estimators. The Wald and multiply robust estimators usually agree, but the variability of the IPW estimators are a bit large. Due to the potentially different recommendations by different estimated policies, one may conservatively suggest a recommendation by the majority rule, and accordingly obtain an ensemble policy. It is also interesting to construct a decision tree to further explore which covariates indicate which treatment level [Qi et al., 2023].

Policies	intercept	born_australia	married	uni_mem	gov_emp	age	year_expe
IV.t0	0.4442	−0.4547	0.1311	−0.1179	−0.5181	0.0080	−0.5444
IV.t1	−0.2518	−0.3103	0.2445	−0.6157	−0.1406	0.2015	−0.5840
IPW1	−0.4203	−0.0847	0.5454	−0.3941	−0.5690	0.0299	0.1969
IPW2	−0.2503	−0.0529	0.6051	−0.4384	−0.5801	0.0207	0.1980
Wald	0.5032	0.3891	0.4738	0.5755	−0.1656	−0.0772	0.0793
MR1	−0.0513	0.1341	−0.6039	0.4127	0.5861	−0.0226	−0.3168
MR2	0.5480	−0.3937	−0.4072	0.4393	0.4167	−0.0302	−0.1064

Table 4.1 – *Coefficients of estimated optimal policy (normalized with L_2 norm 1).* `born_australia`*: whether a person is born in Australia;* `married`*: marital status;* `uni_mem`*: union membership;* `gov_emp`*: government employment;* `age`*: age;* `year_expe`*: work experience.*

CHAPTER 5

EXTENSION AND DISCUSSION

5.1 Extension to panel data

In this section, we consider extending the instrumented DiD approach to the panel data setup where a random sample from the population is followed up over two time points [Abadie, 2005]. The observed data are $O = (X, Z, A_0, Y_0, A_1, Y_1)$. Let $\delta_{Y,z}(x) = E[Y_1 - Y_0 \mid X = x, Z = z]$, $\delta_{A,z}(x) = E[A_1 - A_0 \mid X = x, Z = z]$, and $\pi_Z(x) = Pr(Z = 1 \mid X = x)$. We make the following identification assumptions.

Assumption 5.1.1. Suppose the following assumptions hold: (consistency) $A_t = A_t(Z)$ and $Y_t = Y_t(A_t)$ for $t = 0, 1$; (positivity) $c_3 < \pi_Z(x) < 1 - c_3$ for some $0 < c_3 < 1/2$; (trend relevance) $E[A_1(1) - A_0(1) \mid Z = 1, X] \neq E[A_1(0) - A_0(0) \mid Z = 0, X]$; (stable treatment effect over time) $E[Y_0(1) - Y_0(0) \mid X] = E[Y_1(1) - Y_1(0) \mid X]$; (independence & exclusion restriction) $Z \perp \{A_t(1), A_t(0), Y_t(1) - Y_t(0), Y_1(0) - Y_0(0) : t = 0, 1\} \mid X$; (no unmeasured common effect modifier) $Cov\{A_t(1) - A_t(0), Y_t(1) - Y_t(0) \mid X\} = 0$ for $t = 0, 1$.

Assumption 5.1.1 is the counterpart of Assumptions 2.3.1-2.3.7 for the panel/longitudinal structure. Vo et al. [2022] use a structural mean model and consider alternative assumptions to the no unmeasured common effect modifier assumption above. In Section A.10 of the Supplementary Material, we also prove the identification results under the following assumptions that replaces the no unmeasured common effect modifier assumption: (sequential ignorability) $Y_t(a) \perp A_t \mid U, X, Z$ for $t, a = 0, 1$, and there is no additive interaction of either (i) $E[A_1 - A_0 \mid X, U, Z = 1] - E[A_1 - A_0 \mid X, U, Z = 0] = E[A_1 - A_0 \mid X, Z = 1] - E[A_1 - A_0 \mid X, Z = 0]$ or (ii) $E[Y_t(1) - Y_t(0) \mid U, X] = E[Y_t(1) - Y_t(0) \mid X]$ for $t = 0, 1$. The sequential ignorability is intuitive, and commonly assumed in panel/longitudinal data analysis. We note that the no additive interaction assumption implies the no unmeasured common effect modifier assumption.

Theorem 5.1.2. *Under Assumption 5.1.1, the CATE is nonparametrically identified by*

$$\tau(x) = \frac{E[Y_1 - Y_0 \mid X = x, Z = 1] - E[Y_1 - Y_0 \mid X = x, Z = 0]}{E[A_1 - A_0 \mid X = x, Z = 1] - E[A_1 - A_0 \mid X = x, Z = 0]}, \tag{5.1}$$

and the efficient influence function is

$$\phi_{\text{panel}} = \frac{\delta_{Y,1}(x) - \delta_{Y,0}(x)}{\delta_{A,1}(x) - \delta_{A,0}(x)} - \frac{z - \pi_Z(x)}{\pi_Z(x)(1-\pi_Z(x))(\delta_{A,1}(x) - \delta_{A,0}(x))^2} \{(y_1 - y_0)(\delta_{A,1}(x) - \delta_{A,0}(x)) - (a_1 - a_0)(\delta_{Y,1}(x) - \delta_{Y,0}(x)) + \delta_{Y,1}(x)\delta_{A,0}(x) - \delta_{Y,0}(x)\delta_{A,1}(x)\} - \tau(x).$$

Theorem 5.1.3. *Under Assumption 5.1.1, the optimal policy is nonparametrically identified by*

$$\arg\max_{\mathcal{D}} E\left[\frac{\delta_{Y,1}(X) - \delta_{Y,0}(X)}{\delta_{A,1}(X) - \delta_{A,0}(X)} d(X)\right] = \arg\max_{\mathcal{D}} E\left[\Delta_{\text{panel}}(X) d(X)\right], \tag{5.2}$$

where the uncentered efficient influence function Δ_{panel} *is*

$$\Delta_{\text{panel}} = \frac{\delta_{Y,1}(X) - \delta_{Y,0}(X)}{\delta_{A,1}(X) - \delta_{A,0}(X)} - \frac{Z - \pi_Z(X)}{\pi_Z(X)(1-\pi_Z(X))(\delta_{A,1}(X) - \delta_{A,0}(X))^2} \{(Y_1 - Y_0)(\delta_{A,1}(X) - \delta_{A,0}(X)) - (A_1 - A_0)(\delta_{Y,1}(X) - \delta_{Y,0}(X)) + \delta_{Y,1}(X)\delta_{A,0}(X) - \delta_{Y,0}(X)\delta_{A,1}(X)\}.$$

Estimators of optimal policies can be constructed by the empirical versions of equations in Theorem 5.1.3, and the cross-fitting procedure can also be applied when using the efficient influence function. Similarly, asymptotic analysis of policy learning as Theorems 3.2.2 and 3.2.3 can be established for panel data.

5.2 Discussion

Similar approaches as the instrumented difference-in-differences design has long been employed by econometricians [Duflo, 2001] and has also been formally considered as fuzzy differences-in-differences by De Chaisemartin and d'Haultfoeuille [2018], where the individuals can switch treatment in only one direction within each treatment group. We refer interested readers to Ye et al. [2022] and its rejoinder for discussions on the differences, and applications in biomedicine and epidemiology.

There are several interesting directions for future research and application. Our approach is the first work to systematically study policy learning under the DiD setting. It may be possible to consider alternative assumptions or structures in DiD design to learn the optimal policy. Our instrumented DiD may also be generalized to multiple time points, continuous time, or continuous IV.

Note that Assumption 2.3.7 can be replaced by the *monotonicity* assumption, i.e. $A_t(1) \geq A_t(0)$ for $t = 0, 1$ with probability 1, which identifies the complier treatment effects. Then we can also target complier optimal policies that would optimize the potential outcome among compliers.

Part III

POSITIVITY-FREE POLICY LEARNING WITH OBSERVATIONAL DATA

Policy learning utilizing observational data is pivotal across various domains, with the objective of learning the optimal treatment assignment policy while adhering to specific constraints such as fairness, budget, and simplicity. This study introduces a novel positivity-free (stochastic) policy learning framework designed to address the challenges posed by the impracticality of the positivity assumption in real-world scenarios. This framework leverages incremental propensity score policies to adjust propensity score values instead of assigning fixed values to treatments. We characterize these incremental propensity score policies and establish identification conditions, employing semiparametric efficiency theory to propose efficient estimators capable of achieving rapid convergence rates, even when integrated with advanced machine learning algorithms. This paper provides a thorough exploration of the theoretical guarantees associated with policy learning and validates the proposed framework's finite-sample performance through comprehensive numerical experiments, ensuring the identification of causal effects from observational data is both robust and reliable. [1]

1. co-authored with Antoine Chambaz, Julie Josse and Shu Yang, published in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, PMLR 238:1918-1926, 2024*, and selected as an *oral presentation*.

CHAPTER 6

INCREMENTAL PROPENSITY SCORE

6.1 Introduction

Over the past decade, methodologies for learning treatment assignment policies have seen substantial advancements in fields like biostatistics [Luedtke and van der Laan, 2016b, Tsiatis et al., 2019], computer science [Uehara et al., 2022, Yu et al., 2022], and econometrics [Athey and Wager, 2021, Jia et al., 2023]. The core objective of data-driven policy learning is to learn optimal policies that map individual characteristics to treatment assignments to optimize some utility or outcome functions. This is crucial for deriving robust and trustworthy policies in high-stakes decision-making settings, requiring adherence to standard causal assumptions: consistency, unconfoundedness, and positivity [van der Laan et al., 2011, Imbens and Rubin, 2015].

Various statistical and machine-learning methods have been developed to address policy learning tasks. Popular approaches include model-based methods such as Q-learning and A-learning [Murphy, 2003, Shi et al., 2018], and direct model-free policy search methods such as decision trees and outcome weighted learning [Zhang et al., 2012b, Cui et al., 2017], among others [Bibaut et al., 2021, Zhou et al., 2023b]. Another prevailing line of work concerns heterogeneous treatment effects estimation [Wager and Athey, 2018, Künzel et al., 2019, Nie and Wager, 2021, Kallus and Oprescu, 2023], where the sign of the conditional average treatment effects equivalently determines the optimal policy.

However, most methods depend heavily on the three standard causal assumptions to identify causal effects and optimal policies. Recent progress has been made to relax the consistency and unconfoundedness assumptions [Cortez et al., 2022, Kallus and Zhou, 2018], but advancements addressing the violation of the positivity assumption are scarce. Yang and Ding [2018b] and Branson et al. [2023] provide estimation and asymptotic inference results for propensity score trimming with binary and continuous treatments. Lawrence et al. [2017] consider counterfactual learning from deterministic bandit logs under lack of sufficient exploration. Gui and Veitch [2023] use supervised representation learning to estimate causal effects for text data with apparent overlap violation. Zhang et al. [2023] consider a missing-at-random mechanism without a positivity condition for generalizable and double robust inference for average treatment effects under selection bias with decaying overlap. Jin et al. [2022] use pessimism and generalized empirical Bernstein's inequality to study offline policy learning without

assuming any uniform overlap condition. Khan et al. [2023] provide partial identification results for off-policy evaluation under non-parametric Lipschitz smoothness assumptions on the conditional mean function, and thus avoid assuming either overlap or a well-specified model. Liu et al. [2023] propose the overlap weighted average treatment effect on the treated under lack of positivity. To our knowledge, our work is the first to consider learning treatment assignment policies while avoiding the positivity assumption.

This study introduces a novel positivity-free policy learning framework focusing on dynamic and stochastic policies, which are practical. We propose *incremental propensity score policies* that shift propensity scores by an individualized parameter, requiring only the consistency and unconfoundedness causal assumptions. Our approach enhances the concept of incremental intervention effects, as proposed by Kennedy [2019], adapting it to individual treatment policy contexts.

We also use semiparametric theory to characterize the efficient influence function [Bickel et al., 1993, van der Laan and Robins, 2003], which serves as the foundation to construct estimators with favorable properties, such as double/multiple robustness and asymptotically negligible second-order bias (also called Neyman orthogonality in double machine learning [Chernozhukov et al., 2018] or orthogonal statistical learning [Foster and Syrgkanis, 2023]). Thus, our proposed estimators can attain fast parametric $\sqrt{n}$ convergence rates, even when nuisance parameters are estimated at slower rates such as $n^{1/4}$ via flexible machine learning algorithms.

Based on the above efficient off-policy evaluation results, we propose approaches to learning the optimal policy by maximizing the value function, possibly under application-specific constraints. Several examples are provided in Section 7.1, including fairness and resource limit. While it remains an open problem to provide finite sample or asymptotic regret bounds as Athey and Wager [2021] for stochastic policy learning with constraints, which is out of the scope of this article, we establish asymptotic guarantees for our proposed policy learning methods under alternative (stronger) conditions.

The rest of this article is organized as follows. Section 6.2 introduces the basic setup and notations and proposes the incremental propensity score policy. Our main identification and semiparametric efficiency theory results for off-policy evaluation are presented in Section 6.3. Section 7.1 formally introduces our positivy-free policy learning framework, with several examples. Asymptotic analysis of guarantees for policy evaluation and learning are given in Section 7.2. Finally, we illustrate our methods via simulations and a data application in Section 7.3. The article concludes in Section 7.4 with a discussion of some remarks and future work. All proofs and additional results are provided in the Supplementary Material.

6.2 Statistical Framework

We first introduce the notations and setup. Let X denote the p-dimensional vector of covariates that belongs to a covariate space $\mathcal{X} \subset \mathbb{R}^p$, $A \in \mathcal{A} = \{0, 1\}$ denote the binary treatment, $Y \in \mathbb{R}$ denote the outcome of interest. Without loss of generality,

we assume throughout that larger values of Y are more desirable. Our observed data structure is $O = (X, A, Y)$. Suppose that our collected random sample $(O_1, \ldots, O_n)$ of size n are independent and identically distributed (i.i.d.) observations of $O \sim P$, where P denote the true distribution of the observed data.

Now, we are in the position to introduce different types of policies or interventions commonly used in the literature: (i) under *static* policies, the same treatments would be applied indiscriminately, while *dynamic* policies depend on individual characteristics; (ii) *deterministic* policies recommend one specific treatment and *stochastic* policies output probabilities of prescribing each treatment level. This article focuses on dynamic and stochastic policies, which are more practical in various settings and have received substantial recent interest. Typical examples include point exposures [Dudík et al., 2014], longitudinal studies [Tian, 2008, Murphy et al., 2001, van der Laan and Petersen, 2007], natural stochastic policies in reinforcement learning [Kallus and Uehara, 2020], and particularly interventions that depend on the observational treatment process [Muñoz and van Der Laan, 2012, Haneuse and Rotnitzky, 2013, Young et al., 2014]; but none of the existing intervention effects both avoids positivity conditions entirely and is completely nonparametric.

We use the potential outcomes framework [Neyman, 1923, Rubin, 1974] to define causal effects. Let $Y(a)$ denote the potential outcome had the treatment a been assigned. A policy $d : \mathcal{X} \to \{0, 1\}$ is deterministic if it maps individual characteristics x to a treatment assignment 0 or 1, and the output of a stochastic policy $d : \mathcal{X} \to [0, 1]$ is the probability of assigning treatment 1. Let $\mathcal{D}$ denote a pre-specified class of policies of interest, where each policy $d \in \mathcal{D}$ induces the value function defined by

$$V(d) = E[Y(d)] = E[Y(1)d(X) + Y(0)(1 - d(X))],$$

where $Y(d)$ is the potential outcome under the policy d. In Remark 6.2.3, we briefly review standard (deterministic) policy learning methods. In our framework, we focus on dynamic and stochastic policies. Our goal is to directly search for the optimal policy d^* that maximizes the value function $V(d)$, possibly under application-specific constraints $c(d) \leq 0$. See Section 7.1 for detailed examples.

6.2.1 Causal Assumptions

We make the following identification assumptions.

Assumption 6.2.1 (Consistency)**.** $Y = Y(A)$.

Assumption 6.2.2 (Unconfoundedness)**.** $A \perp Y(a) \mid X$ for $a = 0, 1$.

Assumption 6.2.1 is also known as the stable unit treatment value assumption, which says there should be no multiple versions of the treatment and no interference between units. Assumption 6.2.2 states that there are no unmeasured confounders so that treatment assignment is as good as random conditional on the covariates X. In this article, we entirely avoid the positivity assumption which requires that each unit has a positive probability of receiving both treatment levels, i.e., $c < Pr(A = 1 \mid X) < 1 - c$ for some constant $c > 0$.

Remark 6.2.3. Standard policy learning methods need all of Assumptions 6.2.1, 6.2.2 and the positivity assumption to identify the value function of deterministic policies $d : \mathcal{X} \to \mathcal{A}$ by the outcome regression (OR), inverse probability weighting (IPW) and augmented IPW (AIPW) formulas:

$$V_{\text{OR}}(d) = E[E[Y \mid X, A = d(X)]], \quad V_{\text{IPW}}(d) = E\left[\frac{I\{A = d(X)\}Y}{Pr(A = d(X) \mid X)}\right],$$
$$V_{\text{AIPW}}(d) = E\left[E[Y \mid X, A = d(X)] + \frac{I\{A = d(X)\}(Y - E[Y \mid X, A = d(X)])}{Pr(A = d(X) \mid X)}\right],$$

thus the optimal policies are given by $d^*_{\text{OR}} = \arg\max_{d\in\mathcal{D}} V_{\text{OR}}(d)$, $d^*_{\text{IPW}} = \arg\max_{d\in\mathcal{D}} V_{\text{IPW}}(d)$, and $d^*_{\text{AIPW}} = \arg\max_{d\in\mathcal{D}} V_{\text{AIPW}}(d)$, possibly under application-specific constraints. When the positivity is violated, it is error-prone to rely on the outcome regression model's extrapolation, and the IPW and AIPW estimators would fail due to division by zero.

6.2.2 Incremental Propensity Score Policies

Kennedy [2019] propose a new class of stochastic dynamic intervention, called incremental propensity score interventions, and show that these interventions are non-parametrically identified without requiring any positivity restrictions on the propensity scores. Specifically, their proposed intervention replaces the observational propensity score π with a shifted version based on multiplying the odds of receiving treatment, $\delta\pi(x)/\{\delta\pi(x) + 1 - \pi(x)\}$, where the increment parameter $\delta \in (0, \infty)$ is user-specified and dictates the extent to which the propensity scores fluctuate from their actual observational values. Some motivation and examples, efficiency theory, and estimators for mean outcomes under these interventions are studied in detail by Kennedy [2019].

We propose a positivity-free (stochastic) policy learning framework based on the incremental propensity score interventions. Specifically, we consider the stochastic policy $d : \mathcal{X} \to [0, 1]$ that assigns treatment 1 with probability

$$d(x) = \frac{\delta(x)\pi(x)}{\delta(x)\pi(x) + 1 - \pi(x)}, \tag{6.1}$$

where $\delta(x)$ enables individualized treatment assignment. We note that the choice of $d(x)$ in (6.1) is motivated by its interpretability and positivity-free. In particular, whenever $0 < \pi(x) < 1$, $\delta(x) = [d(x)/\{1 - d(x)\}]/[\pi(x)/\{1 - \pi(x)\}]$ is simply an odds ratio, indicating how the policy changes the odds of receiving treatment. When positivity is violated, we have that $d(x) = 0$ if $\pi(x) = 0$, and $d(x) = 1$ if $\pi(x) = 1$.

6.3 Identification and Efficiency Theory

6.3.1 Identification

We first give formal identification results for the value function of incremental propensity score policies, which require no conditions on the propensity scores.

Proposition 6.3.1 (Identification formulas). *Under Assumptions 6.2.1 and 6.2.2, the value function $V(d)$ can be nonparametrically identified by the outcome regression with incremental propensity score (OR-IPS) formula:*

$$V_{\text{OR-IPS}}(d) = E\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\right], \tag{6.2}$$

where $\mu_a(X) = E[Y \mid X, A = a], a = 0, 1$ are the outcome regression functions or the inverse probability weighting of incremental propensity score (IPW-IPS) formula:

$$V_{\text{IPW-IPS}}(d) = E\left[\frac{Y\{\delta(X)A + 1 - A\}}{\delta(X)\pi(X) + 1 - \pi(X)}\right]. \tag{6.3}$$

Proposition 6.3.1 shows that the value function can be identified by (i) a weighted average of the outcome regression functions μ_0, μ_1, where the weight on μ_1 is given by the incremental propensity score $d(x)$ and the weight on μ_0 is $1 - d(x)$; (ii) inverse probability weighting where each treated is weighted by the (inverse of the) propensity score plus some fractional contribution of its complement, i.e., $\pi(x) + (1 - \pi(x))/\delta(x)$, and untreated units are weighted by this same amount, except the entire weight is further down-weighted by a factor of $\delta(x)$.

6.3.2 Efficient Off-policy Evaluation

Despite that simple plug-in OR-IPS and IPW-IPS estimators can be easily constructed from (6.2) and (6.3), these estimators will only be $\sqrt{n}$-consistent when the outcome regression or propensity score models are correctly specified. This is usually unrealistic in practice. We use semiparametric efficiency theory to study the following statistical functional of P from a nonparametric statistical model $\mathcal{M}$:

$$\Psi(P) = V(d) = E_P\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\right],$$

and propose efficient estimators based on the efficient influence function.

Proposition 6.3.2 (Semiparametric Efficiency). *The efficient influence function of $\Psi(P)$ is*

$$\begin{aligned}\phi(P)(O) = &\frac{A\delta(X)\{Y - \mu_1(X)\} + (1-A)\{Y - \mu_0(X)\} + \delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)} \\ &+ \frac{\delta(X)\tau(X)\{A - \pi(X)\}}{\{\delta(X)\pi(X) + 1 - \pi(X)\}^2} - \Psi(P),\end{aligned} \tag{6.4}$$

where $\tau(x) = \mu_1(x) - \mu_0(x)$.

By Proposition 6.3.2, the one-step bias-corrected estimator is given by

$$\hat{\Psi}_{\text{OS}} = \Psi(\hat{P}) + P_n\phi(\hat{P})(O) = \frac{1}{n}\sum_{i=1}^{n}\xi(\hat{P})(O_i), \tag{6.5}$$

where we estimate P by $\hat{P}$, and let P_n denote the empirical distribution, and $\xi(P)(O) = \phi(P)(O) + \Psi(P)$ is the uncentered efficient influence function. This estimator can converge at fast parametric $\sqrt{n}$ rates and attain the efficiency bound, even when the propensity score $\pi(x)$ and outcome regression functions μ_0, μ_1 are modeled flexibly and estimated at rates slower than $\sqrt{n}$, as long as these nuisance functions are estimated consistently at rates faster than $n^{1/4}$. This allows much more flexible nonparametric methods and modern machine learning algorithms to be employed.

However, characterizing asymptotic properties of the estimator (6.5) requires some empirical process conditions that restrict the flexibility and complexity of the nuisance estimators; otherwise, we will have overfitting bias and intractable asymptotic behaviors. See the asymptotic analysis in Section 7.2 and proofs thereof. To accommodate the wide use of modern machine learning algorithms that usually fail to satisfy the required empirical process conditions, we apply the cross-fitting procedure to obtain asymptotically normal and efficient estimators [Zheng and van der Laan, 2010, Chernozhukov et al., 2018]. Suppose we randomly split the data into K folds. Then the cross-fitting estimator is

$$\hat{\Psi}_{\text{CF}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\Psi}_k = \frac{1}{K}\sum_{k=1}^{K} P_{n,k}\xi(P_{n,-k})(O), \tag{6.6}$$

where $P_{n,k}$ and $P_{n,-k}$ denote the empirical measures on data from the k-fold and excluding the k-fold, respectively. That is, for $k = 1, \ldots, K$, nuisance estimators are constructed excluding the k-fold, and the value function $\hat{\Psi}_k$ is evaluated on the k-th fold; finally, the cross-fitting estimator is the average of the K value estimators from K folds.

CHAPTER 7

POLICY LEARNING AND EXAMPLES

7.1 From Efficient Policy Evaluation to Learning

In this section, we first present our proposed methods for policy learning.

As discussed in Section 6.2, given a pre-specified policy class $\mathcal{D}$ (e.g., linear decision rules), we propose estimating the optimal treatment assignment rule $\hat{d}$ that solves (i) $\hat{d} = \arg\max_{d\in\mathcal{D}} \hat{V}(d)$, where $\hat{V}(d)$ is a value function estimator by OR-IPS (6.2), IPW-IPS (6.3), one-step (6.5) or cross-fitting (6.6); or (ii) $\hat{d} = \arg\max_{d\in\mathcal{D}} \hat{V}(d)$ subject to $\hat{c}(d) \leq c$, when an application-specific constraint $c(d) \leq c$ is imposed, and $\hat{c}(d)$ is a constraint estimator which usually needs to be studied on a case-by-case basis.

We first review important examples of policy learning that fit into our framework.

Vanilla direct policy search. The first example is what most existing work on policy learning has focused on, primarily for deterministic policies with a binary treatment. When the policy class is unrestricted, the optimal treatment assignment rule depends on the sign of the conditional average treatment effect for each individual unit, which cannot be extended to stochastic policies. Our proposed optimal incremental propensity score policies maximize the value function.

Fair policy learning. In many decision-making scenarios, such as hiring, recommendation systems, and criminal justice, concerns have been raised regarding the fairness of decisions from the learning process [Chzhen et al., 2020]. Let $S \in \mathcal{S}$ denote the sensitive attribute. For randomized predictions $f : \mathcal{X} \times \mathcal{S} \to \Delta(\mathcal{A})$, popular fairness criteria include demographic parity (DP) [Calders et al., 2009]:

$$E[f(X,S) \mid S = s] = E[f(X,S) \mid S = s'], \forall s, s' \in \mathcal{S}, \tag{7.1}$$

which says that $f(X,S)$ is independent from S, or equal opportunity (EO) [Hardt et al., 2016]:

$$E[f(X,S) \mid S = s, A = a] = E[f(X,S) \mid S = s', A = a], \forall s, s' \in \mathcal{S}, a \in \mathcal{A}, \tag{7.2}$$

which requires equal true positive and true negative rates. Following the same spirit, we consider fair policy learning tasks as the constrained optimization problem:

$$\max_{d\in\mathcal{D}} V(d), \text{ subject to } f(d) \leq b,$$

where $f(d)$ is either the DP or EO metrics, which can be estimated by

$$\hat{f}_{\mathrm{DP}}(d) = \left(\sum_{s \in \mathcal{S}} \left(\frac{\sum_{i=1}^n d(X_i) I\{S_i = s\}}{\sum_{i=1}^n I\{S_i = s\}} - \frac{\sum_{i=1}^n d(X_i)}{n} \right)^2 \right)^{1/2},$$

or

$$\hat{f}_{\mathrm{EO}}(d) = \left(\sum_{s \in \mathcal{S}} \left(\frac{\sum_{i=1}^n d(X_i) I\{S_i = s, A_i = 1\}}{\sum_{i=1}^n I\{S_i = s, A_i = 1\}} - \frac{\sum_{i=1}^n d(X_i) I\{A_i = 1\}}{\sum_{i=1}^n I\{A_i = 1\}} \right)^2 \right)^{1/2},$$

and b is a pre-specified tuning parameter.

Resource-limited policy learning. In many real-world applications, the proportion of individuals who can receive the treatment is a priori limited due to a budget or a capacity constraint. So we consider the resource-limited policy learning tasks as the constrained optimization problem:

$$\max_{d \in \mathcal{D}} V(d), \text{ subject to } E[d] \leq b,$$

where b is the pre-specified budget or capacity.

Protect the vulnerable. Since the optimal policy is typically defined as the maximizer of the expected potential outcome over the entire population, such a policy may be suboptimal or even detrimental to certain disadvantaged subgroups. Fang et al. [2022] propose the fairness-oriented optimal policy learning framework:

$$\max_{d \in \mathcal{D}} V(d), \text{ subject to } Q_\tau(Y(d)) \geq b,$$

where $Q_\tau(Y(d)) = \inf\{t : F_{Y(d)}(t) \geq \tau\}$ is the τ-th quantile of $Y(d)$, $F_{Y(d)}$ denotes the cumulative distribution function of $Y(d)$, and b is a pre-specified protection threshold. Note that the quantile function can be estimated by $\hat{Q}_\tau(Y(d)) = \arg\min_q n^{-1} \sum_{i=1}^n c_i(d) \rho_\tau(Y_i - q)$, where $\rho_\tau(u) = u(\tau - I\{u < 0\})$ is the quantile loss function, and $c_i(d) = A_i d(X_i) + (1 - A_i)(1 - d(X_i))$.

Other examples in the literature include the counterfactual no-harm criterion by the principal stratification method [Li et al., 2023], (weakly) NP-hard knapsack problem [Luedtke and van der Laan, 2016a], and instrumental variable methods [Qiu et al., 2021].

7.2 Asymptotic Analysis of Policy Evaluation and Learning

In this section, we first characterize the asymptotic distributions of our proposed one-step estimator (6.5) and the cross-fitted estimator (6.6) for off-policy evaluation.

Theorem 7.2.1. *Assume the following conditions hold:* (i) $\|\hat{\pi}(x) - \pi(x)\|_{L_2} = o_p(n^{-1/4})$, $\|\hat{\mu}_a - \mu_a\|_{L_2} = o_p(n^{-1/4})$ *for* $a = 0, 1$; (ii) $\phi(P)$ *belongs to a Donsker class;* (iii) $|Y|$ *and* $|\delta(X)|$ *are bounded in probability. For the one-step estimator, we have that* $\sqrt{n}(\hat{\Psi}_{\mathrm{OS}} - \Psi(P)) \to \mathcal{N}(0, E[\phi^2])$.

Theorem 7.2.2. *Assume the following conditions hold:* (i) $\|\hat{\pi}(x) - \pi(x)\|_{L_2} = o_p(n^{-1/4})$, $\|\hat{\mu}_a - \mu_a\|_{L_2} = o_p(n^{-1/4})$ *for* $a = 0, 1$; (ii) $|Y|$ *and* $|\delta(X)|$ *are bounded in probability. For the cross-fitting estimator, we have that* $\sqrt{n}(\hat{\Psi}_{\text{CF}} - \Psi(P)) \to \mathcal{N}(0, E[\phi^2])$.

Condition (i) of Theorems 7.2.1 and 7.2.2 is commonly assumed such that the second-order remainder term is $o_p(1)$ [Kennedy, 2022]. Condition (ii) of Theorems 7.2.1 ensures the centered empirical process term is $o_p(1)$. Condition (iii) of Theorems 7.2.1 and condition (ii) of Theorems 7.2.2 are mild regularity conditions. The asymptotic variance of the one-step estimator can be consistently estimated by $\frac{1}{n}\sum_{i=1}^{n}\phi^2(\hat{P})(O_i)$, and the asymptotic variance of the cross-fitting estimator can be consistently estimated by $\frac{1}{K}\sum_{k=1}^{K}P_{n,k}\phi^2(\hat{P}_{-k})(O)$.

Next, we prove asymptotic guarantees for the following generic off-policy learning problem:

$$\max_{d\in\mathcal{D}} \hat{V}(d), \text{ subject to } \hat{c}(d) \leq c,$$

where $\hat{V}(d)$ is a value estimator of our proposed incremental propensity score policies, $\hat{c}(d)$ is an estimate of the constraint, and c is a pre-specified criterion.

Consider a parametric policy class $\mathcal{D}(H)$ indexed by $\eta \in H$, where H is a compact set. Let η^* denote the true Euclidean parameter indexing the optimal policy. To simplify the notation, for $d(x;\eta) \in \mathcal{D}(H)$, we define $V(\eta) = V(d(x;\eta))$ and $c(\eta) = c(d(x;\eta))$.

Theorem 7.2.3. *Assume the following conditions hold:* (i) $d(x;\eta)$ *is a continuously differentiable and convex function with respect to* η; (ii) $\hat{V}(\eta)$ *and* $\hat{c}(\eta)$ *converge to* $V(\eta)$ *and* $c(\eta)$ *at rates* $\sqrt{n}$. *We have that* (i) $V(\hat{\eta}) - V(\eta^*) = O_p(n^{-1/2})$; (ii) $\hat{V}(\hat{\eta}) - V(\eta^*) = O_p(n^{-1/2})$.

Theorem 7.2.4. *Assume the following conditions hold:* (i) $\mathcal{D}$ *is a Glivenko–Cantelli class;* (ii) $\hat{\pi}(x)$ *and* $\hat{\mu}_a(x)$ *are uniformly consistent estimators of* $\pi(x)$ *and* $\mu_a(x)$ *for* $a = 0, 1$; (iii) $\forall d \in \mathcal{D}$, $m \in (0, 1)$, *it follows that* $md \in \mathcal{D}$. *We have that* (i) $V(\hat{d}) - V(d) = o_p(1)$; (ii) $\hat{V}(\hat{d}) - V(d) = o_p(1)$.

Theorem 7.2.3 (i) establishes that the regret of the learned policy attains the convergence rate of $n^{-1/2}$, and (ii) shows that $\hat{V}(\hat{\eta})$ is a $\sqrt{n}$-consistent estimator of the optimal value function for parametric and convex policy classes under mild assumptions. Theorem 7.2.4 (i) establishes that the regret of the learned policy vanishes, and (ii) shows $\hat{V}(\hat{\eta})$ is still a consistent estimator for GC classes.

7.3 Experiments

In this section, we conduct extensive experiments to evaluate the performance of our proposed positivity-free policy learning methods by comparison with standard policy learning methods. Replication code is available at GitHub.

7.3.1 Simulation

We consider the fair policy learning task under the demographic parity constraint and simulate

$$
\begin{aligned}
S &\sim \text{Bernoulli}(0.5), \quad (X_1, X_2, X_3) \sim \text{Uniform}(0, 1), \\
A &\sim \text{Bernoulli}(\text{expit}(-1 - X_1 + 1.5X_2 - 0.25X_3 - 3.1S)), \\
Y(0) &\sim \mathcal{N}\{20(1 + X_1 - X_2 + X_3^2 + \exp(X_2)), 20^2\}, \\
Y(1) &\sim \mathcal{N}\{20(1 + X_1 - X_2 + X_3^2 + \exp(X_2)) + 25(3 - 5X_1 + 2X_2 - 3X_3 + S), 20^2\},
\end{aligned}
$$

where $\text{expit} : x \mapsto 1/(1 + \exp(-x))$. We let S denote the sensitive attribute and X_1, X_2, X_3 the common non-sensitive attributes. The treatment assignment mechanism yields variable propensity scores that can degrade the performance of weighting-based estimators in standard policy learning methods. For standard methods, we consider the policy class of linear rules $\mathcal{D}_{\text{linear}} = \{d(s, x) = I\{(1, s, x_1, x_2, x_3)\beta > 0\} : \beta \in \mathbb{R}^5, \|\beta\|_2 = 1\}$. For the incremental propensity score policies, we consider the class $\mathcal{D}_{\text{IPS}} = \{d(s, x) = \delta(s, x; \beta)\pi(s, x)/\{\delta(s, x; \beta)\pi(s, x) + 1 - \pi(s, x)\} : \beta \in \mathbb{R}^5\}$, which is indexed by $\delta(s, x; \beta) = \exp\{(1, s, x_1, x_2, x_3)\beta\}$.

We estimate the outcome regression model $\mu(s, x)$ and the propensity score $\pi(s, x)$ using the generalized random forests [Athey et al., 2019] implemented in the `R` package `grf`. The constrained optimization problems are solved by the derivative-free linear approximations algorithm [Powell, 1994], implemented in the `R` package `nloptr`. The sample size is $n = 1000$, and the demographic parity threshold is $\tau = 0.01$.

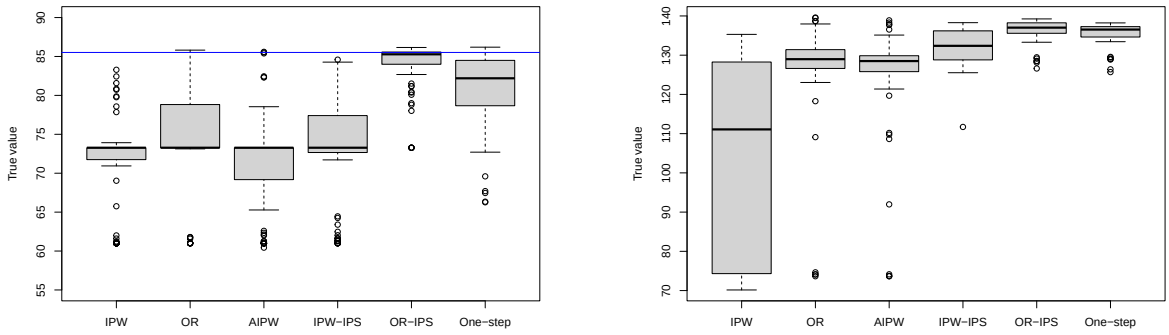
We compare the true values of the estimated optimal policies using test data with sample size $N = 10^5$. The true optimal value is approximated using the test data. Simulation results of 100 Monte Carlo repetition are reported in Figure 7.1(a). When some estimated propensity scores are exactly 0, the IPW and AIPW estimators would fail, and `NA` is returned. Three standard methods IPW, OR, and AIPW have the worst performance. The IPW-IPS estimator also has large variability, which is similarly reported in Kennedy [2019]. The OR-IPS and efficient one-step estimators achieve the best performance with the highest value.

Additional simulation results are given in Section B.7 of the Supplementary Material. Specifically, we illustrate that our proposed policy learning methods have comparable performance when there is no positivity violation, and also illustrate the better performance of our proposed methods when using parametric models.

7.3.2 Data application

We illustrate our proposed methods using semi-synthetic data from the `Fairlearn` open source project [Weerts et al., 2023]. Additional information on our data analysis is provided in Section B.8 of the Supplementary Material.

The Diabetes dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks [Strack et al., 2014], and contains hospital records of patients diagnosed with diabetes who underwent laboratory tests and medications and stayed up to 14 days. Our application aims to learn the optimal



(*a*) *Simulations.* (*b*) *Diabetes data application.*

Figure 7.1 – *Performance of optimal policies under three standard methods* (*IPW, OR, AIPW*) *and our proposed three methods* (*IPW-IPS, OR-IPS, One-step*). *The blue line is the* (*approximate*) *true optimal value.*

policy for prescribing diabetic medication by maximizing the expected outcome under the demographic parity constraint. The sensitive attribute is race, and a violation of positivity exists in the data.

We include 7 baseline covariates: `race`, `gender`, `age`, `time_in_hospital` (number of days between admission and discharge), `num_lab_procedures` (number of lab tests performed during the encounter), `num_medications` (number of distinct generic names administered during the encounter) and `number_diagnoses` (number of diagnoses). Under positivity violation, we are unable to identify the value function, e.g. relying on the outcome regression's extrapolation to learn the counterfactual outcomes on test data. Thus the potential outcomes are simulated as follows: $Y(0) \sim \mathcal{N}\{20(1 + \texttt{gender} - \texttt{age} + \texttt{time\_in\_hospital} + \texttt{num\_lab\_procedures} + \texttt{num\_medications} + \texttt{num\_medications}^2 + \exp(\texttt{number\_diagnoses})), 20^2\}$, and $Y(1) \sim \mathcal{N}\{20(1 + \texttt{gender} - \texttt{age} + \texttt{time\_in\_hospital} + \texttt{num\_lab\_procedures} + \texttt{num\_medications} + \texttt{num\_medications}^2 + \exp(\texttt{number\_diagnoses})) + 25(3 - 5\texttt{age} + 2\texttt{time\_in\_hospital} - 3\texttt{num\_medications} + \texttt{race}), 20^2\}$. The estimation setup and policy classes are the same as previous simulations. We run 50 repetitions; each time we randomly select 500 patients as training data to learn the optimal policy and 2000 patients as test data to evaluate the performance. Empirical results are reported in Figure 7.1(b). When the positivity violation is severer, the IPW estimator has extremely large variability, and we also observe that our proposed methods perform consistently better than the standard methods.

7.4 Discussion

This article proposes a general positivity-free stochastic policy learning framework using observational data, possibly subject to application-specific constraints. There are several interesting directions for future research. It is relevant to extend our methods to the more general case with multiple time points for treatment assignment, multiple

treatment levels, or high-dimensional models [Wei et al., 2023], where positivity is even more likely to be violated. The incremental propensity score approach can also be extended to account for common issues such as covariate shift [Zhao et al., 2023, Lei et al., 2023], censoring and dropout [Cui et al., 2023a], and truncation by death [Chu et al., 2023].

Part IV

EFFICIENT AND ROBUST TRANSFER LEARNING OF OPTIMAL INDIVIDUALIZED TREATMENT REGIMES WITH RIGHT-CENSORED SURVIVAL DATA

An individualized treatment regime (ITR) is a decision rule that assigns treatments based on patients' characteristics. The value function of an ITR is the expected outcome in a counterfactual world had this ITR been implemented. Recently, there has been increasing interest in combining heterogeneous data sources, such as leveraging the complementary features of randomized controlled trial (RCT) data and a large observational study (OS). Usually, a covariate shift exists between the source and target population, rendering the source-optimal ITR unnecessarily optimal for the target population. We present an efficient and robust transfer learning framework for estimating the optimal ITR with right-censored survival data that generalizes well to the target population. The value function accommodates a broad class of functionals of survival distributions, including survival probabilities and restrictive mean survival times (RMSTs). We propose a doubly robust estimator of the value function, and the optimal ITR is learned by maximizing the value function within a pre-specified class of ITRs. We establish the $N^{-1/3}$ rate of convergence for the estimated parameter indexing the optimal ITR, and show that the proposed optimal value estimator is consistent and asymptotically normal even with flexible machine learning methods for nuisance parameter estimation. We evaluate the empirical performance of the proposed method by simulation studies and a real data application of sodium bicarbonate therapy for patients with severe metabolic acidaemia in the intensive care unit (ICU), combining a RCT and an observational study with heterogeneity. [1]

1. co-authored with Julie Josse and Shu Yang, rejected and resubmitted to *Journal of Machine Learning Research*.

CHAPTER 8

INDIVIDUALIZED TREATMENT REGIMES WITH SURVIVAL DATA

8.1 Introduction

Data-driven individualized decision making has recently received increasing interest in many fields, such as precision medicine [Kosorok and Laber, 2019, Tsiatis et al., 2019], mobile health [Trella et al., 2022], precision public health [Rasmussen et al., 2020] and econometrics [Athey and Wager, 2021]. The goal of optimal ITR estimation is to learn a decision rule that assigns the best treatment among possible options to each patient based on their individual characteristics in order to optimize some functional of the counterfactual outcome distribution in the population of interest, also known as the value function. The optimal ITR is the one with the maximal value function, and the value function of the optimal ITR is the optimal value function.

For completely observed data without censoring, one prevailing line of work in the statistical and biomedical literature uses model-based methods to solve the optimal ITR problem, such as Q-learning [Robins, 2004, Qian and Murphy, 2011, Laber et al., 2014] and A-learning [Murphy, 2003, Schulte et al., 2014, Shi et al., 2018]. Alternatively, direct model-free or policy search methods have been proposed recently, including the classification perspective [Zhang et al., 2012a,b, Zhao et al., 2012, Rubin and van der Laan, 2012] and interpretable tree or list-based ITRs [Laber and Zhao, 2015, Zhang et al., 2015, 2018a], among others. In clinical studies, right-censored survival data are frequently observed as primary outcomes. Recent extensions of optimal ITR with survival data have been established in Goldberg and Kosorok [2012], Cui et al. [2017], Jiang et al. [2017], Bai et al. [2017], Díaz et al. [2018], Zhou et al. [2023a].

Researchers have investigated using machine learning algorithms to estimate the optimal ITR from large classes, which cannot be indexed by a finite-dimensional parameter [Luedtke and van der Laan, 2016b,c]. One typical instance is that the optimal ITR can be learned from the blip function, which is defined as the additive effect of a blip in treatment on a counterfactual outcome, conditional on baseline covariates [Robins, 2004]; and most existing regression or supervised learning methods can be directly applied [Künzel et al., 2019]. However, the ITRs learned by machine learning methods can be too complex to inform policy-making and clinical practice; to facilitate the integration of data-driven ITRs into practice, it is crucial that estimated ITRs be

interpretable and parsimonious [Zhang et al., 2015].

Recently, there has been increasing interest in combining heterogeneous data sources, such as leveraging the complementary features of RCT data and a large OS. For example, in biomedical studies and policy research, RCTs are deemed as the gold standard for treatment effects evaluation. However, due to inclusion or exclusion criteria, data availability, and study design, the enrolled participants in RCT who form the source sample may have systematically different characteristics from the target population. Therefore, findings from RCTs cannot be directly extended to the target population of interest [Cole and Stuart, 2010, Dahabreh and Hernán, 2019]. See also Colnet et al. [2020] and Degtiar and Rose [2021] for detailed reviews. Heterogeneity in the populations is of great relevance, and a *covariate shift* usually exists where the covariate distributions differ between the source and target populations; thus, the optimal ITR for the source population is not necessarily optimal for the target population. Zhao et al. [2019] uses data from a single trial study and proposes a two-stage procedure to derive a robust and parsimonious rule for the target population; Mo et al. [2021] proposes a distributionally robust framework that maximizes the worst-case value function under a set of distributions that are "close" to the training distribution; Kallus [2021] tackles the lack of overlap for different actions in policy learning based on retargeting; Wu and Yang [2022] and Chu et al. [2022] develop a calibration weighting framework that tailors a targeted optimal ITR by leveraging the individual covariate data or summary statistics from a target population; Sahoo et al. [2022] uses distributionally robust optimization and sensitivity analysis tools to learn a decision rule that minimizes the worst-case risk incurred under a family of test distributions. However, these methods focus on continuous or binary outcomes and only consider a single sample for worst-case risk minimization; the extension to right-censored survival outcomes within the data integration context has not been studied.

In this paper, we propose a new transfer learning method of finding an optimal ITR from a restricted ITR class under the super population framework where the source sample is subject to selection bias and the target sample is representative of the target population with a known sampling mechanism. Specifically, in our value search method, the value function accommodates a broad class of functionals of survival distributions, including survival probabilities and RMSTs. We characterize the efficient influence function (EIF) of the value function and propose the augmented estimator, which involves models for the survival outcome, propensity score, censoring and sampling processes. The proposed estimator is doubly robust in the sense that it is consistent if either the survival outcome model or the models of the propensity score, censoring, and sampling are correctly specified and is locally efficient when all models are correct. We also consider flexible data-adaptive machine learning algorithms to estimate the nuisance parameters and use the cross-fitting procedure to draw valid inferences under mild regularity conditions and a certain rate of convergence conditions. As we consider a restricted class of ITRs indexed by a Euclidean parameter η, we also establish the $N^{-1/3}$ convergence rate of $\hat{\eta}$, even though its resultant limiting distribution is not standard, and thus very challenging to characterize. Based on this rate of convergence, we show that the proposed estimator for the target value function is consistent and asymptotically normal, even with flexible machine learning methods for

nuisance parameter estimation. Interestingly, when the covariate distributions of the source and target populations are the same, i.e., no covariate shift, the semiparametric efficiency bounds of our method and the standard doubly robust method [Bai et al., 2017] are equal. Moreover, if the true optimal ITR belongs to the restricted class of ITRs, the standard doubly robust method can still learn the optimal ITR despite the covariate shift, but only our method provides valid statistical inference for the value function.

The rest of our paper is organized as follows. In Section 8.2, we introduce the statistical framework of causal survival analysis and transfer learning of optimal ITR. Section 9.1 develops the main methodology of learning the value function and associated optimal ITR. Section 9.2 establishes the asymptotic properties of the proposed value estimator. Extensive simulations are reported in Section 10.1 to demonstrate the empirical performance of the proposed method, followed by a real data application given in Section 10.2. The article concludes in Section 10.3 with a discussion of some remarks and future work. All proofs and additional results are provided in the Supplementary Material.

8.2 Statistical Framework

8.2.1 Causal survival analysis

Let X denote the p-dimensional vector of covariates that belongs to a covariate space $\mathcal{X} \subset \mathbb{R}^p$, $A \in \mathcal{A} = \{0, 1\}$ denote the binary treatment, and $T \in \mathbb{R}^+$ denote the *survival time* to the event of interest. In the presence of right censoring, the outcome T may not be observed. Let $C \in \mathbb{R}^+$ denote the censoring time and $\Delta = I\{T \leq C\}$ where $I\{\cdot\}$ is the indicator function. Let $U = \min\{T, C\}$ be the observed outcome, $N(t) = I\{U \leq t, \Delta = 1\}$ the counting process, and $Y(t) = I\{U \geq t\}$ the at-risk process.

We use the potential outcomes framework [Neyman, 1923, Rubin, 1974], where for $a \in \mathcal{A} = \{0, 1\}$, $T(a)$ is the survival time had the subject received treatment a. The common goal in causal survival analysis is to identify and estimate the counterfactual quantity $\mathbb{E}[y(T(a))]$ for some deterministic transformation function $y(\cdot)$. Such transformations include $y(T) = \min(T, L)$ for the RMST with some pre-specified maximal time horizon L, and $y(T) = I\{T \geq t\}$ for the survival probability at time t.

Under the standard assumptions (a) consistency: $T = T(A)$, (b) positivity: $Pr(A = a \,|\, X) > 0$ for every $a \in \mathcal{A}$ *almost surely*, (c) unconfoundedness: $A \perp\!\!\!\perp \{T(1), T(0)\} \,|\, X$, (d) conditionally independent censoring: $C \perp\!\!\!\perp \{T(1), T(0)\} \,|\, \{X, A\}$, we can nonparametrically identify $\mathbb{E}[y(T(a))]$ by the outcome regression (OR) formula or the inverse probability weighting (IPW) formula [Van der Laan and Robins, 2003].

8.2.2 ITR and value function

Without loss of generality, we assume that larger values of T are more desirable. Typically we aim to identify and estimate an ITR $d(x) : \mathcal{X} \to \mathcal{A}$, which is a mapping from the covariate space $\mathcal{X}$ to the treatment space $\mathcal{A} = \{0, 1\}$, that maximizes the

expected outcome in a counterfactual world had this ITR been implemented. Suppose $\mathcal{D}$ is the class of candidate ITRs of interest, then define the potential outcome $T(d)$ under any $d \in \mathcal{D}$ by $T(d) = d(X)T(1) + (1-d(X))T(0)$, and the value function [Manski, 2004] of d is defined by $V(d) = \mathbb{E}[y(T(d))]$. Then by maximizing $V(d)$ over $\mathcal{D}$, the optimal ITR is defined by $d^{\text{opt}} = \arg\max_{d\in\mathcal{D}} V(d)$. See Qian and Murphy [2011] for more details.

To estimate the value function, we can use the OR or IPW formulas, and also a doubly robust method [Bai et al., 2017]:

$$\begin{aligned} V_{DR}(d) =& \mathbb{E}\Bigg[\frac{I\{A = d(X)\}\Delta\, y(U)}{Pr(A = d(X)\,|\,X)S_C(U\,|\,A, X)} \\ &+ \left(1 - \frac{I\{A = d(X)\}}{Pr(A = d(X)\,|\,X)}\right)\mathbb{E}[y(T)\,|\,A = d(X), X] \\ &+ \frac{I\{A = d(X)\}}{Pr(A = d(X)\,|\,X)}\int_0^\infty \frac{\mathrm{d}M_C(u\,|\,A, X)}{S_C(u\,|\,A, X)}\mathbb{E}[y(T)\,|\,T \geq u, A, X]\Bigg], \end{aligned} \tag{8.1}$$

where $S_C(t\,|\,a, x) = Pr(C > t\,|\,A = a, X = x)$ is the conditional survival function for the censoring process, $\mathrm{d}M_C(u\,|\,A = a, X) = \mathrm{d}N_C(u) - Y(u)\mathrm{d}\Lambda_C(u\,|\,A = a, X)$ is the martingale increment for the censoring process, $N_C(u) = I\{U \leq u, \Delta = 0\}$ and $\Lambda_C(u\,|\,A = a, X) = -\log(S_C(u\,|\,A = a, X))$. The first term in (8.1) is the IPW formula, and the augmentation terms capture additional information from the subjects who do not receive treatment d, and who receive treatment d but are censored.

In (clinical) practice, it is usually desirable to consider a class of ITRs indexed by a Euclidean parameter $\eta = (\eta_1, \ldots, \eta_{p+1})^T \in \mathbb{R}^{p+1}$ for feasibility and interpretability. Let $V(\eta) = V(d_\eta)$. Throughout, we focus on such a class of linear ITRs:

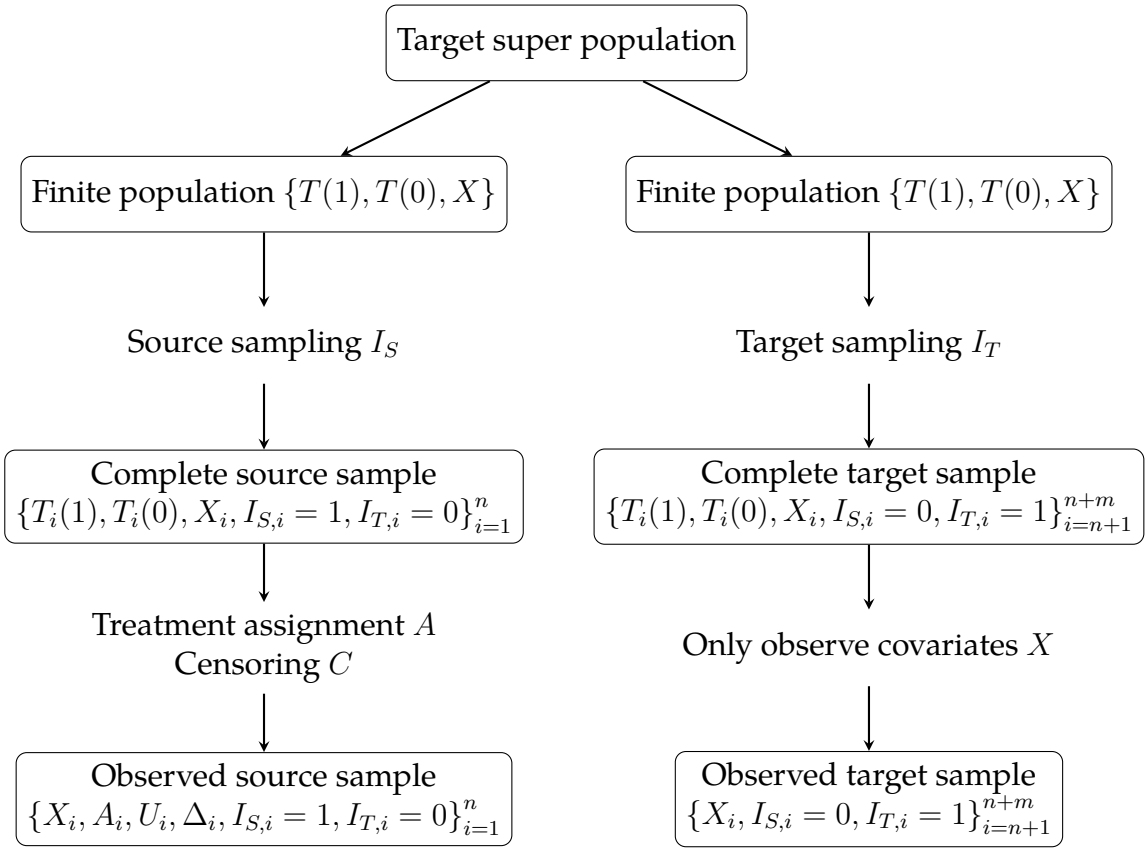
$$\mathcal{D}_\eta = \{d_\eta : d_\eta(X) = I\{\eta^T\tilde{X} \geq 0\}, |\eta_{p+1}| = 1\},$$

where $\tilde{X} = (1, X^T)^T$, and for identifiability we assume there exists a continuous covariate whose coefficient has absolute value one [Zhou et al., 2023a]; without loss of generality, we assume $|\eta_{p+1}| = 1$. Therefore, the population parameter η^* indexing the optimal ITR is $\eta^* = \arg\max_{\eta\in\{\eta\in\mathbb{R}^{p+1}:|\eta_{p+1}|=1\}} V(\eta)$, and the optimal value function is $V(\eta^*)$.

8.2.3 Transfer learning

The performance of such a learned ITR may suffer from a covariate shift in which the population distributions differ [Sugiyama and Kawanabe, 2012]. Instead of minimizing the worst-case risk, here we consider a super population framework. Suppose that a source sample of size n and a target sample of size m are sampled independently from the target super population with different mechanisms. Let I_S and I_T denote the indicator of sampling from source and target populations, respectively. A covariate shift means that $Pr(I_S = 1\,|\,X) \neq Pr(I_T = 1\,|\,X)$. In the source sample, independent and identically distributed (i.i.d.) data $\mathcal{O}_s = \{X_i, A_i, U_i, \Delta_i, I_{S,i} = 1, I_{T,i} = 0\}_{i=1}^n$ are observed from n subjects; in the target sample, it is common that only the covariates information is available, so i.i.d. data $\mathcal{O}_t = \{X_i, I_{S,i} = 0, I_{T,i} = 1\}_{i=n+1}^{n+m}$ are observed from m subjects. The sampling mechanism and data structure are illustrated in Figure 8.1.

Figure 8.1 – *Schematic of the data structure of the source and target samples within the target super population framework.*



In this framework, we assume that the source and target sampling mechanisms are independent, which holds if two separate studies are conducted independently by different research projects in different locations or in two separate time periods, and the target population is sufficiently large. In the context of combining the RCT and observational study, this framework corresponds to the *non-nested* study design [Dahabreh et al., 2021].

Remark 8.2.1. In the framework illustrated in Figure 8.1, we also assume the existence of the finite population of size N, which helps us clarify the sampling mechanism and identification strategy. The two separate finite populations exemplify the independence of the source and target sampling processes. We present the identification formulas in Section 9.1; however, we do not require N to be fixed and known. Equivalently, it is also possible to assume a pooled population consisting of a source population and a target population, and similar identification formulas can be proposed based on the density ratio of the two populations.

CHAPTER 9

TRANSFER LEARNING

9.1 Methodology

9.1.1 Identification and semiparametric efficiency

To identify the causal effects from the observed data, we make the following assumptions.

Assumption 9.1.1. (a) $T = T(A)$ almost surely. (b) $Pr(A = a \,|\, X, I_S = 1) > 0$ for every a almost surely. (c) $A \perp\!\!\!\perp \{T(1), T(0)\} \,|\, \{X, I_S = 1\}$. (d) $C \perp\!\!\!\perp \{T(1), T(0)\} \,|\, \{X, A, I_S = 1\}$.

Assumption 9.1.1 includes the standard assumptions as we have introduced in Section 8.2.1. Here we only assume them in the source population. Assumption 9.1.1(a) implies that the observed outcome is the potential outcome under the actual assigned treatment. Assumption 9.1.1(b) states that each subject has a positive probability of receiving both treatments. Assumption 9.1.1(c) requires that all confounding factors are measured so that treatment assignment is as good as random conditionally on X. Assumption 9.1.1(d) essentially states that the censoring process is non-informative conditionally on X. Furthermore, we require additional assumptions for the source and target populations.

Assumption 9.1.2 (Survival mean exchangeability)**.** $\mathbb{E}[y(T(a)) \,|\, X, I_S = 1] = \mathbb{E}[y(T(a)) \,|\, X]$ for every $a \in \mathcal{A}$.

Assumption 9.1.3 (Positivity of Source Inclusion)**.** $0 < Pr(I_S = 1 \,|\, X) < 1$ almost surely.

Assumption 9.1.4 (Known target design)**.** The target sample design weight $e(x) = \pi_T^{-1}(x) = 1/Pr(I_T = 1 \,|\, X = x)$ is known by design.

Assumption 9.1.2 is similar to the mean exchangeability over trial participation [Dahabreh et al., 2019], and is weaker than the ignorablility assumption [Stuart et al., 2011], i.e., $I_S \perp\!\!\!\perp \{T(1), T(0)\} \,|\, X$. Assumption 9.1.3 states that each subject has a positive probability to be included in the source sample, and implies adequate *overlap* of covariate distributions between the source and target populations. Assumption 9.1.4 is commonly assumed in the survey sampling literature; thus the design-weighted

target sample is representative of the target population. In an observational study with simple random sampling, we have $e(x) = N/m$, where N is the target population size.

Under this framework, we have the following key identity that for any $g(X)$

$$\mathbb{E}\left[\frac{I_S}{\pi_S(X)}g(X)\right] = \mathbb{E}[I_T\, e(X)g(X)] = \mathbb{E}[g(X)], \tag{9.1}$$

where $\pi_S(X) = Pr(I_S = 1 \mid X)$ is the sampling score.

Proposition 9.1.5 (Identification formulas)**.** *Under Assumptions 9.1.1 - 9.1.4, the value function $V(d)$ can be identified by the outcome regression formula:*

$$V(d) = \mathbb{E}[I_T\, e(X)\mathbb{E}[y(T) \mid A = d(X), X, I_S = 1]], \tag{9.2}$$

and the IPW formula:

$$V(d) = \mathbb{E}\left[\frac{I_S}{\pi_S(X)}\frac{I\{A = d(X)\}}{\pi_d(X)}\frac{\Delta\, y(U)}{S_C(U \mid A, X)}\right], \tag{9.3}$$

where $\pi_d(X) = d(X)\pi_A(X) + (1 - d(X))(1 - \pi_A(X))$ with the propensity score $\pi_A(X) = Pr(A = 1 \mid X, I_S = 1)$, and $S_C(t \mid a, x) = Pr(C > t \mid A = a, X = x, I_S = 1)$.

Based on the identification formulas (9.2) and (9.3), we can construct plug-in estimators for $V(d)$, using the sampling score $\pi_S(X)$ or design weights $e(X)$ to account for the sampling bias. By the identity (9.1), the design weights $I_T\, e(X)$ in the OR formula (9.2) with the target sample can also be replaced by the inverse of sampling score $I_S/\pi_S(X)$ using the source sample. However, these estimators are biased if the posited models are misspecified, and extreme weights from π_S, π_A and S_C usually lead to large variability. Therefore, we consider a more efficient and robust approach, motivated by the efficient influence function for $V(d)$.

Proposition 9.1.6. *Under Assumptions 9.1.1 - 9.1.4, the efficient influence function of $V(d)$ is*

$$\begin{aligned}\phi_d =& \frac{I_S}{\pi_S(X)}\frac{I\{A = d(X)\}}{\pi_d(X)}\frac{\Delta\, y(U)}{S_C(U \mid A, X)} - V(d) \\ &+ \left(I_T\, e(X) - \frac{I_S}{\pi_S(X)}\frac{I\{A = d(X)\}}{\pi_d(X)}\right)\mu(d(X), X) \\ &+ \frac{I_S}{\pi_S(X)}\frac{I\{A = d(X)\}}{\pi_d(X)}\int_0^\infty \frac{\mathrm{d}M_C(u \mid A, X)}{S_C(u \mid A, X)}Q(u, A, X).\end{aligned} \tag{9.4}$$

where $\mu(a, x) = \mathbb{E}[y(T) \mid A = a, X = x, I_S = 1]$ and $Q(u, a, x) = \mathbb{E}[y(T) \mid T \geq u, A = a, X = x, I_S = 1]$ [1].

The semiparametric EIF guides us in constructing efficient estimators combining the source and target samples. Compared to (8.1), this EIF captures additional covariates information from the target population via the outcome model and thus removes the sampling bias. An efficient estimation procedure is proposed in the next section, and

1. Note that $\mathbb{E}[y(T) \mid T \geq u, A, X] = -\int_u^\infty y(s)\,\mathrm{d}S(s \mid A, X)/S(u \mid A, X)$. For instance, when $y(T) = I\{T \geq t\}$, we have $\mathbb{E}[y(T) \mid T \geq u, A, X] = S(t \mid A, X)/S(u \mid A, X)$ for $u \leq t$.

we show that it enjoys the double robustness property, i.e., it is consistent if either the survival outcome models $\mu(a,x), Q(u,a,x)$ or the models of propensity score $\pi_A(x)$, sampling score $\pi_S(x)$ and censoring process $S_C(t\,|\,a,x)$ are correct. Moreover, this EIF is Neyman orthogonal in the sense discussed in Chernozhukov et al. [2018]. Therefore, a cross-fitting procedure is also proposed, allowing flexible machine learning methods for the nuisance parameters estimation, and $\sqrt{N}$ rate of convergence can be achieved.

9.1.2 An efficient and robust estimation procedure

In this section, we focus on estimating the survival function $S_d(t) = Pr(T(d) > t)$ as the value function under ITR d. Following the asymptotic linear characterization of survival estimands in Yang et al. [2021], our results are readily extended to a broad class of functionals of survival distributions. For instance, the value function of the RMST under ITR d is simply $\int_0^L S_d(t)\mathrm{d}t$.

Based on the EIF (9.4), we propose an estimator for the survival function

$$\begin{aligned}\hat{S}_d(t) =& \frac{1}{N}\sum_{i=1}^{N}\Bigg\{\frac{I_{S,i}}{\hat{\pi}_S(X_i)}\frac{I\{A_i = d(X_i)\}}{\hat{\pi}_d(X_i)}\frac{\Delta_i\, Y_i(t)}{\hat{S}_C(t\,|\,A_i,X_i)} \\ &+\left(I_{T,i}\, e(X_i) - \frac{I_{S,i}}{\hat{\pi}_S(X_i)}\frac{I\{A_i = d(X_i)\}}{\hat{\pi}_d(X_i)}\right)\hat{S}(t\,|\,A = d(X_i),X_i) \\ &+\frac{I_{S,i}}{\hat{\pi}_S(X_i)}\frac{I\{A_i = d(X_i)\}}{\hat{\pi}_d(X_i)}\int_0^\infty \frac{\hat{S}(t\,|\,A_i,X_i)\mathrm{d}\hat{M}_C(u\,|\,A_i,X_i)}{\hat{S}(u\,|\,A_i,X_i)\hat{S}_C(u\,|\,A_i,X_i)}\Bigg\},\end{aligned} \tag{9.5}$$

where $S(t\,|\,a,x) = Pr(T > t\,|\,A = a, X = x, I_S = 1)$ is the treatment-specific conditional survival function. We posit (semi)parametric models for the nuisance parameters. Let $\pi_A(X;\theta)$ be the posited propensity score model, for example, using logistic regression $\mathrm{logit}\{\pi_\mathrm{A}(\mathrm{X};\theta)\} = \theta^\mathrm{T}\tilde{\mathrm{X}}$, where $\mathrm{logit(x)} = \log\{\mathrm{x}/(1-\mathrm{x})\}$. We use the Cox proportional hazard model $\Lambda(t\,|\,A = a, X = x) = \Lambda_{0,a}(t)\exp(\beta_a^T x)$ to estimate the survival functions $S(t\,|\,a,x) = \exp\{-\Lambda(t\,|\,a,x)\}$ and the cumulative baseline hazard function $\Lambda_{0,a}(t) = \int_0^t \lambda_{0,a}(u)\mathrm{d}u$ can be estimated by the Breslow estimator [Breslow, 1972]. Similarly, we posit a Cox proportional hazard model for the censoring process $\Lambda_C(t\,|\,A = a, X = x) = \Lambda_{C0,a}(t)\exp(\alpha_a^T x)$, and the cumulative baseline hazard function $\Lambda_{C0,a}(t)$ is estimated by the Breslow estimator. The sampling score estimation is discussed in the next section.

Let $\hat{S}(t;\eta) = \hat{S}_{d_\eta}(t)$ be the estimated value function for the ITR class $\mathcal{D}_\eta$, then the optimal ITR is given by $d_{\hat{\eta}}(x)$, where $\hat{\eta} = \arg\max_\eta \hat{S}(t;\eta)$.

9.1.3 Calibration weighting

To correct the bias due to the covariate shift between populations, most existing methods directly model the sampling score [Cole and Stuart, 2010], i.e., inverse probability of sampling weighting (IPSW). However, the IPSW method requires the sampling score model to be correctly specified, and it could also be numerically unstable. Alternatively, we introduce the calibration weighting (CW) approach motivated by the identity (9.1), which is similar to the entropy balancing method [Hainmueller, 2012].

Let $\mathbf{g}(X)$ be a vector of functions of X to be calibrated, such as the moments, interactions, and non-linear transformations of X. Each subject i in the source sample is assigned a weight q_i by solving the following optimization task:

$$\min_{q1,\ldots,q_n} \sum_{i=1}^{n} q_i \log q_i, \tag{9.6}$$

$$\text{subject to } q_i \geq 0, \sum_{i=1}^{n} q_i = 1, \sum_{i=1}^{n} q_i \mathbf{g}(X_i) = \tilde{\mathbf{g}}, \tag{9.7}$$

where $\tilde{\mathbf{g}} = \sum_{i=n+1}^{n+m} e(X_i)\mathbf{g}(X_i) / \sum_{i=n+1}^{n+m} e(X_i)$ is a design-weighted estimate of $\mathbb{E}[\mathbf{g}(X)]$. The objective function (9.6) is the negative entropy of the calibration weights, which ensures that the empirical distribution of the weights is not too far away from the uniform, such that it minimizes the variability due to heterogeneous weights. The final balancing constraint in (9.7) calibrates the covariate distribution of the weighted source sample to the target population in terms of $\mathbf{g}(X)$. By introducing the Lagrange multiplier λ, the minimizer of the optimization task is $q_i = \exp\{\hat{\lambda}^T \mathbf{g}(X_i)\} / \sum_{i=1}^{n} \exp\{\hat{\lambda}^T \mathbf{g}(X_i)\}$, where $\hat{\lambda}$ solves the estimating equation $\sum_{i=1}^{n} \exp\{\lambda^T \mathbf{g}(X_i)\}\{\mathbf{g}(X_i) - \tilde{\mathbf{g}}\} = 0$. Since we only require specifying $\mathbf{g}(X)$, calibration weighting avoids explicitly modeling the sampling score and evades extreme weights.

Moreover, suppose that the sampling score follows a loglinear model $\pi_S(X; \lambda) = \exp\{\lambda^T \tilde{X}\}$, Lee et al. [2021, 2022] show that there is a direct correspondence between the calibration weights and the estimated sampling score, i.e., $q_i = \{N\pi_S(X_i; \hat{\lambda})\}^{-1} + o_p(N^{-1})$. We also note that if the fraction n/N is small, the loglinear model is close to the widely used logistic regression model; our simulation studies show the robustness of calibration weights.

Remark 9.1.7. Other objective functions can also be used for calibration weights estimation. Chu et al. [2022] considers a generic convex distance function $h(q)$ from the Cressie and Read family of discrepancies [Cressie and Read, 1984]. Thus the optimization task is $\min_{q1,\ldots,q_n} \sum_{i=1}^{n} h(q_i)$ under the constraints (9.7), and the correspondence between the sampling score model π_S and the objective function h has also been established.

9.1.4 Cross-fitting

Utilizing the Neyman orthogonality of EIF (9.4), we consider flexible machine learning methods for estimating the nuisance parameters, where we want to remain agnostic on modeling assumptions for the complex treatment assignment, survival, and censoring processes. There is extensive recent literature on nonparametric methods for heterogeneous treatment effect estimation with survival outcomes. Cui et al. [2020] extends the generalized random forests [Athey et al., 2019] to estimate heterogeneous treatment effects in a survival and observational setting. See Xu et al. [2022] for details and practical considerations. A description of the proposed cross-fitting procedure is given below [Schick, 1986, Chernozhukov et al., 2018]. Throughout, we use the subscript CF to denote the cross-fitted version.

Algorithm 1: Pseudo algorithm for the cross-fitting procedure

Step 1 Randomly split the datasets $\mathcal{O}_s$ and $\mathcal{O}_t$ respectively into K-folds with equal size such that $\mathcal{O}_s = \cup_{k=1}^{K} \mathcal{O}_{s,k}$, $\mathcal{O}_t = \cup_{k=1}^{K} \mathcal{O}_{t,k}$. For each $k \in \{1, \ldots, K\}$, let $\mathcal{O}_{s,k}^c = \mathcal{O}_s \backslash \mathcal{O}_{s,k}$, $\mathcal{O}_{t,k}^c = \mathcal{O}_s \backslash \mathcal{O}_{t,k}$.

Step 2 For each $k \in \{1, \ldots, K\}$, estimate the nuisance parameters only using data $\mathcal{O}_{s,k}^c$ and $\mathcal{O}_{t,k}^c$; then obtain an estimate of the value function $\hat{V}_{CF,k}(\eta)$ using data $\mathcal{O}_{s,k}$.

Step 3 Aggregate the estimates from K folds: $\hat{V}_{CF}(\eta) = \frac{1}{K} \sum_{k=1}^{K} \hat{V}_{CF,k}(\eta)$.

Step 4 The estimated optimal ITR is indexed by $\hat{\eta} = \arg\max_{\eta} \hat{V}_{CF}(\eta)$.

9.2 Asymptotic properties

In this section, we present the asymptotic properties of the proposed methods. To establish the asymptotic properties, we require the following assumptions.

Assumption 9.2.1. (i) The value function $V(\eta)$ is twice continuously differentiable in a neighborhood of η^*. (ii) There exists some constant $\delta_0 > 0$ such that $Pr(0 < |\tilde{X}^T \eta| < \delta) = O(\delta)$, where the big-$O$ term is uniform in $0 < \delta < \delta_0$.

Condition (i) is a standard regularity condition to establish uniform convergence. Similar margin conditions as (ii), which state that $Pr(0 < |\gamma(X)| < \delta) = O(\delta^{\alpha})$ [2], are often assumed in the literature of classification [Tsybakov, 2004, Audibert and Tsybakov, 2007], reinforcement learning [Farahmand, 2011, Hu et al., 2021] and optimal treatment regimes [Luedtke and van der Laan, 2016b, Luedtke and Chambaz, 2020], to guarantee a fast convergence rate. Note that $\alpha = 0$ imposes no restriction, which allows $\gamma(X) = 0$ almost surely, i.e., the challenging setting of exceptional laws where the optimal ITR is not uniquely defined [Robins, 2004, Robins and Rotnitzky, 2014], while the case $\alpha = 1$ is of particular interest and would hold if $\gamma(X)$ is absolutely continuous with bounded density.

Theorem 9.2.2. *Under Assumptions 9.1.1 - 9.2.1 and standard regularity conditions provided in the Supplementary Material, if either the survival outcome model, or the models of the propensity score, the sampling score and the censoring process are correct, we have that as* $N \to \infty$, (i) $\hat{S}(t;\eta) \to S(t;\eta)$ *for any* η *and* $0 < t \leq L$; (ii) $\sqrt{N}\left\{\hat{S}(t;\eta) - S(t;\eta)\right\}$ *converges weakly to a mean zero Gaussian process for any* η; (iii) $N^{1/3} \left\|\hat{\eta} - \eta^*\right\|_2 = O_p(1)$; (iv) $\sqrt{N}\left\{\hat{S}(t;\hat{\eta}) - S(t;\eta^*)\right\} \to \mathcal{N}(0, \sigma_{t,1}^2)$, *where* $\sigma_{t,1}$ *is given in the Supplementary Material.*

Next, to characterize the asymptotic behavior of the estimator with the nonparametric estimation of nuisance parameters, we assume the following consistency and convergence rate conditions of the nonparametric plug-in nuisance estimators.

Assumption 9.2.3. Assume the following convergences in probability: $\sup_{x \in \mathcal{X}} |\hat{\pi}_A(x) -$

2. Let $\gamma(X) = \mathbb{E}[T \mid A = 1, X] - \mathbb{E}[T \mid A = 0, X]$ denote the conditional average treatment effect, then the optimal ITR in an unrestricted class is given by $d(X) = I\{\gamma(X) > 0\}$.

$\pi_A(x)| \to 0$, $\sup_{x\in\mathcal{X}} |\hat{\pi}_S(x) - \pi_S(x)| \to 0$, and for $a = 0, 1$,

$$\sup_{x\in\mathcal{X},u\le h} |\hat{S}_C(u \mid a, x) - S_C(u \mid a, x)| \to 0, \sup_{x\in\mathcal{X},u\le h} \left| \frac{\hat{\lambda}_C(u \mid a, x)}{\hat{S}_C(u \mid a, x)} - \frac{\lambda_C(u \mid a, x)}{S_C(u \mid a, x)} \right| \to 0,$$
$$\sup_{x\in\mathcal{X}} |\hat{\mu}(a, x) - \mu(a, x)| \to 0, \sup_{x\in\mathcal{X},u\le h} |\hat{Q}(u, a, x) - Q(u, a, x)| \to 0;$$

and the following rates of convergence: $\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\hat{\pi}_A(x) - \pi_A(x)|\right] = o_p(n^{-1/4})$, $\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\hat{\pi}_S(x) - \pi_S(x)|\right] = o_p(n^{-1/4})$, and for $a = 0, 1$,

$$\sup_{u\le h} \mathbb{E}\left[\sup_{x\in\mathcal{X}} \left|\hat{S}_C(u \mid a, x) - S_C(u \mid a, x)\right|\right] = o_p(n^{-1/4}),$$
$$\sup_{u\le h} \mathbb{E}\left[\sup_{x\in\mathcal{X}} \left| \frac{\hat{\lambda}_C(u \mid a, x)}{\hat{S}_C(u \mid a, x)} - \frac{\lambda_C(u \mid a, x)}{S_C(u \mid a, x)} \right|\right] = o_p(n^{-1/4}),$$
$$\mathbb{E}\left[\sup_{x\in\mathcal{X}} |\hat{\mu}(a, x) - \mu(a, x)|\right] = o(n^{-1/4}), \sup_{u\le h} \mathbb{E}\left[\sup_{x\in\mathcal{X}} |\hat{Q}(u, a, x) - Q(u, a, x)|\right] = o(n^{-1/4}).$$

The rate conditions in Assumption 9.2.3 are generally assumed in the literature [Kennedy, 2022]. This rate can be achieved by many existing methods under certain structural assumptions on the nuisance parameters. Note that the nuisance parameters do not necessarily need to be estimated at the same rates $n^{-1/4}$ for our theorems to hold; it would suffice that the product of rates of any combination of two nuisance parameters is $n^{-1/2}$.

Theorem 9.2.4. *Under Assumptions 9.1.1 - 9.2.3, we have that as $N \to \infty$, (i) $\hat{S}_{CF}(t;\eta) \to S(t;\eta)$ for any η and $0 < t \le L$; (ii) $\sqrt{N}\left\{\hat{S}_{CF}(t;\eta) - S(t;\eta)\right\}$ converges weakly to a mean zero Gaussian process for any η; (iii) $N^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$; (iv) $\sqrt{N}\left\{\hat{S}_{CF}(t;\hat{\eta}) - S(t;\eta^*)\right\} \to \mathcal{N}(0, \sigma_{t,2}^2)$, where $\sigma_{t,2}$ is given in the Supplementary Material.*

Besides the survival functions, another common measure of particular interest in survival analysis is the RMST. Let $V_{\text{RMST}}(\eta) = \mathbb{E}[\min(T(d_\eta), L)]$. We present two corollaries.

Corollary 9.2.5. *Under Assumptions 9.1.1 - 9.2.1 and standard regularity conditions provided in the Supplementary material, if either the survival outcome model or the models of the propensity score, the censoring and sampling processes are correct, we have that as $N \to \infty$, (i) $\hat{V}_{RMST}(\eta) \to V_{RMST}(\eta)$ for any η; (ii) $N^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$; (iii) $\sqrt{N}\left\{\hat{V}_{RMST}(\hat{\eta}) - V_{RMST}(\eta^*)\right\} \to \mathcal{N}(0, \sigma_3^2)$, where σ_3 is given in the Supplementary Material.*

Corollary 9.2.6. *Under Assumptions 9.1.1 - 9.2.3, we have that as $N \to \infty$, (i) $\hat{V}_{RMST,CF}(\eta) \to V_{RMST}(\eta)$ for any η; (ii) $N^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$; (iii) $\sqrt{N}\left\{\hat{V}_{RMST,CF}(\hat{\eta}) - V_{RMST}(\eta^*)\right\} \to \mathcal{N}(0, \sigma_4^2)$, where σ_4 is given in the Supplementary Material..*

Finally, we show that when the covariate distributions of the source and target populations are the same, the semiparametric efficiency bounds of $\hat{V}_{DR}(\eta)$ and $\hat{V}_{CF}(\eta)$ are equal.

Theorem 9.2.7. *Under Assumptions 9.1.1 - 9.2.3, when the covariate distributions of the source and target populations are the same, both* $\sqrt{N}\{\hat{V}_{DR}(\eta) - V(\eta)\}$ *and* $\sqrt{N}\{\hat{V}_{CF}(\eta) - V(\eta)\}$ *are asymptotically normal with mean zero and same variance.*

Theorem 9.2.7 implies that when there is no covariate shift, our proposed estimator does not lose efficiency in comparison to the original double robust estimator since the augmentation term in EIF (9.4) from the target population, $I_T\, e(X)\mu(d(X), X)$, is asymptotically equal to this term evaluated on the source population in this case.

Moreover, when the covariate shift exists, we consider the optimal ITR d^{opt} without restriction on the ITR class.

Theorem 9.2.8. *Under Assumptions 9.1.1 - 9.2.3, If* $d^{opt} \in \mathcal{D}_\eta$*, i.e.,* $d^{opt} = d_{\eta^*}$*, both the maximizers of* $\hat{V}_{DR}(\eta)$ *and* $\hat{V}_{CF}(\eta)$ *converge to* η^**. However,* $\hat{V}_{DR}(\eta)$ *is a biased estimator of* $V(\eta)$*.*

Theorem 9.2.8 implies if the true optimal ITR belongs to the restricted ITR class $\mathcal{D}_\eta$, standard methods, without accounting for the covariate shift, are still able to recover the optimal ITR but fail to be consistent for the value function, due to the covariate shift. And we can only rely on the proposed method to draw valid inferences.

CHAPTER 10

NUMERICAL EXPERIMENTS AND DISCUSSION

10.1 Simulation

In this section, we investigate the finite-sample properties of our method through extensive numerical simulations [1].

Consider a target population of sample size $N = 2 \times 10^5$. The covariates $(X_1, X_2, X_3)^T$ are generated from a multivariate normal distribution with mean 0, unit variance with $corr(X_1, X_3) = 0.2$ and all other pairwise correlations equal to 0, and further truncated below -4 and above 4 to satisfy regularity conditions. The target sample is a random sample of size $m = 8000$ from the target population. The sampling score follows $\pi_S(X) = \text{expit}(-4.5 - 0.5X_1 - 0.5X_2 - 0.4X_3)$; thus the source sampling rate is around 1.6%, and the source sample size around $n = 3000$. The treatment assignment mechanism in the source sample follows $\pi_A(X) = \text{expit}(0.5 + 0.8X_1 - 0.5X_2)$.

The counterfactual survival times $T(a)$ are generated according to the hazard functions $\lambda(t \mid A = 0, X) = \exp(t) \cdot \exp(-2.5 - 1.5X_1 - X_2 - 0.7X_3)$ and $\lambda(t \mid A = 1, X) = \exp(t) \cdot \exp(-1 - X_1 - 0.9X_2 - X_3 - 2X_2^2 + X_1X_3)$. The censoring time C is generated according to the hazard functions $\lambda_C(t \mid A = 0, X) = 0.04 \exp(t) \cdot \exp(-1.6 + 0.8X_1 - 1.1X_2 - 0.7X_3)$ and $\lambda_C(t \mid A = 1, X) = 0.04 \exp(t) \cdot \exp(-1.8 - 0.8X_1 - 1.7X_2 - 1.4X_3)$. The resultant censoring rate is approximately 20%.

We consider the RMST with the maximal time horizon $L = 4$ as the value function. To evaluate the performance of different estimators for optimal ITRs, we compute the corresponding true value functions and percentages of correct decisions (PCD) for the target population. Specifically, we generate a large sample with size $\tilde{N} = 1 \times 10^5$ from the target population. The true value function of any ITR $d(\cdot\,;\,\eta)$ is computed by $V(\eta) = \tilde{N}^{-1} \sum_{i=1}^{\tilde{N}} \min\{d(X_i\,;\,\eta)T_i(1) + (1 - d(X_i\,;\,\eta))T_i(0), L\}$ and its associated PCD is computed by $1 - \tilde{N}^{-1} \sum_{i=1}^{\tilde{N}} |d(X_i\,;\,\eta^*) - d(X_i\,;\,\eta)|$, where $\eta^* = \arg\max_\eta V(\eta)$.

We compare the following estimators for the RMST $\hat{V}(\eta) = \int_0^L \hat{S}(t;\eta)\mathrm{d}t$:

1. The `R` code to replicate all results is available at https://github.com/panzhaooo/transfer-learning-survival-ITR.

— Naive: $\hat{S}^{\text{Naive}}(t;\eta) = \frac{1}{n}\sum_{i=1}^{n} \frac{I\{A_i=d(X_i)\}}{\hat{\pi}_d(X_i)} \frac{\Delta_i Y_i(t)}{\hat{S}_C(U \mid A,X)}$; IPW formula (9.3) without using the sampling score;

— IPSW: $\hat{S}^{\text{IPSW}}(t;\eta) = \frac{1}{n}\sum_{i=1}^{n} \frac{I_{S,i}}{\hat{\pi}_S(X_i)} \frac{I\{A_i=d(X_i)\}}{\hat{\pi}_d(X_i)} \frac{\Delta_i Y_i(t)}{\hat{S}_C(U \mid A,X)}$; IPW formula (9.3) where the sampling score is estimated via logistic regression;

— CW-IPW: $\hat{S}^{\text{CW-IPW}}(t;\eta) = \sum_{i=1}^{n} q_i \frac{I\{A_i=d(X_i)\}}{\hat{\pi}_d(X_i)} \frac{\Delta_i Y_i(t)}{\hat{S}_C(U \mid A,X)}$ IPW formula (9.3) where the sampling score is estimated by calibration weighting;

— CW-OR: $\hat{S}^{\text{CW-OR}}(t;\eta) = \sum_{i=1}^{n} q_i \hat{S}(t \mid A = d(X_i), X_i)$; OR formula (9.2) in combination with calibration weights by the identity (9.1);

— ORt: $\hat{S}^{\text{ORt}}(t;\eta) = \frac{1}{m}\sum_{i=n+1}^{n+m} \hat{S}(t \mid A = d(X_i), X_i)$; OR formula (9.2) evaluated on the target sample;

— ACW: augmented estimator (9.5), where the sampling score is estimated by calibration weighting.

Remark 10.1.1. Since the estimated value functions are non-convex and non-smooth, multiple local optimal may exist in the optimization task, and many derivatives-based algorithms do not work for this challenging setting. Here we utilize the genetic algorithm implemented in the R package `rgenoud` [Mebane Jr and Sekhon, 2011], which performs well in our numerical experiments. We refer to Mitchell [1998] for algorithmic details.

10.1.1 (Semi)parametric models

We first consider the setting where the nuisance parameters are estimated by posited (semi)parametric working models as introduced in Section 9.1.2. To assess the performance of these estimators under model misspecification, we consider four scenarios: (1) all models are correct, (2) only the survival outcome model is correct, (3) only the survival outcome model is wrong, (4) all models are wrong. For the wrong sampling model, the weights are estimated using calibration on e^{X_1}. The wrong propensity score model is fitted on e^{X_3}. The wrong Cox models for survival and censoring times are fitted on $(e^{X_1}, e^{X_2}, e^{X_3})^T$.

Figure 10.1 and Table 10.1 report the simulation results from 350 Monte Carlo replications. Variance is estimated by a bootstrap procedure with $B = 200$ bootstrap replicates. The proposed ACW estimator is unbiased in scenarios (1) - (3), and the 95% coverage probabilities approximately achieve the nominal level, which shows the double robustness property.

10.1.2 Flexible machine learning methods

When utilizing flexible ML methods, we construct the cross-fitted ACW estimator as introduced in Section 9.1.4. The data generation process is the same as above, except that the censoring time C is generated according to the hazard functions $\lambda_C(t \mid A = 0, X) = 0.2\exp(t) \cdot \exp(-1.6 + 0.8X_1 - 1.1X_2 - 0.7X_3)$ and $\lambda_C(t \mid A = 1, X) = 0.2\exp(t) \cdot \exp(-1.8 - 0.8X_1 - 1.7X_2 - 1.4X_3)$ which leads to an increased censoring rate of approximately 33%, so there are enough observations to get an accurate estimate of the censoring process. The propensity score is estimated by the generalized random

Figure 10.1 – *Boxplot of the estimated value, true value and PCD results of estimators under four model specification scenarios. O: survival outcome, S: sampling score, A: propensity score, C: censoring; T: True (correctly specified) model, W: Wrong (misspecified) model.*

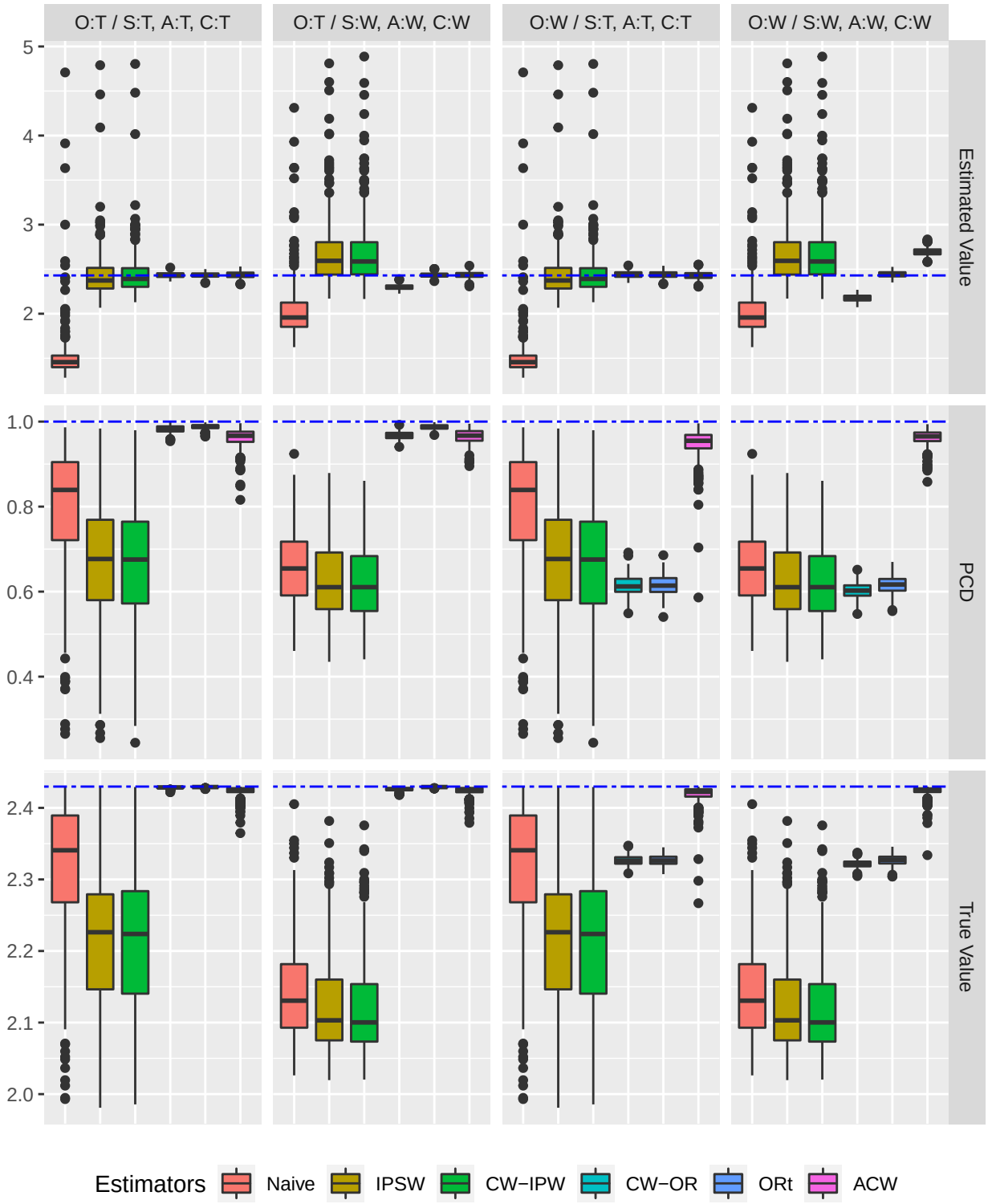


Table 10.1 – *Numerical results under four different model specification scenarios. Bias is the empirical bias of point estimates; SD is the empirical standard deviation of point estimates; SE is the average of bootstrap standard error estimates; CP is the empirical coverage probability of the* 95% *confidence intervals.*

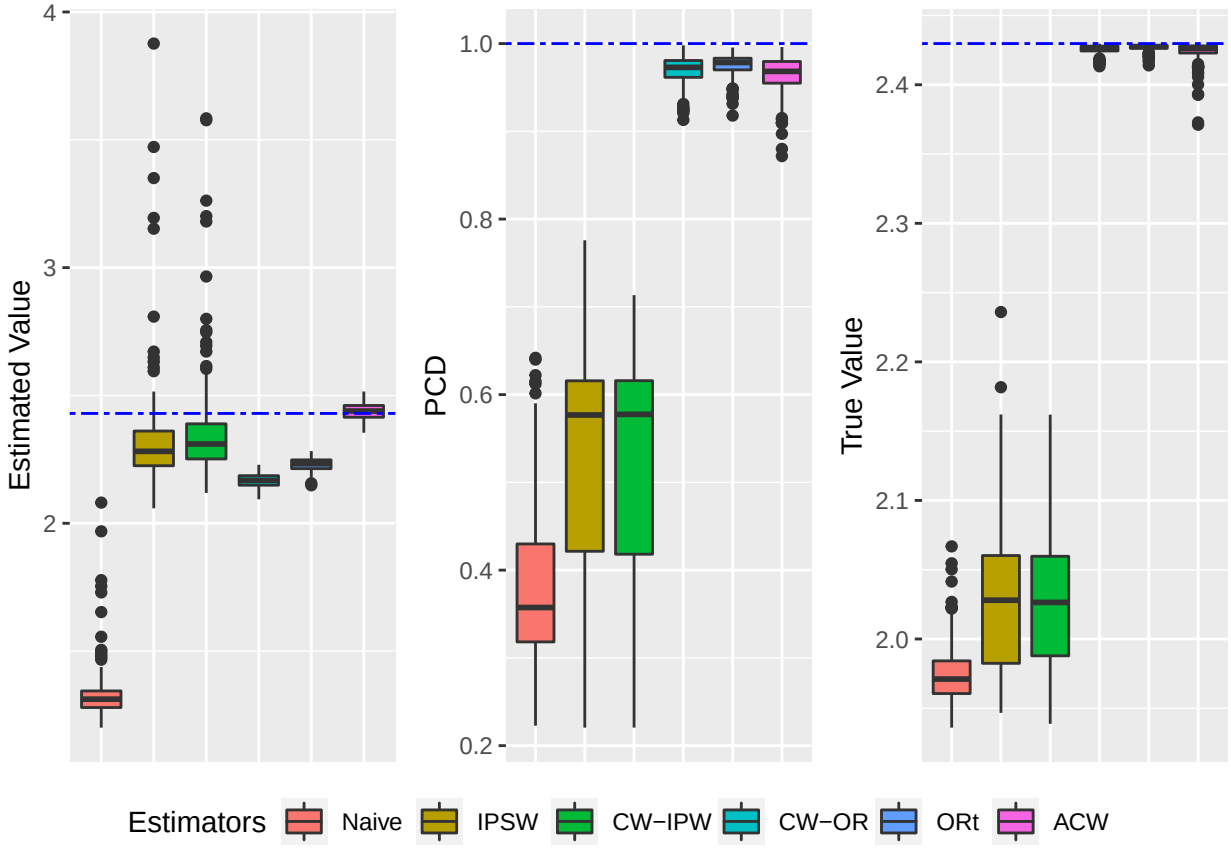
	Bias	SD	SE	CP(%)	Bias	SD	SE	CP(%)
	O:T / S:T, A:T, C:T				O:T / S:W, A:W, C:W			
Naive	-0.8801	0.4595	0.2189	7.43	-0.3528	0.5024	0.4598	37.43
IPSW	0.0185	0.3685	0.2562	87.14	0.3377	0.7144	0.6958	98.29
CW-IPW	0.0378	0.3701	0.2498	88.29	0.3406	0.7144	0.6957	97.71
CW-OR	0.0047	0.0273	0.0286	96.29	-0.1312	0.0269	0.0279	0.57
ORt	0.0041	0.0258	0.0262	95.14	0.0035	0.0258	0.0262	95.71
ACW	0.0070	0.0380	0.0369	94.29	0.0055	0.0316	0.0334	95.43
	O:W / S:T, A:T, C:T				O:W / S:W, A:W, C:W			
Naive	-0.8801	0.4595	0.2207	6.86	-0.3528	0.5024	0.5018	38.57
IPSW	0.0185	0.3685	0.2486	87.71	0.3377	0.7144	0.7586	99.14
CW-IPW	0.0378	0.3701	0.2418	88.86	0.3406	0.7144	0.7570	98.57
CW-OR	0.0103	0.0370	0.0362	92.29	-0.2551	0.0366	0.0391	0.00
ORt	0.0094	0.0365	0.0355	94.00	0.0115	0.0328	0.0355	95.71
ACW	-0.0010	0.0426	0.0419	93.14	0.2644	0.0422	0.0475	0.57

forest. The conditional survival and censoring functions are estimated by the random survival forest. The calibration weighting uses calibration on the first- and second-order moments of X.

First, we study the impact of sample sizes on the performance of the ML methods, and simulation results are given in the Supplementary Material. With a small sample size, the ACW estimator is largely biased, and the bias diminishes as the sample size increases.

Next, we compare the performance of different estimators with target population size $N = 6 \times 10^5$ and target sample size $m = 24000$. Figure 10.2 shows the simulation results from 200 Monte Carlo replications. The two IPW-based estimators are biased and perform poorly due to the large variability of weights. The two OR-based estimators have comparable performance as the ACW estimator in terms of PCD and true value function but still suffer from the overfitting bias. Only the ACW estimator is consistent and provides valid inferences.

Figure 10.2 – *Boxplots of the estimated value, true value, and PCD of different estimators using flexible ML methods.*



10.2 Real Data Analysis

In this section, to illustrate the proposed method, we study the sodium bicarbonate therapy for patients with severe metabolic acidaemia in the intensive care unit by leveraging the RCT data BICAR-ICU [Jaber et al., 2018] and the observational study (OS) data from Jung et al. [2011]. Specifically, we consider the BICAR-ICU data as the source sample and the observational study data as the target sample. The BICAR-ICU is a multi-center, open-label, randomized controlled, phase 3 trial between May 5, 2015, and May 7, 2017, which includes 387 adult patients admitted within 48 hours to the ICU with severe acidaemia. The prospective, multiple-center observational study was conducted over thirteen months in five ICUs, consisting of 193 consecutive patients who presented with severe acidemia within the first 24 hours of their ICU admission. Some heterogeneity exists between the two populations.

Both the RCT and OS datasets contain detailed measurements of ICU patients with severe acidaemia. Motivated by the clinical practice and existing work in the medical literature, we consider ITRs that depend on the following five variables: SEPSIS, AKIN, SOFA, SEX, and AGE. A detailed description of the data preprocessing and variable selection is given in the Supplementary Material. Table 10.2 summarizes the baseline characteristics of the two datasets. The baseline covariates distribution of the patients in the BICAR-ICU differs from the distribution in the observational study; specifically, the BICAR-ICU patients have higher SOFA scores and the more frequent presence of acute kidney injury and sepsis.

Table 10.2 – *Summary of baseline characteristics of the BICAR-ICU trial sample and the OS sample. Mean (standard deviation) for continuous and number (proportion) for the binary covariate.*

	SEPSIS	AKIN	SOFA	SEX	AGE
BICAR-ICU ($n = 387$)	236 (60.98%)	181 (46.77%)	10.12 (3.72)	237 (61.24%)	63.95 (14.41)
OS ($m = 193$)	99(51.30%)	75 (38.86%)	9.10 (4.54)	122 (63.21%)	62.73 (17.49)

We apply our proposed ACW estimator to learn the optimal ITR for the target population. The calibration weights are estimated based on the means of continuous covariates and the proportions of the binary covariates. The propensity score is estimated using a logistic regression model, and the Cox proportional hazard model is fitted for the survival outcome with all covariates. The censoring only occurred on the 28th day when the follow-up in ICU ends. We consider the class of linear ITRs that depend on all five variables:

$$\mathcal{D} = \{I\{\eta_1 + \eta_2 \text{SEPSIS} + \eta_3 \text{AKIN} + \eta_4 \text{SOFA} + \eta_5 \text{SEX} + \eta_6 \text{AGE} > 0\} : \eta_1, \ldots, \eta_6 \in \mathbb{R}, |\eta_6| = 1\},$$

with the aim to maximize the RMST within 28 days in ICU stay. The estimated parameter indexing the optimal ITR is $\hat{\eta}_{\text{ACW}} = (22.9, -36.1, 87.4, -9.8, 33.7, 1.0)^T$, which leads to an estimated value function $\hat{V}(\hat{\eta}_{\text{ACW}}) = 19.52$ days, with confidence interval $[17.74, 21.30]$ given by 200 bootstraps. In contrast, we also use the standard double robust method to estimate the optimal ITR for the RCT, indexed by $\hat{\eta}_{\text{DR.RCT}}$ which maximize the value function $\hat{V}_{\text{DR}}(\eta)$ in (8.1) with $y(T) = \min(T, 28)$. The estimated value function is $\hat{V}(\hat{\eta}_{\text{DR.RCT}}) = 15.37$ days for the target population.

10.3 Discussion

In this paper, we present an efficient and robust transfer learning framework for estimating optimal ITR with right-censored survival data that generalizes well to the target population. The proposed method can be improved or extended in several directions for future work. Construction and estimation of optimal ITRs for multiple decision points with censored survival data are challenging, taking into account the timing of censoring, events and decision points [Jiang et al., 2017, Hager et al., 2018], e.g., using a reinforcement learning method [Cho et al., 2020]. Furthermore, besides the class of ITRs indexed by a Euclidean parameter, it may be possible to consider other classes of ITRs, such as tree or list-based ITRs. The current work focus on value functions in the form $V(d) = \mathbb{E}[y(T(d))]$ and can also be modified in case of optimizing certain easy-to-interpret quantile criteria, which does not require specifying an outcome regression model and is robust for heavy-tailed distributions [Zhou et al., 2023a]. And relaxing the restrictive assumptions such as positivity [Yang and Ding, 2018a, Jin et al., 2022] and unconfoundedness [Cui and Tchetgen Tchetgen, 2021, Qi et al., 2021] for learning optimal ITRs is also a fruitful direction.

BIBLIOGRAPHY

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

Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91 (434):444–455, 1996.

Peter M Aronow and Allison Carnegie. Beyond late: Estimation of the average treatment effect with an instrumental variable. *Political Analysis*, 21(4):492–506, 2013.

Susan Athey and Guido W Imbens. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

Susan Athey and Guido W Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, 31(2):3–32, 2017.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2), 2019.

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.

Xiaofei Bai, Anastasios A Tsiatis, Wenbin Lu, and Rui Song. Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime data analysis*, 23(4):585–604, 2017.

Aurélien Bibaut, Nathan Kallus, Maria Dimakopoulou, Antoine Chambaz, and Mark van Der Laan. Risk minimization from adaptively collected data: Guarantees for supervised and policy learning. *Advances in neural information processing systems*, 34: 19261–19273, 2021.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Zach Branson, Edward H Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Causal effect estimation after propensity score trimming with continuous treatments. *arXiv preprint arXiv:2309.00706*, 2023.

Norman E Breslow. Contribution to discussion of paper by dr cox. *J. Roy. Statist. Soc., Ser. B*, 34:216–217, 1972.

Zongwu Cai, Mitali Das, Huaiyu Xiong, and Xizhi Wu. Functional coefficient instrumental variables models. *Journal of Econometrics*, 133(1):207–241, 2006.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE, 2009.

David Card. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160, 2001.

David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, 84:772–793, 1994.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. doi: 10.1111/ectj.12097.

Hunyong Cho, Shannon T Holloway, David J Couper, and Michael R Kosorok. Multistage optimal dynamic treatment regimes for survival outcomes with dependent censoring. *arXiv preprint arXiv:2012.03294*, 2020.

Jianing Chu, Wenbin Lu, and Shu Yang. Targeted optimal treatment regime learning using summary statistics. *arXiv preprint arXiv:2201.06229*, 2022.

Jianing Chu, Shu Yang, and Wenbin Lu. Multiply robust off-policy evaluation and learning under truncation by death. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 6195–6227. PMLR, 23–29 Jul 2023.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7321–7331. Curran Associates, Inc., 2020.

Stephen R Cole and Elizabeth A Stuart. Generalizing evidence from randomized clinical trials to target populations: the actg 320 trial. *American journal of epidemiology*, 172(1):107–115, 2010.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

Mayleen Cortez, Matthew Eichhorn, and Christina Yu. Staggered rollout designs enable causal inference under interference without network knowledge. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural*

Information Processing Systems, volume 35, pages 7437–7449. Curran Associates, Inc., 2022.

Noel Cressie and Timothy RC Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464, 1984.

Yifan Cui and Eric Tchetgen Tchetgen. A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. *Journal of the American Statistical Association*, 116(533):162–173, 2021.

Yifan Cui, Ruoqing Zhu, and Michael Kosorok. Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic journal of statistics*, 11(2):3927, 2017.

Yifan Cui, Michael R Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv preprint arXiv:2001.09887*, 2020.

Yifan Cui, Michael R Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2): 179–211, 2023a.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, pages 1–12, 2023b.

Issa J Dahabreh and Miguel A Hernán. Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology*, 34(8):719–722, 2019.

Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

Issa J Dahabreh, Sebastien JP A Haneuse, James M Robins, Sarah E Robertson, Ashley L Buchanan, Elizabeth A Stuart, and Miguel A Hernán. Study designs for extending causal inferences from a randomized trial to a target population. *American journal of epidemiology*, 190(8):1632–1642, 2021.

Clément De Chaisemartin and Xavier d'Haultfoeuille. Fuzzy differences-in-differences. *The Review of Economic Studies*, 85(2):999–1028, 2018.

Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *arXiv preprint arXiv:2102.11904*, 2021.

Iván Díaz, Oleksandr Savenkov, and Karla Ballman. Targeted learning ensembles for optimal individualized treatment rules with time-to-event outcomes. *Biometrika*, 105 (3):723–738, 2018.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29:485–511, 2014.

Esther Duflo. Schooling and labor market consequences of school construction in

indonesia: Evidence from an unusual policy experiment. *American economic review*, 91(4):795–813, 2001.

Oliver Dukes, David Richardson, Zach Shahn, and Eric Tchetgen Tchetgen. Semiparametric bespoke instrumental variables. *arXiv preprint arXiv:2204.04119*, 2022.

Ashkan Ertefaie and Robert L Strawderman. Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105(4):963–977, 2018.

Ethan X Fang, Zhaoran Wang, and Lan Wang. Fairness-oriented learning for optimal individualized treatment rules. *Journal of the American Statistical Association*, pages 1–14, 2022.

Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. *Advances in Neural Information Processing Systems*, 24, 2011.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Cambridge University Press, 2021.

Yair Goldberg and Michael R Kosorok. Q-learning with censored data. *Annals of statistics*, 40(1):529, 2012.

Lin Gui and Victor Veitch. Causal estimation for text data with (apparent) overlap violations. In *International Conference on Learning Representations*, 2023.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118, 2021.

Rebecca Hager, Anastasios A Tsiatis, and Marie Davidian. Optimal two-stage dynamic treatment regimes from a classification perspective with censored survival data. *Biometrics*, 74(4):1180–1192, 2018.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20 (1):25–46, 2012.

Sebastian Haneuse and Andrea Rotnitzky. Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine*, 32(30):5260–5277, 2013.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Miguel A Hernán and James M Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, pages 360–372, 2006.

Yichun Hu, Nathan Kallus, and Masatoshi Uehara. Fast rates for the regret of offline reinforcement learning. *arXiv preprint arXiv:2102.00479*, 2021.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Fast rates for contextual linear optimization. *Management Science*, 68(6):4236–4245, 2022.

Kosuke Imai and David A van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99 (467):854–866, 2004.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica: journal of the Econometric Society*, pages 467–475, 1994.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Samir Jaber, Catherine Paugam, Emmanuel Futier, Jean-Yves Lefrant, Sigismond Lasocki, Thomas Lescot, Julien Pottecher, Alexandre Demoule, Martine Ferrandiere, Karim Asehnoune, et al. Sodium bicarbonate therapy for patients with severe metabolic acidaemia in the intensive care unit (bicar-icu): a multicentre, open-label, randomised controlled, phase 3 trial. *The Lancet*, 392(10141):31–40, 2018.

Zeyang Jia, Eli Ben-Michael, and Kosuke Imai. Bayesian safe policy learning with chance constrained optimization: Application to military security assessment during the vietnam war. *arXiv preprint arXiv:2307.08840*, 2023.

Runchao Jiang, Wenbin Lu, Rui Song, and Marie Davidian. On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1165–1185, 2017.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning "without" overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6):e2214889120, 2023.

Boris Jung, Thomas Rimmele, Charlotte Le Goff, Gérald Chanques, Philippe Corne, Olivier Jonquet, Laurent Muller, Jean-Yves Lefrant, Christophe Guervilly, Laurent Papazian, et al. Severe metabolic or mixed acidemia on intensive care unit admission: incidence, prognosis and administration of buffer therapy. a prospective, multiple-center study. *Critical Care*, 15(5):1–9, 2011.

Nathan Kallus. More efficient policy learning via optimal retargeting. *Journal of the American Statistical Association*, 116(534):646–658, 2021.

Nathan Kallus and Miruna Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6037–6060. PMLR, 2023.

Nathan Kallus and Masatoshi Uehara. Efficient evaluation of natural stochastic policies in offline reinforcement learning. *arXiv preprint arXiv:2006.03886*, 2020.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *Advances in neural information processing systems*, 31, 2018.

Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pages 10598–10632. PMLR, 2022.

Luke J Keele, Dylan S Small, Jesse Y Hsu, and Colin B Fogarty. Patterns of effects and sensitivity analysis for differences-in-differences. *arXiv preprint arXiv:1901.01869*, 2019.

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

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Edward H. Kennedy, Sivaraman Balakrishnan, and Max G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48 (4):2008 – 2030, 2020. doi: 10.1214/19-AOS1874. URL https://doi.org/10.1214/19-AOS1874.

Samir Khan, Martin Saveski, and Johan Ugander. Off-policy evaluation beyond overlap: partial identification through smoothness. *arXiv preprint arXiv:2305.11812*, 2023.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*, volume 61. Springer, 2008.

Michael R Kosorok and Eric B Laber. Precision medicine. *Annual review of statistics and its application*, 6:263, 2019.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Eric B Laber and Ying-Qi Zhao. Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):501–514, 2015.

Eric B Laber, Kristin A Linn, and Leonard A Stefanski. Interactive model building for q-learning. *Biometrika*, 101(4):831–847, 2014.

Carolin Lawrence, Artem Sokolov, and Stefan Riezler. Counterfactual learning from bandit feedback under deterministic logging: A case study in statistical machine translation. *arXiv preprint arXiv:1707.09118*, 2017.

Michael Lechner et al. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3):165–224, 2011.

Dasom Lee, Shu Yang, Lin Dong, Xiaofei Wang, Donglin Zeng, and Jianwen Cai. Improving trial generalizability using observational studies. *Biometrics*, 2021.

Dasom Lee, Shu Yang, and Xiaofei Wang. Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population. *Journal of causal inference*, 10(1):415–440, 2022.

Lihua Lei, Roshni Sahoo, and Stefan Wager. Policy learning under biased sample selection. *arXiv preprint arXiv:2304.11735*, 2023.

Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu, and Peng Wu. Trustworthy policy learning under the counterfactual no-harm criterion. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 20575–20598. PMLR, 23–29 Jul 2023.

Kristin A Linn, Eric B Laber, and Leonard A Stefanski. Interactive q-learning for quantiles. *Journal of the American Statistical Association*, 112(518):638–649, 2017.

Yi Liu, Huiyue Li, Yunji Zhou, and Roland Matsouaka. Average treatment effect on the treated, under lack of positivity. *arXiv preprint arXiv:2309.01334*, 2023.

Alexander Luedtke and Antoine Chambaz. Performance guarantees for policy learning. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(3):2162–2188, 2020.

Alexander R Luedtke and Mark J van der Laan. Optimal individualized treatments in resource-limited settings. *The international journal of biostatistics*, 12(1):283–303, 2016a.

Alexander R Luedtke and Mark J van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016b.

Alexander R Luedtke and Mark J van der Laan. Super-learning of an optimal dynamic treatment rule. *The international journal of biostatistics*, 12(1):305–332, 2016c.

Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

Walter R Mebane Jr and Jasjeet S Sekhon. Genetic optimization using derivatives: the rgenoud package for r. *Journal of Statistical Software*, 42:1–26, 2011.

Melanie Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

Weibin Mo, Zhengling Qi, and Yufeng Liu. Learning optimal distributionally robust individualized treatment rules. *Journal of the American Statistical Association*, 116 (534):659–674, 2021.

Iván Díaz Muñoz and Mark van Der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

Susan A Murphy, Mark J van der Laan, James M Robins, and Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *Journal of the American Statistical Association*, 116(533):392–409, 2021.

Elizabeth L Ogburn, Andrea Rotnitzky, and James M Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 77(2):373, 2015.

Chan Park and Eric Tchetgen Tchetgen. A universal difference-in-differences approach for causal inference, 2023.

Judea Pearl. *Causality*. Cambridge university press, 2009.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

Michael JD Powell. *A direct search optimization method that models the objective and constraint functions by linear interpolation*. Springer, 1994.

Hongming Pu and Bo Zhang. Estimating optimal treatment rules with an instrumental variable: A partial identification learning approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):318–345, 2021.

Zhengling Qi, Rui Miao, and Xiaoke Zhang. Proximal learning for individualized treatment regimes under unmeasured confounding. *arXiv preprint arXiv:2105.01187*, 2021.

Zhengling Qi, Rui Miao, and Xiaoke Zhang. Proximal learning for individualized treatment regimes under unmeasured confounding. *Journal of the American Statistical Association*, pages 1–14, 2023.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

Hongxiang Qiu, Marco Carone, Ekaterina Sadikova, Maria Petukhova, Ronald C Kessler, and Alex Luedtke. Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, 116(533): 174–191, 2021.

Sonja A Rasmussen, Muin J Khoury, and Carlos Del Rio. Precision public health as a key tool in the covid-19 response. *JAMA*, 324(10):933–934, 2020.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128 (30):2013, 2013.

James Robins and Andrea Rotnitzky. Discussion of "Dynamic treatment regimes: Technical challenges and applications". *Electronic Journal of Statistics*, 8(1):1273 – 1289, 2014.

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

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pages 550–560, 2000.

Paul R Rosenbaum. *Observational Studies*. Springer, 2002.

Jonathan Roth, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 2023.

Daniel B Rubin and Mark J van der Laan. Statistical issues and limitations in personalized medicine research with clinical trials. *The international journal of biostatistics*, 8 (1):18, 2012.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Roshni Sahoo, Lihua Lei, and Stefan Wager. Learning from a biased sample. *arXiv preprint arXiv:2209.01754*, 2022.

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

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, pages 1139–1151, 1986.

Phillip J Schulte, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):640, 2014.

Jasjeet S Sekhon and Walter R Mebane. Genetic optimization using derivatives. *Political Analysis*, 7:187–210, 1998.

Alexander Shapiro. Asymptotic analysis of stochastic programs. *Annals of Operations Research*, 30:169–186, 1991.

Eli Sherman, David Arbour, and Ilya Shpitser. General identification of dynamic treatment regimes under interference. In *International Conference on Artificial Intelligence and Statistics*, pages 3917–3927. PMLR, 2020.

Chengchun Shi, Alin Fan, Rui Song, and Wenbin Lu. High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of statistics*, 46(3):925, 2018.

Chengchun Shi, Jin Zhu, Shen Ye, Shikai Luo, Hongtu Zhu, and Rui Song. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, pages 1–12, 2022.

Dylan S Small. Protocols for observational studies: Methods and open problems. *arXiv preprint arXiv:2403.19807*, 2024.

Tamar Sofer, David B Richardson, Elena Colicino, Joel Schwartz, and Eric J Tchetgen Tchetgen. On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(3):348, 2016.

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, John N Clore, et al. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

Liangjun Su, Irina Murtazashvili, and Aman Ullah. Local linear gmm estimation of functional coefficient iv models with an application to estimating the rate of return to schooling. *Journal of Business & Economic Statistics*, 31(2):184–207, 2013.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Zhiqiang Tan. Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association*, 101(476):1607–1618, 2006.

Jin Tian. Identifying dynamic sequential plans. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008.

Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. Designing reinforcement learning algorithms for digital interventions: Pre-implementation guidelines. *arXiv preprint arXiv:2206.03944*, 2022.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Anastasios A Tsiatis, Marie Davidian, Shannon T Holloway, and Eric B Laber. *Dynamic treatment regimes: Statistical methods for precision medicine*. CRC press, 2019.

Alexander B Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Mark J van der Laan and Maya L Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *The international journal of biostatistics*, 3 (1), 2007.

Mark J van der Laan and James M Robins. *Unified methods for censored longitudinal data and causality*. Springer, 2003.

Mark J Van der Laan and James M Robins. *Unified methods for censored longitudinal data and causality*, volume 5. Springer, 2003.

Mark J van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer New York, 1996.

Francis Vella. Gender roles and human capital investment: The relationship between traditional attitudes and female labour market performance. *Economica*, pages 191–211, 1994.

Davide Viviano. Policy targeting under network interference. *arXiv preprint arXiv:1906.10258*, 2019.

Tat-Thang Vo, Ting Ye, Ashkan Ertefaie, Samrat Roy, James Flory, Sean Hennessy, Stijn Vansteelandt, and Dylan S Small. Structural mean models for instrumented difference-in-differences. *arXiv preprint arXiv:2209.10339*, 2022.

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

Linbo Wang and Eric Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):531–550, 2018.

Zengri Wang and Thomas A Louis. Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika*, 90 (4):765–775, 2003.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.

Waverly Wei, Yuqing Zhou, Zeyu Zheng, and Jingshen Wang. Inference on the best policies with many covariates. *Journal of Econometrics*, page 105460, 2023. ISSN 0304-4076.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pages 1–25, 1982.

Lili Wu and Shu Yang. Transfer learning of individualized treatment rules from experimental to real-world data. *Journal of Computation and Graphical Statistics*, page doi.org/10.1080/10618600.2022.2141752, 2022.

Yang Xu, Jin Zhu, Chengchun Shi, Shikai Luo, and Rui Song. An instrumental variable approach to confounded off-policy evaluation. In *Proceedings of the 40th International*

Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 38848–38880, 2023.

Yizhe Xu, Nikolaos Ignatiadis, Erik Sverdrup, Scott Fleming, Stefan Wager, and Nigam Shah. Treatment heterogeneity with survival outcomes. *arXiv preprint arXiv:2207.07758*, 2022.

S. Yang and P. Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):1–7, 2018a.

Shu Yang and Peng Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493, 03 2018b. URL https://doi.org/10.1093/biomet/asy008.

Shu Yang, Yilong Zhang, Guanghan Frank Liu, and Qian Guan. SMIM: A unified framework of survival sensitivity analysis using multiple imputation and martingale. *Biometrics*, page doi: 10.1111/biom.13555, 2021.

Ting Ye, Luke Keele, Raiden Hasegawa, and Dylan S Small. A negative correlation strategy for bracketing in difference-in-differences. *arXiv preprint arXiv:2006.02423*, 2020.

Ting Ye, Ashkan Ertefaie, James Flory, Sean Hennessy, and Dylan S Small. Instrumented difference-in-differences. *Biometrics*, 2022.

Jessica G Young, Miguel A Hernán, and James M Robins. 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, 2014.

Christina Lee Yu, Edoardo M Airoldi, Christian Borgs, and Jennifer T Chayes. Estimating the total treatment effect in randomized experiments with unknown network structure. *Proceedings of the National Academy of Sciences*, 119(44):e2208975119, 2022.

Jakub Závada, Eric Hoste, Rodrigo Cartin-Ceba, Paolo Calzavacca, Ognjen Gajic, Gilles Clermont, Rinaldo Bellomo, John A Kellum, and AKI6 investigators. A comparison of three methods to estimate baseline creatinine for rifle classification. *Nephrology Dialysis Transplantation*, 25(12):3911–3918, 2010.

Baqun Zhang, Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1): 103–114, 2012a.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012b.

Bo Zhang, Jordan Weiss, Dylan S Small, and Qingyuan Zhao. Selecting and ranking individualized treatment rules with unmeasured confounding. *Journal of the American Statistical Association*, 116(533):295–308, 2021.

Yichi Zhang, Eric B Laber, Anastasios Tsiatis, and Marie Davidian. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, 71(4): 895–904, 2015.

Yichi Zhang, Eric B Laber, Marie Davidian, and Anastasios A Tsiatis. Interpretable dynamic treatment regimes. *Journal of the American Statistical Association*, 113(524): 1541–1549, 2018a.

Yuqian Zhang, Abhishek Chakrabortty, and Jelena Bradic. Semi-supervised causal inference: Generalizable and double robust inference for average treatment effects under selection bias with decaying overlap. *arXiv preprint arXiv:2305.12789*, 2023.

Zhongheng Zhang, Carlie Zhu, Lei Mo, and Yucai Hong. Effectiveness of sodium bicarbonate infusion on mortality in septic patients with metabolic acidosis. *Intensive care medicine*, 44(11):1888–1895, 2018b.

Pan Zhao, Julie Josse, and Shu Yang. Efficient and robust transfer learning of optimal individualized treatment regimes with right-censored survival data. *arXiv preprint arXiv:2301.05491*, 2023.

Ying-Qi Zhao, Donglin Zeng, Catherine M Tangen, and Michael L Leblanc. Robustifying trial-derived optimal treatment rules for a target population. *Electronic journal of statistics*, 13(1):1717, 2019.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Wenjing Zheng and Mark J. van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. Technical Report Working Paper 273, U.C. Berkeley Division of Biostatistics Working Paper Series, November 2010. URL https://biostats.bepress.com/ucbbiostat/paper273.

Yu Zhou, Lan Wang, Rui Song, and Tuoyi Zhao. Transformation-invariant learning of optimal individualized decision rules with time-to-event outcomes. *Journal of the American Statistical Association*, 118(544):2632–2644, 2023a.

Zhengyuan Zhou, Susan Athey, and Stefan Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023b.

Part V

SUPPLEMENTARY MATERIAL

CHAPTER A

APPENDIX TO PART II

A.1 Directed acyclic graphs

In this section, we present the directed acyclic graphs (DAGs) in Figures A.1 and A.2 illustrating the causal structure of the proposed instrumented DiD. The IV Z is associated with the trend in treatment $A_1 - A_0$, is independent of the unmeasured confounders U_0, U_1, cannot have direct effect on the trend in outcome $Y_1 - Y_0$, and does not modofy the treatment effect. But in comparison to a standard IV, here Z is allowed to have a direct effect on the outcomes Y_0, Y_1, as illustrated by the edges $Z \to Y_0$ and $Z \to Y_1$ in Figure A.2.

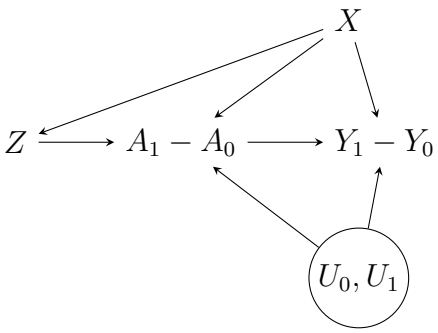


Figure A.1 – *DAG for instrumented DiD on the trend scale.*

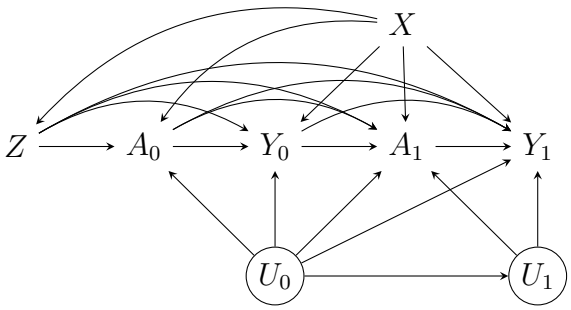


Figure A.2 – *DAG for instrumented DiD over two time points.*

A.2 Proof of Theorem 2.3.8

In this section, we provide a proof of Theorem 2.3.8 for completeness. Similar proof can be found at Ye et al. [2022].

We first note that

$$
\begin{aligned}
&\delta_Y(X) = \mu_Y(1,1,X) - \mu_Y(0,1,X) - \mu_Y(1,0,X) + \mu_Y(0,0,X) \\
&= \sum_{z=0,1} (2z-1)(E[Y \mid T=1, Z=z, X] - E[Y \mid T=0, Z=z, X]) \\
&= \sum_{z=0,1} (2z-1)(E[Y_1(A_1(z)) \mid T=1, Z=z, X] - E[Y_0(A_0(z)) \mid T=0, Z=z, X]) \\
&= \sum_{z=0,1} (2z-1)(E[Y_1(A_1(z)) \mid Z=z, X] - E[Y_0(A_0(z)) \mid Z=z, X]) \\
&= \sum_{z=0,1} (2z-1)E[Y_1(A_1(z)) - Y_0(A_0(z)) \mid Z=z, X] \\
&= \sum_{z=0,1} (2z-1)E[A_1(z)Y_1(1) + (1-A_1(z))Y_1(0) - A_0(z)Y_0(1) - (1-A_0(z))Y_0(0) \mid Z=z, X] \\
&= \sum_{z=0,1} (2z-1)E[A_1(z)(Y_1(1) - Y_1(0)) - A_0(z)(Y_0(1) - Y_0(0)) + Y_1(0) - Y_0(0) \mid Z=z, X] \\
&= \sum_{z=0,1} (2z-1)(E[A_1(z)(Y_1(1) - Y_1(0)) \mid X] - E[A_0(z)(Y_0(1) - Y_0(0)) \mid X] + E[Y_1(0) - Y_0(0) \mid X]) \\
&= E[(A_1(1) - A_1(0))(Y_1(1) - Y_1(0)) \mid X] - E[(A_0(1) - A_0(0))(Y_0(1) - Y_0(0)) \mid X] \\
&= E[A_1(1) - A_1(0) \mid X]E[Y_1(1) - Y_1(0) \mid X] - E[A_0(1) - A_0(0) \mid X]E[Y_0(1) - Y_0(0) \mid X] \\
&= E[A_1(1) - A_1(0) - A_0(1) + A_0(0) \mid X]\tau(X).
\end{aligned}
$$

Then note that

$$
\begin{aligned}
&\delta_A(X) = \mu_A(1,1,X) - \mu_A(0,1,X) - \mu_A(1,0,X) + \mu_A(0,0,X) \\
&= \sum_{z=0,1} (2z-1)(E[A \mid T=1, Z=z, X] - E[A \mid T=0, Z=z, X]) \\
&= \sum_{z=0,1} (2z-1)(E[A_1(z) \mid T=1, Z=z, X] - E[A_0(z) \mid T=0, Z=z, X]) \\
&= E[A_1(1) - A_1(0) - A_0(1) + A_0(0) \mid X].
\end{aligned}
$$

Hence we have that $\delta_Y(X) = \delta_A(X)\tau(X)$. That is, the CATE $\tau(X)$ can be identified by $\delta_Y(X)/\delta_A(X)$. It follows that the optimal policy is nonparametrically identified by

$$
\arg\max_{d \in \mathcal{D}} E[\tau(X)d(X)] = \arg\max_{d \in \mathcal{D}} E\left[\frac{\delta_Y(X)}{\delta_A(X)} d(X)\right],
$$

which completes the proof.

A.3 Proof of Theorem 2.3.9

In this section, we prove our first novel identification results of the optimal policy.

First we note that

$$
\begin{aligned}
&E\left[\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right] \\
&= E\left[\sum_{a=0,1}\frac{(2Z-1)(2T-1)(2a-1)Y_T(a)I\{A=a\}I\{d(X)=a\}}{\pi(T,Z,X)\delta_A(X)}\right] \\
&= E\left[\sum_{a=0,1}\frac{(2Z-1)(2T-1)(2a-1)E[Y_T(a)\mid X,U]I\{A=a\}I\{d(X)=a\}}{\pi(T,Z,X)\delta_A(X)}\right] \\
&= E\left[\sum_{a=0,1}\frac{(2Z-1)(2T-1)(2a-1)E[Y_T(a)\mid X,U]Pr(A=a\mid X,U,T,Z)I\{d(X)=a\}}{\pi(T,Z,X)\delta_A(X)}\right] \\
&= E\left[\frac{Pr(A=1\mid X,U,T=1,Z=1)I\{d(X)=1\}E[Y_1(1)\mid X,U]}{\delta_A(X)}\right] \\
&\quad - E\left[\frac{Pr(A=1\mid X,U,T=0,Z=1)I\{d(X)=1\}E[Y_0(1)\mid X,U]}{\delta_A(X)}\right] \\
&\quad - E\left[\frac{Pr(A=1\mid X,U,T=1,Z=0)I\{d(X)=1\}E[Y_1(1)\mid X,U]}{\delta_A(X)}\right] \\
&\quad + E\left[\frac{Pr(A=1\mid X,U,T=0,Z=0)I\{d(X)=1\}E[Y_0(1)\mid X,U]}{\delta_A(X)}\right] \\
&\quad - E\left[\frac{Pr(A=0\mid X,U,T=1,Z=1)I\{d(X)=0\}E[Y_1(0)\mid X,U]}{\delta_A(X)}\right] \\
&\quad + E\left[\frac{Pr(A=0\mid X,U,T=0,Z=1)I\{d(X)=0\}E[Y_0(0)\mid X,U]}{\delta_A(X)}\right] \\
&\quad + E\left[\frac{Pr(A=0\mid X,U,T=1,Z=0)I\{d(X)=0\}E[Y_1(0)\mid X,U]}{\delta_A(X)}\right] \\
&\quad - E\left[\frac{Pr(A=0\mid X,U,T=0,Z=0)I\{d(X)=0\}E[Y_0(0)\mid X,U]}{\delta_A(X)}\right] \\
&= E\left[\frac{[Pr(A=1\mid X,U,T=1,Z=1)-Pr(A=1\mid X,U,T=1,Z=0)]I\{d(X)=1\}E[Y_1(1)\mid X,U]}{\delta_A(X)}\right] \\
&\quad + E\left[\frac{[Pr(A=1\mid X,U,T=1,Z=1)-Pr(A=1\mid X,U,T=1,Z=0)]I\{d(X)=0\}E[Y_1(0)\mid X,U]}{\delta_A(X)}\right] \\
&\quad - E\left[\frac{[Pr(A=1\mid X,U,T=0,Z=1)-Pr(A=1\mid X,U,T=0,Z=0)]I\{d(X)=1\}E[Y_0(1)\mid X,U]}{\delta_A(X)}\right] \\
&\quad - E\left[\frac{[Pr(A=1\mid X,U,T=0,Z=1)-Pr(A=1\mid X,U,T=0,Z=0)]I\{d(X)=0\}E[Y_0(0)\mid X,U]}{\delta_A(X)}\right]
\end{aligned}
$$

Since we have that for $t=0,1$,

$$
\begin{aligned}
&I\{d(X)=1\}E[Y_t(1)\mid X,U]+I\{d(X)=0\}E[Y_t(0)\mid X,U] \\
&= d(X)(E[Y_t(1)\mid X,U]-E[Y_t(0)\mid X,U])+E[Y_t(0)\mid X,U] \\
&= d(X)\tau(X)+E[Y_t(0)\mid X,U],
\end{aligned}
$$

we continue by Assumption 2.3.7 that

$$E\left[\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right]$$
$$= E[d(X)\tau(X)] + E[\nu(X,U)],$$

where the second term $E[\nu(X,U)]$ does not depend on the policy d. That is,

$$\arg\max_{d\in\mathcal{D}} E\left[\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right] = \arg\max_{d\in\mathcal{D}} E[\tau(X)d(X)],$$

which completes the proof.

A.4 Proof of Theorem 2.3.10

In this section, we prove our second novel identification results of the optimal policy.

First we note that

$$E\left[\frac{(2T-1)YI\{Z=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right]$$
$$= E\left[\sum_{a=0,1}\frac{(2T-1)I\{Z=d(X)\}Y_T(a)I\{A=a\}}{\pi(T,Z,X)\delta_A(X)}\right]$$
$$= E\left[\sum_{a=0,1}\frac{(2T-1)I\{Z=d(X)\}E[Y_T(a)\mid X,U]Pr(A=a\mid X,U,T,Z)}{\pi(T,Z,X)\delta_A(X)}\right]$$
$$= E\left[\frac{Pr(A=1\mid X,U,T=1,Z=1)I\{d(X)=1\}E[Y_1(1)\mid X,U]}{\delta_A(X)}\right]$$
$$- E\left[\frac{Pr(A=1\mid X,U,T=0,Z=1)I\{d(X)=1\}E[Y_0(1)\mid X,U]}{\delta_A(X)}\right]$$
$$+ E\left[\frac{Pr(A=1\mid X,U,T=1,Z=0)I\{d(X)=0\}E[Y_1(1)\mid X,U]}{\delta_A(X)}\right]$$
$$- E\left[\frac{Pr(A=1\mid X,U,T=0,Z=0)I\{d(X)=0\}E[Y_0(1)\mid X,U]}{\delta_A(X)}\right]$$
$$+ E\left[\frac{Pr(A=0\mid X,U,T=1,Z=1)I\{d(X)=1\}E[Y_1(0)\mid X,U]}{\delta_A(X)}\right]$$
$$- E\left[\frac{Pr(A=0\mid X,U,T=0,Z=1)I\{d(X)=1\}E[Y_0(0)\mid X,U]}{\delta_A(X)}\right]$$
$$+ E\left[\frac{Pr(A=0\mid X,U,T=1,Z=0)I\{d(X)=0\}E[Y_1(0)\mid X,U]}{\delta_A(X)}\right]$$
$$- E\left[\frac{Pr(A=0\mid X,U,T=0,Z=0)I\{d(X)=0\}E[Y_0(0)\mid X,U]}{\delta_A(X)}\right]$$
$$= E\left[\frac{[Pr(A=1\mid X,U,T=1,Z=1)-Pr(A=1\mid X,U,T=1,Z=0)]I\{d(X)=1\}E[Y_1(1)\mid X,U}{\delta_A(X)}\right]$$

$$
\begin{aligned}
&+ E\left[\frac{[Pr(A=1 \mid X,U,T=1,Z=1) - Pr(A=1 \mid X,U,T=1,Z=0)]I\{d(X)=0\}E[Y_1(0) \mid X,U]}{\delta_A(X)}\right] \\
&- E\left[\frac{[Pr(A=1 \mid X,U,T=0,Z=1) - Pr(A=1 \mid X,U,T=0,Z=0)]I\{d(X)=1\}E[Y_0(1) \mid X,U]}{\delta_A(X)}\right] \\
&- E\left[\frac{[Pr(A=1 \mid X,U,T=0,Z=1) - Pr(A=1 \mid X,U,T=0,Z=0)]I\{d(X)=0\}E[Y_0(0) \mid X,U]}{\delta_A(X)}\right] \\
&+ E\left[\frac{Pr(A=1 \mid X,U,T=1,Z=0)E[Y_1(1) \mid X,U] + Pr(A=0 \mid X,U,T=1,Z=1)E[Y_1(0) \mid X,U]}{\delta_A(X)}\right] \\
&- E\left[\frac{Pr(A=1 \mid X,U,T=0,Z=0)E[Y_0(1) \mid X,U] + Pr(A=0 \mid X,U,T=0,Z=1)E[Y_0(0) \mid X,U]}{\delta_A(X)}\right]
\end{aligned}
$$

Then by the same arguments as in Section A.3, we have that

$$
E\left[\frac{(2T-1)YI\{Z=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right] = E[d(X)\tau(X)] + E[\tilde{\nu}(X,U)],
$$

where the second term does not depend on the policy d. That is,

$$
\arg\max_{d\in\mathcal{D}} E\left[\frac{(2T-1)YI\{Z=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right] = \arg\max_{d\in\mathcal{D}} E[\tau(X)d(X)],
$$

which completes the proof.

A.5 Proof of Theorem 3.1.1

In this section, we prove our identification results of the optimal policy using the efficient influence functions.

First we note that

$$
\begin{aligned}
&E[W_1 I\{A=d(X)\}] \\
&= \frac{1}{2}E[W_1(2I\{A=d(X)\}-1)] + \frac{1}{2}E[W_1] \\
&= \frac{1}{2}E[W_1(2A-1)(2d(X)-1)] + \frac{1}{2}E[W_1] \\
&= \frac{1}{2}E[\Delta(O)(2d(X)-1)] + \frac{1}{2}E[W_1] \\
&= E[\Delta(O)d(X)] + \frac{1}{2}E[W_1 - \Delta(O)] \\
&= E[\tau(X)d(X)] + \frac{1}{2}E[W_1 - \Delta(O)],
\end{aligned}
$$

where the last equality holds under the union model $\mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$. The proof of the multiple robustness is omitted since it simply follows the same arguments of Theorem 1 in Ye et al. [2022].

We also note that

$$\begin{aligned}
&E[W_2 I\{Z = d(X)\}] \\
&= \frac{1}{2}E[W_2(2I\{Z = d(X)\} - 1)] + \frac{1}{2}E[W_2] \\
&= \frac{1}{2}E[W_2(2Z - 1)(2d(X) - 1)] + \frac{1}{2}E[W_2] \\
&= \frac{1}{2}E[\Delta(O)(2d(X) - 1)] + \frac{1}{2}E[W_2] \\
&= E[\Delta(O)d(X)] + \frac{1}{2}E[W_2 - \Delta(O)] \\
&= E[\tau(X)d(X)] + \frac{1}{2}E[W_2 - \Delta(O)],
\end{aligned}$$

where the last equality holds under the union model $\mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$.

A.6 A locally efficient and multiply robust estimator

In this section, we present the semiparametric efficiency results for our proposed IPW formula:

$$\Psi(P) = E\left[\frac{(2Z-1)(2T-1)(2A-1)YI\{A = d(X)\}}{\pi(T, Z, X)\delta_A(X)}\right].$$

We first characterize the efficient influence function, and then propose the multiply robust estimator.

Theorem A.6.1. *The efficient influence function of* $\Psi(P)$ *is*

$$\begin{aligned}
\phi_P = &\frac{(2Z-1)(2T-1)(2A-1)YI\{A = d(X)\}}{\pi(T, Z, X)\delta_A(X)} \\
&- \frac{(2Z-1)(2T-1)E[(2A-1)YI\{A = d(X)\} \mid T, Z, X]}{\pi(T, Z, X)\delta_A(X)} + \gamma(X) \\
&- \frac{(2Z-1)(2T-1)(A - \mu_A(T, Z, X))\gamma(X)}{\pi(T, Z, X)\delta_A(X)} - \Psi(P),
\end{aligned}$$

where $\gamma(x) = \sum_{t,z}(2z-1)(2t-1)E[(2A-1)YI\{A = d(X)\} \mid T = t, Z = z, X = x]/\delta_A(x)$.

By Theorem A.6.1, we conclude that the optimal policy is nonparametrically identified by $\arg\max_{\mathcal{D}} \psi_P$, where

$$\begin{aligned}
\psi_P = E\Bigg[&\frac{(2Z-1)(2T-1)(2A-1)YI\{A = d(X)\}}{\pi(T, Z, X)\delta_A(X)} \\
&- \frac{(2Z-1)(2T-1)E[(2A-1)YI\{A = d(X)\} \mid T, Z, X]}{\pi(T, Z, X)\delta_A(X)} + \gamma(X) \\
&- \frac{(2Z-1)(2T-1)(A - \mu_A(T, Z, X))\gamma(X)}{\pi(T, Z, X)\delta_A(X)}\Bigg].
\end{aligned}$$

In Theorem A.6.2, we show the multiple robustness of the above formula under models:
$\tilde{\mathcal{M}}_1$: models for $\pi(t,z,x)$ and $\delta_A(x)$ are correct;
$\tilde{\mathcal{M}}_2$: models for $\pi(t,z,x)$ and $\gamma(x)$ are correct;
$\tilde{\mathcal{M}}_3$: models for $\mu_A(t,z,x)$, $\gamma(x)$ and $\nu(t,z,x)$ are correct, where $\nu(t,z,x) = E[(2A-1)YI\{A=d(X)\} \mid T=t, Z=z, X=x]$.

Theorem A.6.2. *Under standard regularity conditions, we have that*

$$P_n\psi(\hat{P}) = P_n\Bigg[\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\hat{\pi}(T,Z,X)\hat{\delta}_A(X)} - \frac{(2Z-1)(2T-1)\hat{E}[(2A-1)YI\{A=d(X)\} \mid T,Z,X]}{\hat{\pi}(T,Z,X)\hat{\delta}_A(X)} + \hat{\gamma}(X) - \frac{(2Z-1)(2T-1)(A-\hat{\mu}_A(T,Z,X))\hat{\gamma}(X)}{\hat{\pi}(T,Z,X)\hat{\delta}_A(X)}\Bigg]$$

is a consistent and asymptotically normal estimator of $\Psi(P)$ under the union model $\tilde{\mathcal{M}}_1 \cup \tilde{\mathcal{M}}_2 \cup \tilde{\mathcal{M}}_3$. Furthermore, it is locally efficient under the intersection model $\tilde{\mathcal{M}}_1 \cap \tilde{\mathcal{M}}_2 \cap \tilde{\mathcal{M}}_3$.

Despite the fact that we characterize the efficient influence function and propose a multiply robust estimator, note that it is not straightforward to posit models for $\gamma(x)$ and $\nu(t,z,x)$.

A.7 Proof of Theorem A.6.1 and Theorem A.6.2

We first prove Theorem A.6.1 by deriving the efficient influence function.

For a given distribution P in the nonparametric statistical model $\mathcal{M}$, we let p denote the density of P with respect to some dominating measure ν. For all bounded $h \in L_2(P)$, define the parametric submodel $p_\epsilon = (1+\epsilon h)p$, which is valid for small enough ϵ and has score h at $\epsilon = 0$.

We study the following statistical functional

$$\Psi(P) = E_P\left[\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right],$$

and would establish that $\Psi(P)$ is pathwise differentiable with respect to $\mathcal{M}$ at P with efficient influence function ϕ_P if we have that for any $P \in \mathcal{M}$

$$\frac{\partial}{\partial\epsilon}\Psi(P_\epsilon)\Big|_{\epsilon=0} = \int \phi_P(o)h(o)dP(o).$$

We denote $\pi_\epsilon(t,z,x) = E_{P_\epsilon}[I\{T=t,Z=z\} \mid X=x]$, $\delta_{A,\epsilon}(x) = \mu_{A,\epsilon}(1,1,x) - \mu_{A,\epsilon}(0,1,x) - \mu_{A,\epsilon}(1,0,x) + \mu_{A,\epsilon}(0,0,x)$, $\mu_{A,\epsilon}(t,z,x) = E_{P_\epsilon}[A \mid T=t,Z=z,X=x]$,

$S = \partial \log p_\epsilon / \partial \epsilon$, and compute

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\Psi(P_\epsilon)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} E_{P_\epsilon}\left[\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi_\epsilon(T,Z,X)\delta_{A,\epsilon}(X)}\right]\Big|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon} E_P\left[(1+\epsilon S)\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi_\epsilon(T,Z,X)\delta_{A,\epsilon}(X)}\right]\Big|_{\epsilon=0} \\
&= E_P\left[S\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi(T,Z,X)\delta_A(X)}\right] \\
&\quad - E_P\left[\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi^2(T,Z,X)\delta_A^2(X)}\left(\delta_A(X)\frac{\partial}{\partial \epsilon}\pi_\epsilon(T,Z,X)\Big|_{\epsilon=0}\right.\right. \\
&\quad \left.\left. +\pi(T,Z,X)\frac{\partial}{\partial \epsilon}\delta_{A,\epsilon}(X)\Big|_{\epsilon=0}\right)\right].
\end{aligned}$$

Then we need to compute

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\pi_\epsilon(t,z,X)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} E_{P_\epsilon}[I\{T=t,Z=z\} \mid X]\Big|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon}\frac{\pi(t,z,X)+\epsilon E_P[SI\{T=t,Z=z\} \mid X]}{1+\epsilon E_P[S \mid X]}\Big|_{\epsilon=0} \\
&= E_P[SI\{T=t,Z=z\} \mid X] - \pi(t,z,X)E_P[S \mid X],
\end{aligned}$$

$$\frac{\partial}{\partial \epsilon}\delta_{A,\epsilon}(X)\Big|_{\epsilon=0} = \frac{\partial}{\partial \epsilon}\{\mu_{A,\epsilon}(1,1,X) - \mu_{A,\epsilon}(0,1,X) - \mu_{A,\epsilon}(1,0,X) + \mu_{A,\epsilon}(0,0,X)\}\Big|_{\epsilon=0},$$

and

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\mu_{A,\epsilon}(t,z,X)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} E_{P_\epsilon}[A \mid T=t,Z=z,X]\Big|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon}\frac{\mu_A(t,z,X)+\epsilon E_P[SA \mid T=t,Z=z,X]}{1+\epsilon E_P[S \mid T=t,Z=z,X]}\Big|_{\epsilon=0} \\
&= E_P[SA \mid T=t,Z=z,X] - \mu_A(t,z,X)E_P[S \mid T=t,Z=z,X] \\
&= E_P\left[S\frac{(A-\mu_A(t,z,X))I\{T=t,Z=z\}}{\pi(t,z,X)} \mid X\right].
\end{aligned}$$

In summary, we obtain the efficient influence function

$$\begin{aligned}
\phi_P = &\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi(T,Z,X)\delta_A(X)} \\
&- \frac{(2Z-1)(2T-1)E[(2A-1)YI\{A=d(X)\} \mid T,Z,X]}{\pi(T,Z,X)\delta_A(X)} + \gamma(X) \\
&- \frac{(2Z-1)(2T-1)(A-\mu_A(T,Z,X))\gamma(X)}{\pi(T,Z,X)\delta_A(X)} - \Psi(P),
\end{aligned}$$

which completes the proof of Theorem A.6.1.

Next, we prove Theorem A.6.2 by verifying the multiple robustness property.

We first note the facts that $\mu_A(T,Z,X) = \mu_A(0,0,x) + Z(\mu_A(0,1,x) - \mu_A(0,0,x)) + T(\mu_A(1,0,x) - \mu_A(0,0,x)) + TZ\delta_A(X)$, $\nu(T,Z,X) = \nu(0,0,x) + Z(\nu(0,1,x) - \nu(0,0,x)) + T(\nu(1,0,x) - \nu(0,0,x)) + TZ\delta_A(X)$, $E[(2Z-1)(2T-1)/\pi(T,Z,X) \mid T,X] = E[(2Z-1)(2T-1)/\pi(T,Z,X) \mid Z,X] = 0$, and $E[\gamma(X)] = \Psi(P)$.

If $\tilde{\mathcal{M}}_1$ is correctly specified, we have that

$$\begin{aligned} E[\phi_P(O)] &= E\left[\frac{(2Z-1)(2T-1)(A-\mu_A(T,Z,X))\gamma(X)}{\pi(T,Z,X)\delta_A(X)}\right] \\ &= E\left[\frac{(2Z-1)(2T-1)\gamma(X)}{\pi(T,Z,X)\delta_A(X)}(A-\mu_A(T,Z,X))\right] = 0. \end{aligned}$$

If $\tilde{\mathcal{M}}_2$ is correctly specified, we have that

$$\begin{aligned} E[\phi_P(O)] = E\Bigg[&\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi(T,Z,X)\delta_A(X)} \\ &-\frac{(2Z-1)(2T-1)E[(2A-1)YI\{A=d(X)\} \mid T,Z,X]}{\pi(T,Z,X)\delta_A(X)} \\ &-\frac{(2Z-1)(2T-1)(A-\mu_A(T,Z,X))\gamma(X)}{\pi(T,Z,X)\delta_A(X)}\Bigg] = 0. \end{aligned}$$

If $\tilde{\mathcal{M}}_3$ is correctly specified, we have that

$$\begin{aligned} E[\phi_P(O)] = E\Bigg[&\frac{(2Z-1)(2T-1)(2A-1)YI\{A=d(X)\}}{\pi(T,Z,X)\delta_A(X)} \\ &-\frac{(2Z-1)(2T-1)E[(2A-1)YI\{A=d(X)\} \mid T,Z,X]}{\pi(T,Z,X)\delta_A(X)}\Bigg] = 0, \end{aligned}$$

which completes the proof.

A.8 Proof of Theorem 3.2.2

We study the following maximization problem:

$$\begin{aligned} \hat{\eta} = \arg\max_{\eta\in\mathbb{H}} \frac{1}{n}\sum_{i=1}^{n}\Bigg(&\frac{\delta_Y(X_i;\hat{\beta})}{\delta_A(X_i;\hat{\alpha})} + \frac{(2Z_i-1)(2T_i-1)}{\pi(T_i,Z_i,X_i;\hat{\theta})\delta_A(X_i;\hat{\alpha})}\Big\{Y_i - \mu_Y(T_i,Z_i,X_i;\hat{\beta}) \\ &-\frac{\delta_Y(X_i;\hat{\beta})}{\delta_A(X_i;\hat{\alpha})}(A_i - \mu_A(T_i,Z_i,X_i;\hat{\alpha}))\Big\}\Bigg) d(X_i;\eta), \end{aligned}$$

where $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\theta}$ are estimated by posited parametric models. We let $\hat{M}(\eta)$ denote the estimated objective function above, i.e. $\hat{\eta} = \arg\max_{\eta\in\mathbb{H}} \hat{M}(\eta)$.

Under standard regularity conditions, we have that

$$\sqrt{n}(\hat{\alpha}-\alpha^*)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{\alpha,i}+o_p(1),$$
$$\sqrt{n}(\hat{\beta}-\beta^*)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{\beta,i}+o_p(1),$$
$$\sqrt{n}(\hat{\theta}-\theta^*)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{\theta,i}+o_p(1),$$

where α^*, β^* and θ^* are the probability limits, ϕ_α, ϕ_β and ϕ_θ are the influence functions.

Now we start our proof which has three main parts as follows.

PART 1. First we note that, by the multiple robustness property, the strong law of large numbers and uniform consistency, $\hat{M}(\eta)=M(\eta)+o_p(1)$.

We denote

$$M_n^*(\eta)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\delta_Y(X_i;\beta^*)}{\delta_A(X_i;\alpha^*)}+\frac{(2Z_i-1)(2T_i-1)}{\pi(T_i,Z_i,X_i;\theta^*)\delta_A(X_i;\alpha^*)}\{Y_i-\mu_Y(T_i,Z_i,X_i;\beta^*)\right.$$
$$\left.-\frac{\delta_Y(X_i;\beta^*)}{\delta_A(X_i;\alpha^*)}(A_i-\mu_A(T_i,Z_i,X_i;\alpha^*))\right\}\Bigg)d(X_i;\eta),$$

and apply the Taylor expansion on $\hat{M}(\eta)$ at $(\alpha^*,\beta^*,\theta^*)$,

$$\hat{M}(\eta)=M_n^*(\eta)+H_{\alpha^*}^{\mathrm{T}}(\hat{\alpha}-\alpha^*)+H_{\beta^*}^{\mathrm{T}}(\hat{\beta}-\beta^*)+H_{\theta^*}^{\mathrm{T}}(\hat{\theta}-\theta^*)+o_p(n^{-1/2}),$$

where $H_{\alpha^*}=\lim_{n\to\infty}\partial\hat{M}(\eta)/\partial\alpha|_{\alpha=\alpha^*}$, $H_{\beta^*}=\lim_{n\to\infty}\partial\hat{M}(\eta)/\partial\beta|_{\beta=\beta^*}$, and $H_{\theta^*}=\lim_{n\to\infty}\partial\hat{M}(\eta)/\partial\theta|_{\theta=\theta^*}$.

Hence, we obtain that

$$\sqrt{n}\left\{\hat{M}(\eta)-M(\eta)\right\}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(M_n^*(\eta)-M(\eta)+H_{\alpha^*}^{\mathrm{T}}\phi_{\alpha,i}+H_{\beta^*}^{\mathrm{T}}\phi_{\beta,i}+H_{\theta^*}^{\mathrm{T}}\phi_{\theta,i}\right)+o_p(1). \tag{A.1}$$

PART 2. We prove that $n^{1/3}\|\hat{\eta}-\eta^*\|_2=O_p(1)$.

First we note that, by Condition 1 (iii), $M(\eta)$ is twice continuously differentiable at a neighborhood of η^*. In PART 1, we show that $\hat{M}(\eta)=M(\eta)+o_p(1),\forall\eta$. Since $\hat{\eta}$ maximizes $\hat{M}(\eta)$, we have that $\hat{M}(\hat{\eta})\geq\sup_\eta\hat{M}(\eta)$; thus by the Argmax theorem, we obtain that $\hat{\eta}\xrightarrow{p}\eta^*$ as $n\to\infty$.

Then we apply Theorem 14.4 (Rate of convergence) of Kosorok [2008] to establish the $n^{-1/3}$ rate of convergence of $\hat{\eta}$, and need to find the suitable rate that satisfies three conditions below.

Condition 1 For every η in a neighborhood of η^* such that $\|\eta-\eta^*\|_2<\delta$, by Condition 1 (iii), we apply the second-order Taylor expansion,

$$M(\eta)-M(\eta^*)=M'(\eta^*)\|\eta-\eta^*\|_2+\frac{1}{2}M''(\eta^*)\|\eta-\eta^*\|_2^2+o(\|\eta-\eta^*\|_2^2)$$
$$=\frac{1}{2}S''(\eta^*)\|\eta-\eta^*\|_2^2+o(\|\eta-\eta^*\|_2^2),$$

and as $S''(\eta^*) < 0$, there exists $c_0 = -\frac{1}{2}S''(\eta^*) > 0$ such that $S(t;\eta) - S(t;\eta^*) \leq c_0\|\eta - \eta^*\|_2^2$.

Condition 2 For all n large enough and sufficiently small δ, we consider the centered process $\hat{M} - M$, and have that

$$
\begin{aligned}
&E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{M}(\eta) - M(\eta) - \left\{\hat{M}(\eta^*) - M(\eta^*)\right\}\right|\right] \\
&= E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{M}(\eta) - M_n^*(\eta) + M_n^*(\eta) - M(\eta) - \left\{\hat{M}(\eta^*) - M_n^*(\eta^*) + M_n^*(\eta^*) - M(\eta^*)\right\}\right|\right] \\
&\leq E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{M}(\eta) - M_n^*(\eta) - \left\{\hat{M}(\eta^*) - M_n^*(\eta^*)\right\}\right|\right] \\
&\quad + E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}|M_n^*(\eta) - M(\eta) - \{M_n^*(\eta^*) - M(\eta^*)\}|\right] \\
&= (I) + (II),
\end{aligned}
$$

where $E^*(\cdot)$ denote the outer expectation, and we bound (I) and (II) respectively as follows.

Condition 2.1 To bound (II), we note that

$$
\begin{aligned}
M_n^*(\eta) - M_n^*(\eta^*) &= \frac{1}{n}\sum_{i=1}^n \Delta^*(O_i)(d(X_i;\eta) - d(X_i;\eta*)) \\
&= \frac{1}{n}\sum_{i=1}^n \Delta^*(O_i)(I\{X_i^{\mathrm{T}}\eta > 0\} - I\{X_i^{\mathrm{T}}\eta^* > 0\}),
\end{aligned}
$$

where

$$
\Delta^*(o) = \frac{\delta_Y(x;\beta^*)}{\delta_A(x;\alpha^*)} + \frac{(2z-1)(2t-1)}{\pi(t,z,x;\theta^*)\delta_A(x;\alpha^*)}\left\{y - \mu_Y(t,z,x;\beta^*) - \frac{\delta_Y(x;\beta^*)}{\delta_A(x;\alpha^*)}(a - \mu_A(t,z,x;\alpha^*))\right\}.
$$

We define a class of functions

$$
\mathcal{F}_\eta^1(o) = \left\{\Delta^*(o)(I\{x^{\mathrm{T}}\eta > 0\} - I\{x^{\mathrm{T}}\eta^* > 0\}) : \|\eta - \eta^*\|_2 < \delta\right\},
$$

and let $B_1 = \sup|\Delta^*(o)|$. By Assumption 2.3.2 and Condition 1, we have that $B_1 < \infty$.

When $\|\eta - \eta^*\|_2 < \delta$, by Condition 1 (i), there exists a constant $0 < k_0 < \infty$ such that $|x^{\mathrm{T}}(\eta - \eta^*)| < k_0\delta$. Furthermore, we show that $|d(x;\eta) - d(x;\eta^*)| = |I\{x^{\mathrm{T}}\eta > 0\} - I\{x^{\mathrm{T}}\eta^* > 0\}| \leq I\{-k_0\delta \leq x^{\mathrm{T}}\eta^* \leq k_0\delta\}$, by considering the three cases:

- when $-k_0\delta \leq x^{\mathrm{T}}\eta^* \leq k_0\delta$, we have $|d(x;\eta) - d(x;\eta^*)| \leq 1 = I\{-k_0\delta \leq x^{\mathrm{T}}\eta^* \leq k_0\delta\}$;
- when $x^{\mathrm{T}}\eta^* > k_0\delta > 0$, we have $x^{\mathrm{T}}\eta = x^{\mathrm{T}}(\eta-\eta^*)+x^{\mathrm{T}}\eta^* > 0$, so $|d(x;\eta)-d(x;\eta^*)| = 0 = I\{-k_0\delta \leq x^{\mathrm{T}}\eta^* \leq k_0\delta\}$;
- when $x^{\mathrm{T}}\eta^* < -k_0\delta < 0$, we have $x^{\mathrm{T}}\eta = x^{\mathrm{T}}(\eta - \eta^*) + x^{\mathrm{T}}\eta^* < 0$, so $|d(x;\eta) - d(x;\eta^*)| = 0 = I\{-k_0\delta \leq x^{\mathrm{T}}\eta^* \leq k_0\delta\}$.

Thus we define the envelope of $\mathcal{F}_\eta^1$ as $F_1 = B_1 I\{-k_0\delta \le x^{\mathrm{T}}\eta^* \le k_0\delta\}$. By Condition 1 (iv), there exists a constant $0 < k_1 < \infty$ such that

$$\|F_1\|_{P,2} \le B_1\sqrt{Pr(-k_0\delta \le x^{\mathrm{T}}\eta^* \le k_0\delta)} \le B_1\sqrt{2k_0k_1}\delta^{1/2} < \infty.$$

By Lemma 9.6 and Lemma 9.9 of Kosorok [2008], we have that $\mathcal{F}_\eta^1$, a class of indicator functions, is a Vapnik-Cervonenkis (VC) class with bounded bracketing entropy $J_{[]}^*(1, \mathcal{F}_\eta^1) < \infty$.

Next, we note that

$$\begin{aligned}\mathbb{G}_n\mathcal{F}_\eta^1 &= n^{-1/2}\sum_{i=1}^n \left\{\mathcal{F}_\eta^1(O_i) - E[\mathcal{F}_\eta^1(O)]\right\} \\ &= \sqrt{n}\left(M_n^*(\eta) - M_n^*(\eta^*) - \{M(\eta) - M(\eta^*)\}\right),\end{aligned}$$

and by Theorem 11.2 of Kosorok [2008], we obtain that there exists a constant $0 < c_1 < \infty$,

$$(II) = E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta} |\mathbb{G}_n\mathcal{F}_\eta^1|\right] \le c_1 J_{[]}^*(1, \mathcal{F}_\eta^1)\|F_1\|_{P,2} \le c_1 J_{[]}^*(1, \mathcal{F}_\eta^1)B_1\sqrt{2k_0k_1}\delta^{1/2} = \tilde{c}_1\delta^{1/2},$$

hence we conclude that $(II) \le \tilde{c}_1\delta^{1/2}$, where $\tilde{c}_1 > 0$ is a finite constant.

Condition 2.2 To bound (I), first we note that

$$\begin{aligned}&\hat{M}(\eta) - M_n^*(\eta) - \{\hat{M}(\eta^*) - M_n^*(\eta^*)\} = \hat{M}(\eta) - \hat{M}(\eta^*) - \{M_n^*(\eta) - M_n^*(\eta^*)\} \\ &= \frac{1}{n}\sum_{i=1}^n (d(X_i;\eta) - d(X_i;\eta*))(\hat{\Delta}(O_i) - \Delta^*(O_i)),\end{aligned}$$

and then apply the Taylor expansion at $(\alpha^*, \beta^*, \theta^*)$

$$\begin{aligned}&\hat{M}(\eta) - M_n^*(\eta) - \{\hat{M}(\eta^*) - M_n^*(\eta^*)\} \\ &= \frac{1}{n}\sum_{i=1}^n (d(X_i;\eta) - d(X_i;\eta*)) \left\{\left[g_1^*(O_i)\left(\frac{\partial\delta_A(X_i;\alpha^*)}{\partial\alpha}\right)^{\mathrm{T}} + g_2^*(O_i)\left(\frac{\partial\mu_A(T_i,Z_i,X_i;\alpha^*)}{\partial\alpha}\right)^{\mathrm{T}}\right](\hat{\alpha}-\alpha^*)\right. \\ &\quad + \left[g_3^*(O_i)\left(\frac{\partial\delta_Y(X_i;\beta^*)}{\partial\beta}\right)^{\mathrm{T}} + g_4^*(O_i)\left(\frac{\partial\mu_Y(T_i,Z_i,X_i;\beta^*)}{\partial\beta}\right)^{\mathrm{T}}\right](\hat{\beta}-\beta^*) \\ &\quad \left. + g_5^*(O_i)\left(\frac{\partial\pi(T_i,Z_i,X_i;\theta^*)}{\partial\theta}\right)^{\mathrm{T}}(\hat{\theta}-\theta^*)\right\} + o_p(n^{-1/2}),\end{aligned} \tag{A.2}$$

where

$$g_1^*(o) = -\frac{\delta_Y(x;\beta^*)}{\delta_A^2(x;\alpha^*)} - \frac{(2z-1)(2t-1)(y-\mu_Y(t,z,x;\beta^*))}{\pi(t,z,x;\theta^*)\delta_A^2(x;\alpha^*)} + \frac{2(2z-1)(2t-1)\delta_Y(x;\beta^*)}{\pi(t,z,x;\theta^*)\delta_A^3(x;\alpha^*)}(a-\mu_A(t,z,x;$$

$$g_2^*(o) = \frac{(2z-1)(2t-1)\delta_Y(x;\beta^*)}{\pi(t,z,x;\theta^*)\delta_A^2(x;\alpha^*)},$$

$$g_3^*(o) = \frac{1}{\delta_A^2(x;\alpha^*)} - \frac{2(2z-1)(2t-1)}{\pi(t,z,x;\theta^*)\delta_A^2(x;\alpha^*)}(a - \mu_A(t,z,x;\alpha^*)),$$

$$g_4^*(o) = -\frac{(2z-1)(2t-1)}{\pi(t,z,x;\theta^*)\delta_A(x;\alpha^*)},$$

$$g_5^*(o) = -\frac{(2z-1)(2t-1)}{\pi^2(t,z,x;\theta^*)\delta_A(x;\alpha^*)}\left\{y - \mu_Y(t,z,x;\beta^*) - \frac{\delta_Y(x;\beta^*)}{\delta_A(x;\alpha^*)}(a - \mu_A(t,z,x;\alpha^*))\right\}.$$

Similarly, we define the following classes of functions

$$\mathcal{F}_\eta^2(o) = \left\{\left[g_1^*(o)\left(\frac{\partial\delta_A(x;\alpha^*)}{\partial\alpha}\right)^{\mathrm{T}} + g_2^*(o)\left(\frac{\partial\mu_A(t,z,x;\alpha^*)}{\partial\alpha}\right)^{\mathrm{T}}\right](I\{x^{\mathrm{T}}\eta > 0\} - I\{x^{\mathrm{T}}\eta^* > 0\}) : \|\eta - \eta^*\|_2 <$$

$$\mathcal{F}_\eta^3(o) = \left\{\left[g_3^*(o)\left(\frac{\partial\delta_Y(x;\beta^*)}{\partial\beta}\right)^{\mathrm{T}} + g_4^*(o)\left(\frac{\partial\mu_Y(t,z,x;\beta^*)}{\partial\beta}\right)^{\mathrm{T}}\right](I\{x^{\mathrm{T}}\eta > 0\} - I\{x^{\mathrm{T}}\eta^* > 0\}) : \|\eta - \eta^*\|_2 <$$

$$\mathcal{F}_\eta^4(o) = \left\{g_5^*(o)\left(\frac{\partial\pi(t,z,x;\theta^*)}{\partial\theta}\right)^{\mathrm{T}}(I\{x^{\mathrm{T}}\eta > 0\} - I\{x^{\mathrm{T}}\eta^* > 0\}) : \|\eta - \eta^*\|_2 < \delta\right\},$$

and let $B_2 = \sup|g_1^*(o)\partial\delta_A(x;\alpha^*)/\partial\alpha + g_2^*(o)\partial\mu_A(t,z,x;\alpha^*)/\partial\alpha|$, $B_3 = \sup|g_3^*(o)\partial\delta_Y(x;\beta^*)/\partial\beta + g_4^*(o)\partial\mu_Y(t,z,x;\beta^*)/\partial\beta|$, and $B_4 = \sup|g_5^*(o)\partial\pi(t,z,x;\theta^*)/\partial\theta|$, where $B_2, B_3, B_4 > 0$ and the supremum is taken over all the coordinates. By Assumption 2.3.2 and Condition 1, we have that $B_2, B_3, B_4 < \infty$.

Using the same technique as in **Condition 2.1**, we define the envelop of $\mathcal{F}_\eta^j$ as $F_j = B_j I\{-k_0\delta \le x^{\mathrm{T}}\eta^* \le k_0\delta\}$ for $j = 2,3,4$, and obtain that

$$\|F_j\|_{P,2} \le \tilde{B}_j\delta^{1/2} < \infty, \quad j = 2,3,4,$$

where $\tilde{B}_2, \tilde{B}_3, \tilde{B}_4$ are some finite constants, and that $\mathcal{F}_\eta^j$ is a VC class with bounded bracketing entropy $J_{[]}^*(1,\mathcal{F}_\eta^j) < \infty$, for $j = 2,3,4$. By Theorem 11.2 of Kosorok [2008], we obtain that

$$E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}\left\|\mathbb{G}_N\mathcal{F}_\eta^j\right\|_1\right] \le c_j J_{[]}^*(1,\mathcal{F}_\eta^j)\|F_j\|_{P,2}, \quad j = 2,3,4,$$

where $c_2, c_3, c_4 > 0$ are some finite constants.

Furthermore, by Theorem 2.14.5 of van der Vaart and Wellner [1996], we obtain that

$$\begin{aligned}\left\{E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_n\mathcal{F}_\eta^j\|_2^2\right]\right\}^{1/2} &\le l_j\left\{E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_n\mathcal{F}_\eta^j\|_1\right] + \|F_j\|_{P,2}\right\} \\ &\le l_j\{c_j J_{[]}^*(1,\mathcal{F}_\eta^j) + 1\}\|F_j\|_{P,2} \\ &\le \tilde{c}_j\delta^{1/2}, \qquad j = 2,3,4,\end{aligned}$$

where l_2, l_3, l_4 and $\tilde{c}_2, \tilde{c}_3, \tilde{c}_4$ are some finite constants.

By Equation (A.2), we have that

$$
\begin{aligned}
(I) &= E^*\left[n^{1/2}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{M}(\eta)-M_n^*(\eta)-\{\hat{M}(\eta^*)-M_n^*(\eta^*)\}\right|\right] \\
&\le E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}\left\{|\mathbb{G}_n\mathcal{F}_\eta^2(\hat{\alpha}-\alpha^*)|+|\mathbb{G}_n\mathcal{F}_\eta^3(\hat{\beta}-\beta^*)|+|\mathbb{G}_n\mathcal{F}_\eta^4(\hat{\theta}-\theta^*)|+o_p(1)\right\}\right] \\
&\le n^{-1/2}\left\{E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^2\cdot n^{1/2}(\hat{\alpha}-\alpha^*)|\right]+E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^3\cdot n^{1/2}(\hat{\beta}-\beta^*)|\right]\right. \\
&\qquad \left.+E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^4\cdot n^{1/2}(\hat{\theta}-\theta^*)|\right]\right\},
\end{aligned}
$$

and then by the Cauchy-Schwarz inequality, we obtain that

$$
\begin{aligned}
(I) \le\, & n^{-1/2}\left\{E[n\|\hat{\alpha}-\alpha^*\|_2^2]\right\}^{1/2}\left\{E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_n\mathcal{F}_\eta^2\|_2^2\right]\right\}^{1/2} \\
&+ n^{-1/2}\left\{E[n\|\hat{\beta}-\beta^*\|_2^2]\right\}^{1/2}\left\{E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_n\mathcal{F}_\eta^3\|_2^2\right]\right\}^{1/2} \\
&+ n^{-1/2}\left\{E[n\|\hat{\theta}-\theta^*\|_2^2]\right\}^{1/2}\left\{E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_n\mathcal{F}_\eta^4\|_2^2\right]\right\}^{1/2}.
\end{aligned}
$$

By Condition 2, we have that $B_\alpha = \{E[n\|\hat{\alpha}-\alpha^*\|_2^2]\}^{1/2} < \infty$, $B_\beta = \left\{E[n\|\hat{\beta}-\beta^*\|_2^2]\right\}^{1/2} < \infty$, $B_\theta = \left\{E[n\|\hat{\theta}-\theta^*\|_2^2]\right\}^{1/2} < \infty$, hence

$$
(I) \le n^{-1/2}(B_\alpha\tilde{c}_2 + B_\beta\tilde{c}_3 + B_\theta\tilde{c}_4)\delta^{1/2}.
$$

In summary, we conclude that as $n \to \infty$, the centered process satisfies

$$
E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{M}(\eta)-M(\eta)-\{\hat{M}(\eta^*)-M(\eta^*)\}\right|\right] \le (I)+(II) \le \tilde{c}_1\delta^{1/2}. \quad \text{(A.3)}
$$

Let $\phi_n(\delta) = \delta^{1/2}$ and $b = \frac{3}{2} < 2$, thus we have $\frac{\phi_n(\delta)}{\delta^b} = \delta^{-1}$ is decreasing, and b does not depend on n.

Condition 3 By the facts that $\hat{\eta} \xrightarrow{p} \eta^*$ as $n \to \infty$, and that $\hat{M}(\hat{\eta}) \ge \sup_\eta \hat{M}(\eta)$, we choose $r_n = n^{1/3}$ such that $r_n^2\phi_n(r_n^{-1}) = n^{2/3}\phi_n(n^{-1/3}) = n^{1/2}$.

In the end, the three conditions are satisfied with $r_n = n^{1/3}$; thus we conclude that $n^{1/3}\|\hat{\eta}-\eta^*\|_2 = O_p(1)$, which completes the proof of (i) of Theorem 3.2.2.

PART 3. We characterize the asymptotic distribution of $\hat{M}(\hat{\eta})$. First we note that

$$
\sqrt{n}\{\hat{M}(\hat{\eta})-M(\eta^*)\} = \sqrt{n}\{\hat{M}(\hat{\eta})-\hat{M}(\eta^*)\}+\sqrt{n}\{\hat{M}(\eta^*)-M(\eta^*)\},
$$

and then study the two terms in two steps.

Step 3.1 To establish $\sqrt{n}\{\hat{M}(\hat{\eta})-\hat{M}(\eta^*)\}=o_p(1)$, it suffices to show that $\sqrt{n}\{M(\hat{\eta})-M(\eta^*)\}=o_p(1)$ and $\sqrt{n}(\hat{M}(\hat{\eta})-\hat{M}(\eta^*)-\{M(\hat{\eta})-M(\eta^*)\})=o_p(1)$.

First, as $n^{1/3}\|\hat{\eta}-\eta^*\|_2=O_p(1)$, we apply the second-order Taylor expansion

$$\begin{aligned}\sqrt{n}\{M(\hat{\eta})-M(\eta^*)\} &= \sqrt{n}\left\{M'(\eta^*)\|\hat{\eta}-\eta^*\|_2+\frac{1}{2}M''(\eta^*)\|\hat{\eta}-\eta^*\|_2^2+o_p(\|\hat{\eta}-\eta^*\|_2^2)\right\}\\ &= \sqrt{n}\left\{\frac{1}{2}M''(\eta^*)\|\hat{\eta}-\eta^*\|_2^2+o_p(\|\hat{\eta}-\eta^*\|_2^2)\right\}\\ &= \sqrt{n}\left\{\frac{1}{2}M''(\eta^*)O_p(n^{-2/3})+o_p(n^{-2/3})\right\}=o_p(1),\end{aligned}$$

which proves (ii) of Theorem 3.2.2.

Next, we follow the result (A.3) obtained in **PART 2**. As $n^{1/3}\|\hat{\eta}-\eta^*\|_2=O_p(1)$, there exists $\tilde{\delta}=c_5n^{-1/3}$, where $c_5<\infty$ is a finite constant, such that $\|\hat{\eta}-\eta^*\|_2\leq\tilde{\delta}$. Therefore we have

$$\begin{aligned}&\sqrt{n}(\hat{M}(\hat{\eta})-\hat{M}(\eta^*)-\{M(\hat{\eta})-M(\eta^*)\})\\ &\leq E^*\left[\sqrt{n}\sup_{\|\hat{\eta}-\eta^*\|_2<\tilde{\delta}}\left|\hat{M}(\hat{\eta})-M(\hat{\eta})-\{\hat{M}(\eta^*)-M(\eta^*)\}\right|\right]\\ &\leq \tilde{c}_1\tilde{\delta}^{1/2}=\tilde{c}_1\sqrt{c_5}n^{-1/6}=o_p(1),\end{aligned}$$

which yields the result.

Step 3.2 To derive the asymptotic distribution of $\sqrt{n}\{\hat{M}(\eta^*)-M(\eta^*)\}$, we follow the result (A.1) obtained in **PART 1** and have that

$$\sqrt{n}\left\{\hat{M}(\eta^*)-M(\eta^*)\right\}\xrightarrow{D}\mathcal{N}(0,\sigma_1^2),$$

where $\sigma_1^2=E[(M^*-M+H_{\alpha^*}^{\mathrm{T}}\phi_{\alpha,i}+H_{\beta^*}^{\mathrm{T}}\phi_{\beta,i}+H_{\theta^*}^{\mathrm{T}}\phi_{\theta,i})^2]$.

Therefore we obtain in the end

$$\begin{aligned}\sqrt{n}\{\hat{M}(\hat{\eta})-M(\eta^*)\} &= \sqrt{n}\{\hat{M}(\hat{\eta})-\hat{M}(\eta^*)\}+\sqrt{n}\{\hat{M}(\eta^*)-M(\eta^*)\}\\ &= o_p(1)+\sqrt{n}\{\hat{M}(\eta^*)-M(\eta^*)\}\\ &\xrightarrow{D}\mathcal{N}(0,\sigma_1^2),\end{aligned}$$

which completes the proof.

A.9 Proof of Theorem 3.2.3

We first review a useful lemma from Kennedy et al. [2020], which illustrates the basic technique of cross-fitting.

Lemma A.9.1. *Consider two independent samples $\mathcal{O}_1=(O_1,\ldots,O_n)$ and $\mathcal{O}_2=(O_{n+1},\ldots,O_{\tilde{n}})$, let $\hat{f}(o)$ be a function estimated from $\mathcal{O}_2$ and $\mathbb{P}_n$ the empirical measure over $\mathcal{O}_1$, then we have*

$$(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)=O_{\mathbb{P}}\left(\frac{\|\hat{f}-f\|}{\sqrt{n}}\right)$$

Proof. First note that by conditioning on $\mathcal{O}_2$ we obtain

$$\mathbb{E}\left\{\mathbb{P}_n(\hat{f}-f)\,\middle|\,\mathcal{O}_2\right\} = \mathbb{E}(\hat{f}-f\,|\,\mathcal{O}_2) = \mathbb{P}(\hat{f}-f)$$

and the conditional variance is

$$var\{(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)\,|\,\mathcal{O}_2\} = var\{\mathbb{P}_n(\hat{f}-f)\,|\,\mathcal{O}_2\} = \frac{1}{n}var(\hat{f}-f\,|\,\mathcal{O}_2) \le \|\hat{f}-f\|^2/n$$

therefore by Chebyshev's inequality we have

$$\mathbb{P}\left\{\frac{|(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)|}{\|\hat{f}-f\|^2/n} \ge t\right\} = \mathbb{E}\left[\mathbb{P}\left\{\frac{|(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)|}{\|\hat{f}-f\|^2/n} \ge t\,\middle|\,\mathcal{O}_2\right\}\right] \le \frac{1}{t^2}$$

thus for any $\epsilon > 0$ we can pick $t = 1/\sqrt{\epsilon}$ so that the probability above is no more than ϵ, which yields the result. □

We randomly split data into K folds. For $k = 1, \ldots, K$,

$$\hat{M}(\eta) = \frac{1}{K}\sum_{k=1}^{K}\hat{M}_k(\eta) = \frac{1}{K}\sum_{k=1}^{K}P_{n,k}\{\Delta(O;\hat{\mu}_{A,-k},\hat{\mu}_{Y,-k},\hat{\pi}_{-k})d(X)\},$$

where $P_{n,k}$ denote empirical averages only over the k-th fold, and $\hat{\mu}_{A,-k}$, $\hat{\mu}_{Y,-k}$ and $\hat{\pi}_{-k}$ denote the nuisance estimators constructed excluding the k-th fold.

Now we start our proof which has three main parts as follows.

PART 1. We prove that $\hat{M}(\eta) - M_n(\eta) = o_p(n^{-1/2})$, where $M_n(\eta) = P_n\{\Delta(O)d(X,\eta)\}$. Essentially it suffices to prove that $\hat{M}_k(\eta) - M_{n,k}(\eta) = o_p(n^{-1/2})$, where $M_{n,k}(\eta) = P_{n,k}\{\Delta(O)d(X,\eta)\}$.

First we note the following decomposition

$$\begin{aligned}
&\hat{M}_k(\eta) - M_{n,k}(\eta)\\
&= P_{n,k}\, d(\eta)\left\{\frac{\hat{\delta}_{Y,-k}}{\hat{\delta}_{A,-k}} - \frac{\delta_Y}{\delta_A} + (2Z-1)(2T-1)\left[\left(\frac{1}{\hat{\pi}_{-k}}-\frac{1}{\pi}\right)\left(\frac{1}{\hat{\delta}_{A,-k}}-\frac{1}{\delta_A}\right)\left(Y-\hat{\mu}_{Y,-k}-\frac{\hat{\delta}_{Y,-k}}{\hat{\delta}_{A,-k}}(A-\right.\right.\right.\\
&\quad +\frac{1}{\delta_A}\left(\frac{1}{\hat{\pi}_{-k}}-\frac{1}{\pi}\right)G_1 + \frac{1}{\pi}\left(\frac{1}{\hat{\delta}_{A,-k}}-\frac{1}{\delta_A}\right)G_1 + \frac{1}{\pi\delta_A}G_2\\
&\quad +\frac{1}{\delta_A}\left(\frac{1}{\hat{\pi}_{-k}}-\frac{1}{\pi}\right)\left(Y-\mu_Y-\frac{\delta_Y}{\delta_A}(A-\mu_A)\right) + \frac{1}{\pi}\left(\frac{1}{\hat{\delta}_{A,-k}}-\frac{1}{\delta_A}\right)\left(Y-\mu_Y-\frac{\delta_Y}{\delta_A}(A-\mu_A)\right)\\
&\quad +\frac{1}{\pi\delta_A}\left(\mu_Y-\hat{\mu}_{Y,-k}-\frac{1}{\delta_A}(\hat{\delta}_{Y,-k}-\delta_Y)(A-\mu_A)-\delta_Y\left(\frac{1}{\hat{\delta}_{A,-k}}-\frac{1}{\delta_A}\right)(A-\mu_A)+\frac{\delta_Y}{\delta_A}(\hat{\mu}_{A,-k}-\mu_A)\right.
\end{aligned}$$

where we omit the arguments of the nuisance functions to simplify the notation, and

denote

$$G_1 = \mu_Y - \hat{\mu}_{Y,-k} - (\hat{\delta}_{Y,-k} - \delta_Y)\left(\frac{1}{\hat{\delta}_{A,-k}} - \frac{1}{\delta_A}\right)(A - \mu_A) - \frac{1}{\delta_A}(\hat{\delta}_{Y,-k} - \delta_Y)(A - \mu_A)$$
$$+ \frac{1}{\delta_A}(\hat{\delta}_{Y,-k} - \delta_Y)(\hat{\mu}_{A,-k} - \mu_A) - \delta_Y\left(\frac{1}{\hat{\delta}_{A,-k}} - \frac{1}{\delta_A}\right)(A - \mu_A)$$
$$+ \delta_Y\left(\frac{1}{\hat{\delta}_{A,-k}} - \frac{1}{\delta_A}\right)(\hat{\mu}_{A,-k} - \mu_A) + \frac{\delta_Y}{\delta_A}(\hat{\mu}_{A,-k} - \mu_A),$$

$$G_2 = \frac{\hat{\delta}_{Y,-k} - \delta_Y}{\delta_A}(\hat{\mu}_{A,-k} - \mu_A) + \delta_Y\left(\frac{1}{\hat{\delta}_{A,-k}} - \frac{1}{\delta_A}\right)(\hat{\mu}_{A,-k} - \mu_A) - (\hat{\delta}_{Y,-k} - \delta_Y)\left(\frac{1}{\hat{\delta}_{A,-k}} - \frac{1}{\delta_A}\right)(A - \mu_A).$$

In summary, we have two types of terms from this decomposition: product terms and mean zero terms (by multiple robustness). The product terms are $o_p(n^{-1/2})$ by Cauchy-Schwarz inequality and Condition 3 (rate of convergence). The mean zero terms are $o_p(n^{-1/2})$ by Lemma A.9.1.

PART 2. We prove that $n^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$.

First we note that, by Condition 1 (iii), $M(\eta)$ is twice continuously differentiable at a neighborhood of η^*. In PART 1, we show that $\hat{M}(\eta) = M(\eta) + o_p(1), \forall \eta$. Since $\hat{\eta}$ maximizes $\hat{M}(\eta)$, we have that $\hat{M}(\hat{\eta}) \geq \sup_\eta \hat{M}(\eta)$; thus by the Argmax theorem, we obtain that $\hat{\eta} \xrightarrow{p} \eta^*$ as $n \to \infty$.

Then we apply Theorem 14.4 (Rate of convergence) of Kosorok [2008] to establish the $n^{-1/3}$ rate of convergence of $\hat{\eta}$, and need to find the suitable rate that satisfies three conditions below.

Condition 1 For every η in a neighborhood of η^* such that $\|\eta - \eta^*\|_2 < \delta$, by Condition 1 (iii), we apply the second-order Taylor expansion,

$$M(\eta) - M(\eta^*) = M'(\eta^*)\|\eta - \eta^*\|_2 + \frac{1}{2}M''(\eta^*)\|\eta - \eta^*\|_2^2 + o(\|\eta - \eta^*\|_2^2)$$
$$= \frac{1}{2}S''(\eta^*)\|\eta - \eta^*\|_2^2 + o(\|\eta - \eta^*\|_2^2),$$

and as $S''(\eta^*) < 0$, there exists $c_0 = -\frac{1}{2}S''(\eta^*) > 0$ such that $S(t;\eta) - S(t;\eta^*) \leq c_0\|\eta - \eta^*\|_2^2$.

Condition 2 For all n large enough and sufficiently small δ, we consider the centered

process $\hat{M} - M$, and have that

$$
\begin{aligned}
&E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{M}(\eta)-M(\eta)-\left\{\hat{M}(\eta^*)-M(\eta^*)\right\}\right|\right] \\
&= E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{M}(\eta)-M_n^*(\eta)+M_n^*(\eta)-M(\eta)-\left\{\hat{M}(\eta^*)-M_n^*(\eta^*)+M_n^*(\eta^*)-M(\eta^*)\right\}\right|\right] \\
&\leq E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{M}(\eta)-M_n^*(\eta)-\left\{\hat{M}(\eta^*)-M_n^*(\eta^*)\right\}\right|\right] \\
&\quad + E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}\left|M_n^*(\eta)-M(\eta)-\{M_n^*(\eta^*)-M(\eta^*)\}\right|\right] \\
&= (I)+(II),
\end{aligned}
$$

where $E^*(\cdot)$ denote the outer expectation, and we bound (I) and (II) respectively as follows.

It follows from the result in **PART 1** that $(I) = o_p(1)$.

To bound (II), we note that

$$
\begin{aligned}
M_n^*(\eta) - M_n^*(\eta^*) &= \frac{1}{n}\sum_{i=1}^n \Delta^*(O_i)(d(X_i;\eta) - d(X_i;\eta*)) \\
&= \frac{1}{n}\sum_{i=1}^n \Delta^*(O_i)(I\{X_i^{\mathrm{T}}\eta > 0\} - I\{X_i^{\mathrm{T}}\eta^* > 0\}),
\end{aligned}
$$

where

$$
\Delta^*(o) = \frac{\delta_Y(x)}{\delta_A(x)} + \frac{(2z-1)(2t-1)}{\pi(t,z,x)\delta_A(x)}\left\{y - \mu_Y(t,z,x) - \frac{\delta_Y(x)}{\delta_A(x)}(a-\mu_A(t,z,x))\right\}.
$$

We define a class of functions

$$
\mathcal{F}_\eta^5(o) = \left\{\Delta^*(o)(I\{x^{\mathrm{T}}\eta > 0\} - I\{x^{\mathrm{T}}\eta^* > 0\}) : \|\eta-\eta^*\|_2 < \delta\right\},
$$

and let $B_5 = \sup|\Delta^*(o)|$. By Assumption 2.3.2 and Condition 1, we have that $B_5 < \infty$.

Using the same technique as in Section **Condition 2.1**, we define the envelop of $\mathcal{F}_\eta^5$ as $F_5 = B_5 I\{-k_0\delta \leq x^T\eta^* \leq k_0\delta\}$, and obtain that $\|F_5\|_{P,2} \leq \tilde{B}_9\delta^{1/2} < \infty$, where $\tilde{B}_9$ is a finite constant, and that $\mathcal{F}_\eta^5$ is a VC class with bounded entropy $J_{[]}^*(1,\mathcal{F}_\eta^5) < \infty$. By Theorem 11.2 of Kosorok [2008], we obtain that there exists a constant $0 < c_6 < \infty$,

$$
(II) = E^*\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^5|\right] \leq c_6 J_{[]}^*(1,\mathcal{F}_\eta^5)\|F_5\|_{P,2} \leq c_6 J_{[]}^*(1,\mathcal{F}_\eta^5)B_5\sqrt{2k_0k_1}\delta^{1/2} = \tilde{c}_5\delta^{1/2}.
$$

In summary, we conclude that as $n \to \infty$, the centered process satisfies

$$
E^*\left[\sqrt{n}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{M}(\eta)-M(\eta)-\{\hat{M}(\eta^*)-M(\eta^*)\}\right|\right] \leq (I)+(II) \leq \tilde{c}_5\delta^{1/2}. \quad \text{(A.4)}
$$

Let $\phi_n(\delta) = \delta^{1/2}$ and $b = \frac{3}{2} < 2$, thus we have $\frac{\phi_n(\delta)}{\delta^b} = \delta^{-1}$ is decreasing, and b does not depend on n.

Condition 3 By the facts that $\hat{\eta} \xrightarrow{p} \eta^*$ as $n \to \infty$, and that $\hat{M}(\hat{\eta}) \geq \sup_\eta \hat{M}(\eta)$, we choose $r_n = n^{1/3}$ such that $r_n^2 \phi_n(r_n^{-1}) = n^{2/3}\phi_n(n^{-1/3}) = n^{1/2}$.

In the end, the three conditions are satisfied with $r_n = n^{1/3}$; thus we conclude that $n^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$, which completes the proof of (i) of Theorem 3.2.3.

PART 3. We characterize the asymptotic distribution of $\hat{M}(\hat{\eta})$. First we note that

$$\sqrt{n}\{\hat{M}(\hat{\eta}) - M(\eta^*)\} = \sqrt{n}\{\hat{M}(\hat{\eta}) - \hat{M}(\eta^*)\} + \sqrt{n}\{\hat{M}(\eta^*) - M(\eta^*)\},$$

and then study the two terms in two steps.

Step 3.1 To establish $\sqrt{n}\{\hat{M}(\hat{\eta}) - \hat{M}(\eta^*)\} = o_p(1)$, it suffices to show that $\sqrt{n}\{M(\hat{\eta}) - M(\eta^*)\} = o_p(1)$ and $\sqrt{n}(\hat{M}(\hat{\eta}) - \hat{M}(\eta^*) - \{M(\hat{\eta}) - M(\eta^*)\}) = o_p(1)$.

First, as $n^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$, we apply the second-order Taylor expansion

$$\begin{aligned}
\sqrt{n}\{M(\hat{\eta}) - M(\eta^*)\} &= \sqrt{n}\left\{M'(\eta^*)\|\hat{\eta} - \eta^*\|_2 + \frac{1}{2}M''(\eta^*)\|\hat{\eta} - \eta^*\|_2^2 + o_p(\|\hat{\eta} - \eta^*\|_2^2)\right\} \\
&= \sqrt{n}\left\{\frac{1}{2}M''(\eta^*)\|\hat{\eta} - \eta^*\|_2^2 + o_p(\|\hat{\eta} - \eta^*\|_2^2)\right\} \\
&= \sqrt{n}\left\{\frac{1}{2}M''(\eta^*)O_p(n^{-2/3}) + o_p(n^{-2/3})\right\} = o_p(1),
\end{aligned}$$

which proves (ii) of Theorem 3.2.3.

Next, we follow the result (A.4) obtained in **PART 2**. As $n^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$, there exists $\tilde{\delta} = c_7 n^{-1/3}$, where $c_7 < \infty$ is a finite constant, such that $\|\hat{\eta} - \eta^*\|_2 \leq \tilde{\delta}$. Therefore we have

$$\begin{aligned}
&\sqrt{n}(\hat{M}(\hat{\eta}) - \hat{M}(\eta^*) - \{M(\hat{\eta}) - M(\eta^*)\}) \\
&\leq E^*\left[\sqrt{n}\sup_{\|\hat{\eta}-\eta^*\|_2<\tilde{\delta}}\left|\hat{M}(\hat{\eta}) - M(\hat{\eta}) - \{\hat{M}(\eta^*) - M(\eta^*)\}\right|\right] \\
&\leq \tilde{c}_5\tilde{\delta}^{1/2} = \tilde{c}_5\sqrt{c_7}n^{-1/6} = o_p(1),
\end{aligned}$$

which yields the result.

Step 3.2 To derive the asymptotic distribution of $\sqrt{n}\{\hat{M}(\eta^*) - M(\eta^*)\}$, we follow the result obtained in **PART 1** and have that

$$\sqrt{n}\left\{\hat{M}(\eta^*) - M(\eta^*)\right\} \xrightarrow{D} \mathcal{N}(0, \sigma_2^2),$$

where $\sigma_2^2 = E[(\Delta(O_i)d(X_i;\eta^*) - M(\eta^*))^2]$.

Therefore we obtain in the end

$$\begin{aligned}
\sqrt{n}\{\hat{M}(\hat{\eta}) - M(\eta^*)\} &= \sqrt{n}\{\hat{M}(\hat{\eta}) - \hat{M}(\eta^*)\} + \sqrt{n}\{\hat{M}(\eta^*) - M(\eta^*)\} \\
&= o_p(1) + \sqrt{n}\{\hat{M}(\eta^*) - M(\eta^*)\} \\
&\xrightarrow{D} \mathcal{N}(0, \sigma_2^2),
\end{aligned}$$

which completes the proof.

A.10 Proof of Theorem 5.1.2 and 5.1.3

We first prove the identification result.

First we note that

$$
\begin{aligned}
&\delta_{Y,1}(X) - \delta_{Y,0}(X) = E[Y_1 - Y_0 \mid X, Z = 1] - E[Y_1 - Y_0 \mid X, Z = 0] \\
&= \sum_{z=0,1} (2z-1) E[Y_1 - Y_0 \mid X, Z = z] \\
&= \sum_{z=0,1} (2z-1) E[Y_1(A_1(z)) - Y_0(A_0(z)) \mid X, Z = z] \\
&= \sum_{z=0,1} (2z-1) E[A_1(z)Y_1(1) + (1 - A_1(z))Y_1(0) - A_0(z)Y_0(1) - (1 - A_0(z))Y_0(0) \mid Z = z, X] \\
&= \sum_{z=0,1} (2z-1) E[A_1(z)(Y_1(1) - Y_1(0)) - A_0(z)(Y_0(1) - Y_0(0)) + Y_1(0) - Y_0(0) \mid Z = z, X] \\
&= \sum_{z=0,1} (2z-1) (E[A_1(z)(Y_1(1) - Y_1(0)) \mid X, Z = z] - E[A_0(z)(Y_0(1) - Y_0(0)) \mid X, Z = z] \\
&\qquad\qquad + E[Y_1(0) - Y_0(0) \mid X, Z = z]) \\
&= \sum_{z=0,1} (2z-1) (E[A_1(z)(Y_1(1) - Y_1(0)) \mid X, Z = z] - E[A_0(z)(Y_0(1) - Y_0(0)) \mid X, Z = z] \\
&= \sum_{z=0,1} (2z-1) (E[A_1(z)(Y_1(1) - Y_1(0)) \mid X] - E[A_0(z)(Y_0(1) - Y_0(0)) \mid X] \\
&= E[(A_1(1) - A_1(0))(Y_1(1) - Y_1(0)) \mid X] - E[(A_0(1) - A_0(0))(Y_0(1) - Y_0(0)) \mid X] \\
&= E[A_1(1) - A_1(0) \mid X]\tau(X) - E[A_0(1) - A_0(0) \mid X]\tau(X) \\
&= E[A_1(1) - A_1(0) - A_0(1) + A_0(0) \mid X]\tau(X).
\end{aligned}
$$

We also note that

$$
\begin{aligned}
&\delta_{A,1}(X) - \delta_{A,0}(X) = E[A_1 - A_0 \mid X, Z = 1] - E[A_1 - A_0 \mid X, Z = 0] \\
&= \sum_{z=0,1} (2z-1) E[A_1 - A_0 \mid X, Z = z] \\
&= \sum_{z=0,1} (2z-1) E[A_1(z) - A_0(z) \mid X, Z = z] \\
&= E[A_1(1) - A_1(0) - A_0(1) + A_0(0) \mid X].
\end{aligned}
$$

Combining the above derivations, we obtain that $\delta_{Y,1}(X) - \delta_{Y,0}(X) = (\delta_{A,1}(X) - \delta_{A,0}(X))\tau(X)$. That is, the CATE is identified by

$$\tau(X) = \frac{\delta_{Y,1}(X) - \delta_{Y,0}(X)}{\delta_{A,1}(X) - \delta_{A,0}(X)}.$$

Alternatively, we consider the following assumptions: (sequential ignorability) $Y_t(a) \perp A_t \mid U, X, Z$ for $t, a = 0, 1$, and there is no additive interaction of either (i) $E[A_1 - A_0 \mid X, U, Z = 1] - E[A_1 - A_0 \mid X, U, Z = 0] = E[A_1 - A_0 \mid X, Z = 1] - E[A_1 - A_0 \mid X, Z = 0]$ or (ii) $E[Y_t(1) - Y_t(0) \mid U, X] = E[Y_t(1) - Y_t(0) \mid X]$ for $t = 0, 1$.

We can continue that

$$
\begin{aligned}
&\delta_{Y,1}(X) - \delta_{Y,0}(X) \\
&= \sum_{z=0,1} (2z-1)(E[A_1(z)(Y_1(1) - Y_1(0)) \mid X, Z = z] - E[A_0(z)(Y_0(1) - Y_0(0)) \mid X, Z = z]) \\
&= E_U \sum_{z=0,1} (2z-1)(E[A_1(z)(Y_1(1) - Y_1(0)) \mid X, U, Z = z] - E[A_0(z)(Y_0(1) - Y_0(0)) \mid X, U, Z = z] \\
&= E_U \sum_{z=0,1} (2z-1)(E[A_1(z) \mid X, U, Z = z] E[Y_1(1) - Y_1(0) \mid X, U, Z = z] \\
&\qquad\qquad - E[A_0(z) \mid X, U, Z = z] E[Y_0(1) - Y_0(0) \mid X, U, Z = z]) \\
&= E_U[E[Y_t(1) - Y_t(0) \mid U, X](E[A_1 - A_0 \mid X, U, Z = 1] - E[A_1 - A_0 \mid X, U, Z = 0])].
\end{aligned}
$$

Under Assumption (i), we have that

$$
\begin{aligned}
&E[A_1 - A_0 \mid X, U, Z = 1] - E[A_1 - A_0 \mid X, U, Z = 0] \\
&= E[A_1 - A_0 \mid X, Z = 1] - E[A_1 - A_0 \mid X, Z = 0] \\
&= \delta_{A,1}(X) - \delta_{A,0}(X);
\end{aligned}
$$

or under Assumption (ii), we have that

$$
E[Y_t(1) - Y_t(0) \mid U, X] = E[Y_t(1) - Y_t(0) \mid X], t = 0, 1,
$$

and also

$$
E_U[E[A_1 - A_0 \mid X, U, Z = 1] - E[A_1 - A_0 \mid X, U, Z = 0]] = \delta_{A,1}(X) - \delta_{A,0}(X).
$$

Hence combining the above derivations, we obtain the same identification results.

Next, we derive the efficient influence function.

For a given distribution P in the nonparametric statistical model $\mathcal{M}$, we let p denote the density of P with respect to some dominating measure ν. For all bounded $h \in L_2(P)$, define the parametric submodel $p_\epsilon = (1 + \epsilon h)p$, which is valid for small enough ϵ and has score h at $\epsilon = 0$.

We study the following statistical functional

$$
\Psi(P) = E_P \left[\frac{E_P[Y_1 - Y_0 \mid X = x, Z = 1] - E_P[Y_1 - Y_0 \mid X = x, Z = 0]}{E_P[A_1 - A_0 \mid X = x, Z = 1] - E_P[A_1 - A_0 \mid X = x, Z = 0]} \right],
$$

and would establish that $\Psi(P)$ is pathwise differentiable with respect to $\mathcal{M}$ at P with efficient influence function ϕ_P if we have that for any $P \in \mathcal{M}$

$$
\frac{\partial}{\partial \epsilon} \Psi(P_\epsilon) \bigg|_{\epsilon=0} = \int \phi_P(o) h(o) dP(o).
$$

We denote $\delta_{Y,z,\epsilon}(x) = E_{P_\epsilon}[Y_1 - Y_0 \mid X = x, Z = z]$, $\delta_{A,z,\epsilon}(x) = E_{P_\epsilon}[A_1 - A_0 \mid X = x, Z = z]$, $S = \partial \log p_\epsilon / \partial \epsilon$, and compute

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\Psi(P_\epsilon)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} E_{P_\epsilon}\left[\frac{\delta_{Y,1,\epsilon}(X) - \delta_{Y,0,\epsilon}(X)}{\delta_{A,1,\epsilon}(X) - \delta_{A,0,\epsilon}(X)}\right]\Big|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon} E_P\left[(1+\epsilon S)\frac{\delta_{Y,1,\epsilon}(X) - \delta_{Y,0,\epsilon}(X)}{\delta_{A,1,\epsilon}(X) - \delta_{A,0,\epsilon}(X)}\right]\Big|_{\epsilon=0} \\
&= E_P\left[S\frac{\delta_{Y,1}(X) - \delta_{Y,0}(X)}{\delta_{A,1}(X) - \delta_{A,0}(X)}\right] \\
&\quad + E_P\left[\frac{1}{\delta_{A,1}(X) - \delta_{A,0}(X)}\left(\frac{\partial}{\partial \epsilon}\delta_{Y,1,\epsilon}(X)\Big|_{\epsilon=0} - \frac{\partial}{\partial \epsilon}\delta_{Y,0,\epsilon}(X)\Big|_{\epsilon=0}\right)\right] \\
&\quad - E_P\left[\frac{\delta_{Y,1}(X) - \delta_{Y,0}(X)}{\{\delta_{A,1}(X) - \delta_{A,0}(X)\}^2}\left(\frac{\partial}{\partial \epsilon}\delta_{A,1,\epsilon}(X)\Big|_{\epsilon=0} - \frac{\partial}{\partial \epsilon}\delta_{A,0,\epsilon}(X)\Big|_{\epsilon=0}\right)\right].
\end{aligned}$$

Then we need to compute

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\delta_{Y,z,\epsilon}(X)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} E_{P_\epsilon}[Y_1 - Y_0 \mid X, Z = z]\Big|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon}\frac{\delta_{Y,z}(X) + \epsilon E_P[S(Y_1 - Y_0) \mid X, Z = z]}{1 + \epsilon E_P[S \mid X, Z = z]}\Big|_{\epsilon=0} \\
&= E_P[S(Y_1 - Y_0) \mid X, Z = z] - \delta_{Y,z}(X) E_P[S \mid X, Z = z] \\
&= E_P\left[S\frac{(Y_1 - Y_0 - \delta_{Y,z}(X))I\{Z = z\}}{z\pi_Z(X) + (1-z)(1-\pi_Z(X))} \mid X\right],
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial \epsilon}\delta_{A,z,\epsilon}(X)\Big|_{\epsilon=0} &= \frac{\partial}{\partial \epsilon} E_{P_\epsilon}[A_1 - A_0 \mid X, Z = z]\Big|_{\epsilon=0} \\
&= \frac{\partial}{\partial \epsilon}\frac{\delta_{A,z}(X) + \epsilon E_P[S(A_1 - A_0) \mid X, Z = z]}{1 + \epsilon E_P[S \mid X, Z = z]}\Big|_{\epsilon=0} \\
&= E_P[S(A_1 - A_0) \mid X, Z = z] - \delta_{A,z}(X) E_P[S \mid X, Z = z] \\
&= E_P\left[S\frac{(A_1 - A_0 - \delta_{A,z}(X))I\{Z = z\}}{z\pi_Z(X) + (1-z)(1-\pi_Z(X))} \mid X\right].
\end{aligned}$$

In summary, we obtain the efficient influence function

$$\begin{aligned}
\phi_P(O) = &\frac{E[Y_1 - Y_0 \mid X, Z = 1] - E[Y_1 - Y_0 \mid X, Z = 0]}{E[A_1 - A_0 \mid X, Z = 1] - E[A_1 - A_0 \mid X, Z = 0]} \\
&+ \frac{Z - \pi_Z(X)}{\pi_Z(X)(1-\pi_Z(X))(\delta_{A,1}(X) - \delta_{A,0}(X))^2}\{(Y_1 - Y_0)(\delta_{A,1}(X) - \delta_{A,0}(X)) \\
&-(A_1 - A_0)(\delta_{Y,1}(X) - \delta_{Y,0}(X)) + \delta_{Y,1}(X)\delta_{A,0}(X) - \delta_{Y,0}(X)\delta_{A,1}(X)\} - \Psi(P).
\end{aligned}$$

Finally, it follows to prove Theorem 5.1.3 by Equation (2.1).

A.11 Additional simulations

In this section, we report additional simulation results to illustrate how different sample sizes and the strength of the IV affect the performance of the estimated policies.

A.11.1 Sensitivity analysis

In this section, we study how the strength of the IV affects the performance of the estimated policies. The data generation process is the same as Section 4.1, except that the treatment assignment mechanism is given by

$$Pr(A_0 = 1 \mid Z, U, X) = \text{expit}(1.5 - 3Z + 0.2U_0 + 2X_1),$$
$$Pr(A_1 = 1 \mid Z, U, X) = \text{expit}(-1.5 + 2Z - 0.15U_1 + 1.5X_2),$$

for weak IV strength, and

$$Pr(A_0 = 1 \mid Z, U, X) = \text{expit}(3 - 7Z + 0.2U_0 + 2X_1),$$
$$Pr(A_1 = 1 \mid Z, U, X) = \text{expit}(-3 + 5Z - 0.15U_1 + 1.5X_2),$$

for strong IV strength, respectively. Simulation results are reported in Figures A.3 and A.4.

A.11.2 Sample size

In this section, we study how different sample sizes affect the performance of the estimated policies. The data generation process is the same as Section 4.1. The sample sizes are $n = 2500, 10000$ when using parametric models, and $n = 5000, 20000$ when using machine learning. Simulation results are reported in Figures A.5 and A.6.

A.12 Australian Longitudinal Survey

In this section, we provide supplementary information on our data analysis of the Australian Longitudinal Survey. The data can be accessed by making a request to the Australian Data Archive (Australian National University).

We follow Su et al. [2013], Cai et al. [2006] and use an index of labor market attitudes as the instrumental variable in our analysis. The survey includes seven questions about work, social roles and school attitudes towards working women. Individuals respond to these questions with scores (1) strongly agree, (2) agree, (3) don't know, (4) disagree, and (5) strongly disagree. This survey design implies that a response with a higher score indicates more positive attitude towards the education benefit of women and also their active role in the labor market. Following Su et al. [2013], we use only six out of the seven questions to construct our attitudes index, since questions 2 and 3 are actually very similar, thus might be repetitive. We choose question 2 over question 3. Summary statistics of our data from the 1984 and 1985 waves are reported in Table A.1 and A.2, respectively. Replication code is available at GitHub.

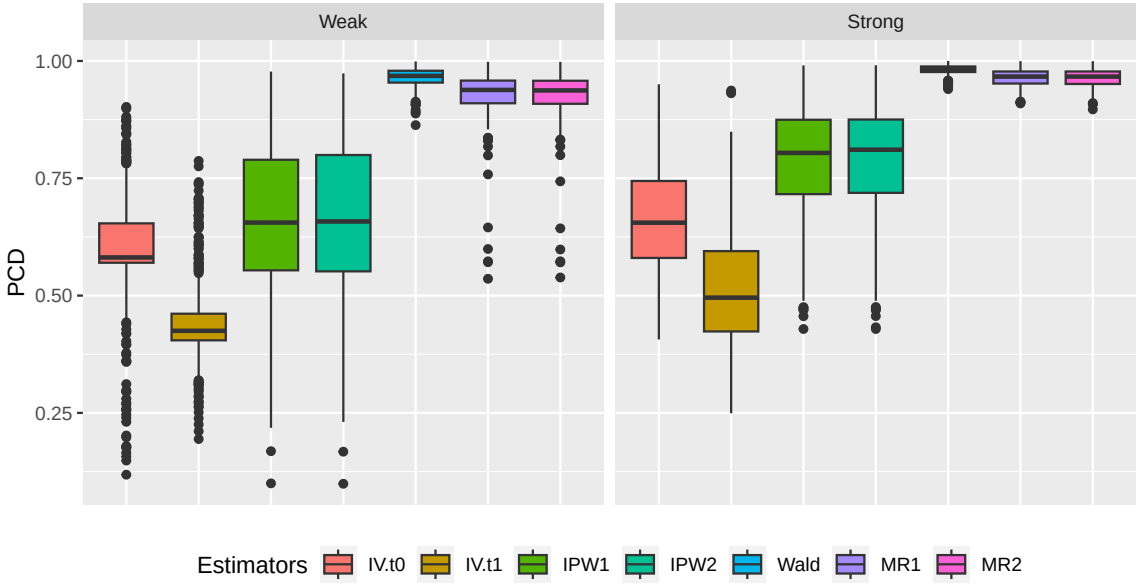


Figure A.3 – *The percentage of correct decisions (PCD) results of the estimated optimal policies using parametric models, under weak (left) or strong (right) IV strength.*

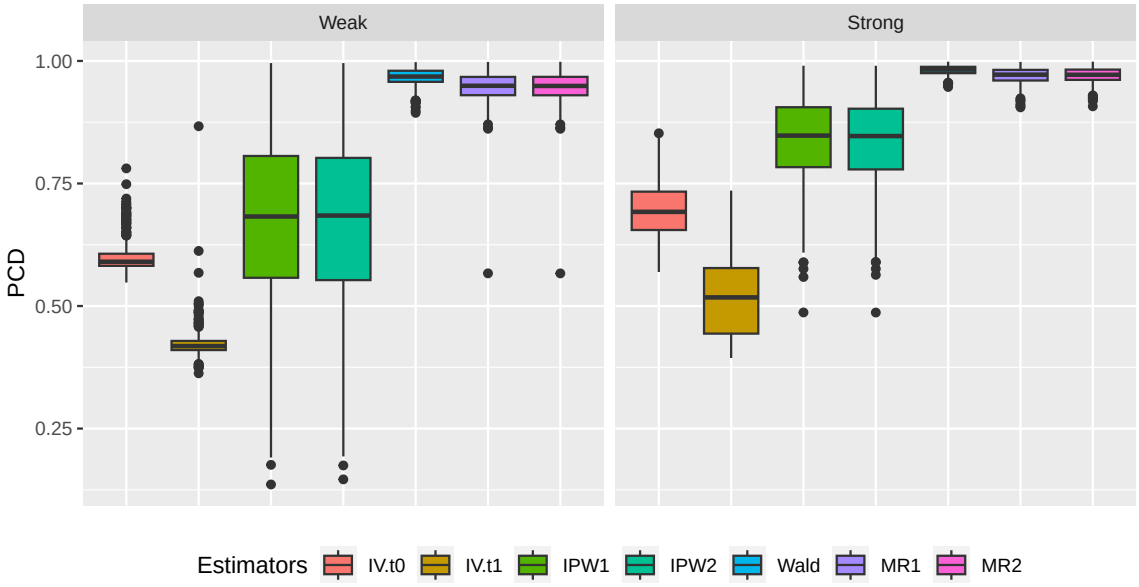


Figure A.4 – *The percentage of correct decisions (PCD) results of the estimated optimal policies using machine learning, under weak (left) or strong (right) IV strength.*

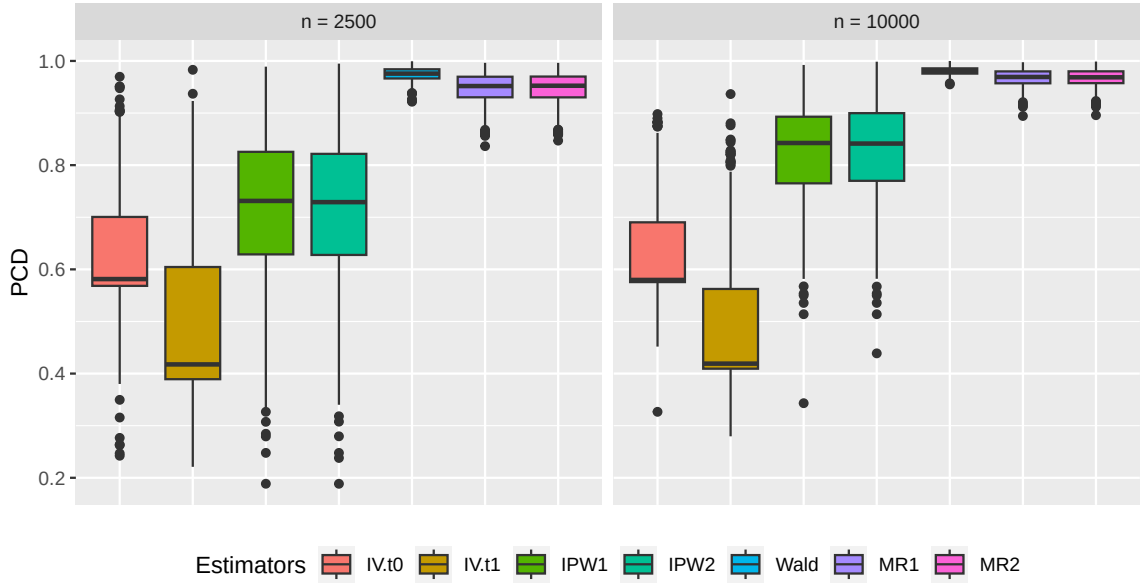


Figure A.5 – *The percentage of correct decisions (PCD) results of the estimated optimal policies, using parametric models with sample size* $n = 2500$ *(left) or* $n = 10000$ *(right).*

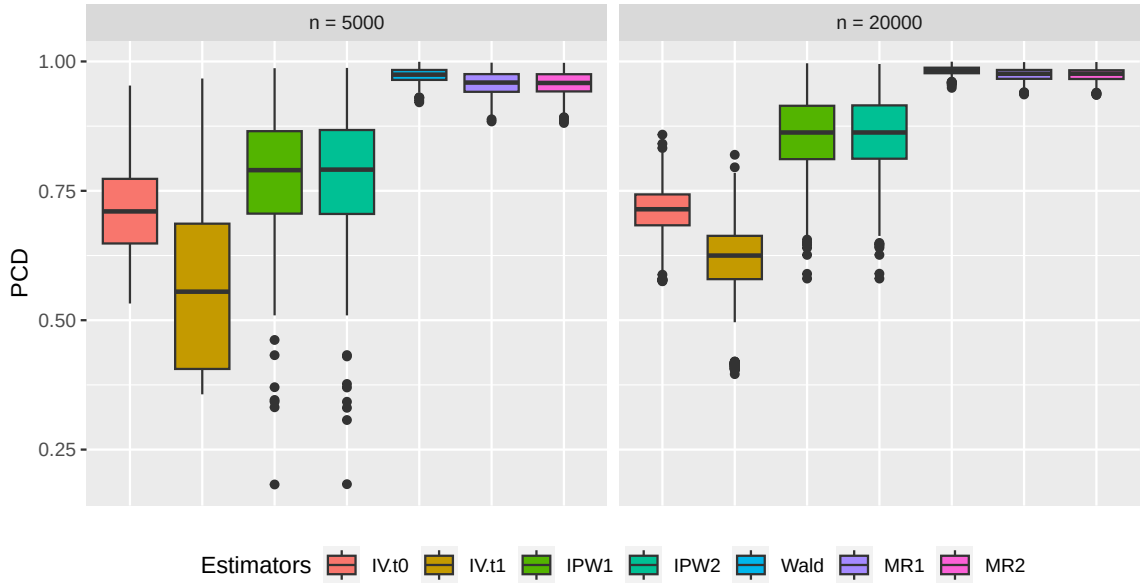


Figure A.6 – *The percentage of correct decisions (PCD) results of the estimated optimal policies, using machine learning with sample size* $n = 5000$ *(left) or* $n = 20000$ *(right).*

Variable	Source	Mean	SD	Min	Max
born_australia	A12	0.82	0.38	0	1
married	A9	0.07	0.25	0	1
uni_mem	G10	0.34	0.48	0	1
gov_emp	G9	0.21	0.41	0	1
age	A4	20.07	2.45	14	26
year_expe	F3-4, F7-10, F31-33, G21-23	0.94	1.40	0	11
attitude	O1-7	17.94	3.48	6	28
year_edu	E4, E7, E10, E14, E16, E23, E25	11.14	1.93	3	20
wage_hour	G3-5, G7-8	4.83	2.01	0.57	21.43

Table A.1 – *The 1984 wave summary statistics of variables* `born_australia`*: whether a person is born in Australia;* `married`*: marital status;* `uni_mem`*: union membership;* `gov_emp`*: government employment;* `age`*: age;* `year_expe`*: work experience;* `attitude`*: index of labor market attitudes;* `year_edu`*: education levels;* `wage_hour`*: hourly wage. Source indicates which questions in the survey provide the information.*

Variable	Source	Mean	SD	Min	Max
born_australia	B3	0.84	0.36	0	1
married	A7	0.15	0.36	0	1
uni_mem	G11	0.38	0.49	0	1
gov_emp	G10	0.22	0.42	0	1
age	A4	20.22	2.87	15	26
year_expe	F3-4, F7-10, F31-33, F23-25	1.82	2.13	0	16
attitude	O1-7	18.75	3.49	6	30
year_edu	E3, E5, E8, E12, E14, E21, E23	11.69	2.11	2	20
wage_hour	G3-5, G7-8	7.48	2.94	0.375	75.00

Table A.2 – *The 1985 wave summary statistics of variables* `born_australia`*: whether a person is born in Australia;* `married`*: marital status;* `uni_mem`*: union membership;* `gov_emp`*: government employment;* `age`*: age;* `year_expe`*: work experience;* `attitude`*: index of labor market attitudes;* `year_edu`*: education levels;* `wage_hour`*: hourly wage. Source indicates which questions in the survey provide the information.*

CHAPTER B

APPENDIX TO PART III

B.1 Proof of Proposition 6.3.1

The proof of our identification results is straightforward, following similar arguments in Kennedy [2019]. First, we prove the OR-IPS formula:

$$
\begin{aligned}
V(d) &= E[Y(d)] \\
&= E[Y(1)d(X) + Y(0)(1 - d(X))] \\
&= E[E[Y(1)d(X) + Y(0)(1 - d(X)) \mid X]] \\
&= E[d(X)E[Y(1) \mid X] + (1 - d(X))E[Y(0) \mid X]] \\
&= E[d(X)E[Y \mid X, A = 1] + (1 - d(X))E[Y \mid X, A = 0]] \\
&= E\left[\frac{\delta(X)\pi(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\mu_1(X) + \frac{1 - \pi(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\mu_0(X)\right] \\
&= E\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1 - \pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\right].
\end{aligned}
$$

Next, we prove the IPW-IPS formula:

$$
\begin{aligned}
&E\left[\frac{Y\{\delta(X)A + 1 - A\}}{\delta(X)\pi(X) + 1 - \pi(X)}\right] \\
&= E\left[\frac{YA\delta(X)}{\delta(X)\pi(X) + 1 - \pi(X)} + \frac{Y(1 - A)}{\delta(X)\pi(X) + 1 - \pi(X)}\right] \\
&= E\left[\frac{Y(1)A\delta(X)}{\delta(X)\pi(X) + 1 - \pi(X)} + \frac{Y(0)(1 - A)}{\delta(X)\pi(X) + 1 - \pi(X)}\right] \\
&= E\left[E\left[\frac{Y(1)A\delta(X)}{\delta(X)\pi(X) + 1 - \pi(X)} + \frac{Y(0)(1 - A)}{\delta(X)\pi(X) + 1 - \pi(X)} \mid X\right]\right] \\
&= E\left[\frac{E[Y(1)A \mid X]\delta(X)}{\delta(X)\pi(X) + 1 - \pi(X)} + \frac{E[Y(0)(1 - A) \mid X]}{\delta(X)\pi(X) + 1 - \pi(X)}\right] \\
&= E\left[Y(1)\frac{E[A \mid X]\delta(X)}{\delta(X)\pi(X) + 1 - \pi(X)} + Y(0)\frac{E[(1 - A) \mid X]}{\delta(X)\pi(X) + 1 - \pi(X)}\right] \\
&= E[E[Y(1)d(X) + Y(0)(1 - d(X)) \mid X]] \\
&= V(d).
\end{aligned}
$$

B.2 Proof of Proposition 6.3.2

We derive the efficient influence function for the following statistical functional:

$$\Psi(P) = E_P\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\right].$$

For a given distribution P in the nonparametric statistical model $\mathcal{M}$, we let p denote the density of P with respect to some dominating measure ν. For all bounded $h \in L_2(P)$, define the parametric submodel $p_\epsilon = (1+\epsilon h)p$, which is valid for small enough ϵ and has score h at $\epsilon = 0$. We would establish that $\Psi(P)$ is pathwise differentiable with respect to $\mathcal{M}$ at P with efficient influence function $\phi(P)$ if we have that for any $P \in \mathcal{M}$,

$$\frac{\partial}{\partial\epsilon}\Psi(P_\epsilon)\bigg|_{\epsilon=0} = \int \phi(P)(o)h(o)dP(o).$$

We denote $\pi_\epsilon(x) = E_{P_\epsilon}[A \mid X = x]$, $\mu_{a,\epsilon}(x) = E_{P_\epsilon}[Y \mid X = x, A = a]$, $S = \partial \log p_\epsilon/\partial\epsilon$, and can compute

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\Psi(P_\epsilon)\bigg|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}E_{P_\epsilon}\left[\frac{\delta(X)\pi_\epsilon(X)\mu_{1,\epsilon}(X) + \{1-\pi_\epsilon(X)\}\mu_{0,\epsilon}(X)}{\delta(X)\pi_\epsilon(X) + 1 - \pi_\epsilon(X)}\right]\bigg|_{\epsilon=0} \\
&= \frac{\partial}{\partial\epsilon}E_P\left[(1+\epsilon S)\frac{\delta(X)\pi_\epsilon(X)\mu_{1,\epsilon}(X) + \{1-\pi_\epsilon(X)\}\mu_{0,\epsilon}(X)}{\delta(X)\pi_\epsilon(X) + 1 - \pi_\epsilon(X)}\right]\bigg|_{\epsilon=0} \\
&= E_P\left[S\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\right] \\
&\quad + E_P\left[\frac{1}{\delta(X)\pi(X) + 1 - \pi(X)}\left(\pi(X)\frac{\partial}{\partial\epsilon}\mu_{1,\epsilon}(X)\bigg|_{\epsilon=0} + \mu_1(X)\frac{\partial}{\partial\epsilon}\pi_\epsilon(X)\bigg|_{\epsilon=0}\right)\right] \\
&\quad + E_P\left[\frac{1}{\delta(X)\pi(X) + 1 - \pi(X)}\left(\{1-\pi(X)\}\frac{\partial}{\partial\epsilon}\mu_{0,\epsilon}(X)\bigg|_{\epsilon=0} - \mu_0(X)\frac{\partial}{\partial\epsilon}\pi_\epsilon(X)\bigg|_{\epsilon=0}\right)\right] \\
&\quad - E_P\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\{\delta(X)\pi(X) + 1 - \pi(X)\}^2}\left(\delta(X)\frac{\partial}{\partial\epsilon}\pi_\epsilon(X)\bigg|_{\epsilon=0} - \frac{\partial}{\partial\epsilon}\pi_\epsilon(X)\bigg|_{\epsilon=0}\right)\right].
\end{aligned}$$

Then we need to compute

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\pi_\epsilon(X)\bigg|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\frac{\pi(X) + \epsilon E_P[SA \mid X]}{1 + \epsilon E_P[S \mid X]}\bigg|_{\epsilon=0} \\
&= E_P[SA \mid X] - \pi(X)E_P[S \mid X] \\
&= E_P[S(A - \pi(X)) \mid X],
\end{aligned}$$

and for $a = 0, 1$,

$$\begin{aligned}
\frac{\partial}{\partial\epsilon}\mu_{a,\epsilon}(X)\bigg|_{\epsilon=0} &= \frac{\partial}{\partial\epsilon}\frac{\mu_a(X) + \epsilon E_P[SY \mid X, A = a]}{1 + \epsilon E_P[S \mid X, A = a]}\bigg|_{\epsilon=0} \\
&= E_P[SY \mid X, A = a] - \mu_a(X)E_P[S \mid X, A = a] \\
&= E_P[S(Y - \mu_a(X)) \mid X, A = a].
\end{aligned}$$

Combining the above derivations, we obtain that

$$\phi(P)(O) = \frac{A\delta(X)\{Y-\mu_1(X)\} + (1-A)\{Y-\mu_0(X)\} + \delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)} + \frac{\delta(X)\tau(X)\{A-\pi(X)\}}{\{\delta(X)\pi(X) + 1 - \pi(X)\}^2} - \Psi(P),$$

which yields the result.

B.3 Proof of Theorem 7.2.1

We first outline the inferential strategy from semiparametric theory. Consider a statistical model $\mathcal{M}$ for distributions $\tilde{P}$, with P denoting the true distribution. Under sufficient smoothness conditions, we have the following von Mises expansion for $\Psi(\tilde{P})$:

$$\Psi(\tilde{P}) = \Psi(P) - \int \phi(\tilde{P})(o)dP(o) + \mathrm{Rem}(\tilde{P}, P),$$

where $\phi(P)$ is the influence function derived in Section B.2 such that $\int \phi(P)(o)dP(o) = 0$, and $\mathrm{Rem}(\tilde{P}, P) = O(\|\tilde{P} - P\|^2)$ is a second-order reminder term that we will analyze later.

Let $\hat{P}$ be an estimator of P, then we obtain the following one-step estimator of $\Psi(P)$:

$$\hat{\Psi} = \Psi(\hat{P}) + \int \phi(\hat{P})(o)dP_n(o),$$

where P_n is the empirical distribution.

Next, we characterize the asymptotic properties of $\hat{\Psi}$. Note that

$$\begin{aligned}
\hat{\Psi} - \Psi(P) &= \left\{\Psi(\hat{P}) + \int \phi(\hat{P})(o)dP_n(o)\right\} - \Psi(P) \\
&= \left\{\Psi(\hat{P}) - \Psi(P)\right\} + \int \phi(\hat{P})(o)dP_n(o) \\
&= -\int \phi(\hat{P})(o)dP(o) + \mathrm{Rem}(\hat{P}, P) + \int \phi(\hat{P})(o)dP_n(o) \\
&= \int \phi(\hat{P})(o)d\left\{P_n(o) - P(o)\right\} + \mathrm{Rem}(\hat{P}, P) \\
&= \int \phi(P)(o)dP_n(o) + \int \left\{\phi(\hat{P})(o) - \phi(P)(o)\right\} d\left\{P_n(o) - P(o)\right\} + \mathrm{Rem}(\hat{P}, P).
\end{aligned}$$

Therefore, $\sqrt{n}\left\{\hat{\Psi} - \Psi(P)\right\}$ is expressed as the following three terms:

$$\begin{aligned}
\sqrt{n}\left\{\hat{\Psi} - \Psi(P)\right\} = &\sqrt{n}\int \phi(P)(o)dP_n(o) \\
&+ \sqrt{n}\int \left\{\phi(\hat{P})(o) - \phi(P)(o)\right\} d\left\{P_n(o) - P(o)\right\} \\
&+ \sqrt{n}\mathrm{Rem}(\hat{P}, P).
\end{aligned}$$

By the central limit theorem, $\sqrt{n}\int \phi(P)(o)dP_n(o)$ is asymptotically normal with the asymptotic variance given by $E[\phi^2(P)(O)]$.

We assume that $\phi(P)$ belongs to a Donsker class, so we have that the centered empirical process

$$\sqrt{n}\int \left\{\phi(\hat{P})(o) - \phi(P)(o)\right\} d\left\{P_n(o) - P(o)\right\} = o_p(1).$$

Finally, we characterize the second-order remainder term:

$$\text{Rem}(\hat{P}, P) = \Psi(\hat{P}) - \Psi(P) + E_P[\phi(\hat{P})(O)].$$

We have that

$$\Psi(P) = E_P\left[\frac{\delta(X)\pi(X)\mu_1(X) + \{1-\pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)}\right],$$

and

$$\begin{aligned}
&E_P[\phi(\hat{P})(O)] \\
&= E_P\left[\frac{A\delta(X)\{Y-\hat{\mu}_1(X)\} + (1-A)\{Y-\hat{\mu}_0(X)\} + \delta(X)\hat{\pi}(X)\hat{\mu}_1(X) + \{1-\hat{\pi}(X)\}\hat{\mu}_0(X)}{\delta(X)\hat{\pi}(X) + 1 - \hat{\pi}(X)}\right. \\
&\quad \left. + \frac{\delta(X)\hat{\tau}(X)\{A-\hat{\pi}(X)\}}{\{\delta(X)\hat{\pi}(X) + 1 - \hat{\pi}(X)\}^2}\right] - \Psi(\hat{P}).
\end{aligned}$$

Combining the derivations above, we have that

$$\begin{aligned}
\left|\text{Rem}(\hat{P}, P)\right| \leq\ & \hat{C}_1\|\hat{\mu}_1(X) - \mu_1(X)\|_{L_2} \times \|\hat{\pi}(X) - \pi(X)\|_{L_2} \\
& + \hat{C}_2\|\hat{\mu}_0(X) - \mu_0(X)\|_{L_2} \times \|\hat{\pi}(X) - \pi(X)\|_{L_2} \\
& + \hat{C}_3\|\hat{\pi}(X) - \pi(X)\|_{L_2}^2,
\end{aligned}$$

where $\hat{C}_1$, $\hat{C}_2$ and $\hat{C}_3$ are $O_p(1)$. We assume that $\|\hat{\pi}(x) - \pi(x)\|_{L_2} = o_p(n^{-1/4})$, and $\|\hat{\mu}_a - \mu_a\|_{L_2} = o_p(n^{-1/4})$ for $a = 0, 1$. Therefore, we have that $\sqrt{n}\text{Rem}(\hat{P}, P) = o_p(1)$. That is, we conclude that

$$\sqrt{n}\left\{\hat{\Psi} - \Psi(P)\right\} \to \mathcal{N}(0, E[\phi^2(P)(O)]),$$

which completes the proof.

B.4 Proof of Theorem 7.2.2

Essentially, we need to prove that the centered empirical process is $o_p(1)$, when we avoid Donsker conditions by using the cross-fitting technique. We first review a useful lemma from Kennedy et al. [2020].

Lemma B.4.1. *Consider two independent samples $\mathcal{O}_1 = (O_1, \ldots, O_n)$ and $\mathcal{O}_2 = (O_{n+1}, \ldots, O_N)$ drawn from the distribution $\mathbb{P}$. Let $\hat{f}(o)$ be a function estimated from $\mathcal{O}_2$, and $\mathbb{P}_n$ the empirical measure over $\mathcal{O}_1$, then we have*

$$(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\hat{f} - f\|}{\sqrt{n}}\right).$$

Proof. First note that by conditioning on $\mathcal{O}_2$, we obtain that

$$\mathbb{E}\left\{\mathbb{P}_n(\hat{f} - f) \,\middle|\, \mathcal{O}_2\right\} = \mathbb{E}(\hat{f} - f \,|\, \mathcal{O}_2) = \mathbb{P}(\hat{f} - f),$$

and the conditional variance is

$$var\{(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) \,|\, \mathcal{O}_2\} = var\{\mathbb{P}_n(\hat{f} - f) \,|\, \mathcal{O}_2\} = \frac{1}{n} var(\hat{f} - f \,|\, \mathcal{O}_2) \leq \|\hat{f} - f\|^2/n,$$

therefore by the Chebyshev's inequality we have that

$$\mathbb{P}\left\{\frac{|(\mathbb{P}_n - \mathbb{P})(\hat{f} - f)|}{\|\hat{f} - f\|^2/n} \geq t\right\} = \mathbb{E}\left[\mathbb{P}\left\{\frac{|(\mathbb{P}_n - \mathbb{P})(\hat{f} - f)|}{\|\hat{f} - f\|^2/n} \geq t \,\middle|\, \mathcal{O}_2\right\}\right] \leq \frac{1}{t^2},$$

thus for any $\epsilon > 0$ we can pick $t = 1/\sqrt{\epsilon}$ so that the probability above is no more than ϵ, which yields the result. □

Next, we characterize the asymptotic properties of the cross-fitted estimator $\hat{\Psi}_{\mathrm{CF}}$. Following similar steps as Section B.3, we have that

$$\sqrt{n}\left\{\hat{\Psi}_{\mathrm{CF}} - \Psi(P)\right\} = \sqrt{n}\int \phi(P)(o) dP_n(o) + \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sqrt{n_k}(R_{k,1} + R_{k,2}),$$

where $R_{k,1} = \int \left\{\phi(\hat{P}_{-k})(o) - \phi(P)(o)\right\} d\left\{P_{n,k}(o) - P(o)\right\}$, $R_{k,2} = \mathrm{Rem}(\hat{P}_{-k}, P)$.

We note that

$$\begin{aligned} R_{k,1} &= \int \left\{\phi(\hat{P}_{-k})(o) - \phi(P)(o)\right\} d\left\{P_{n,k}(o) - P(o)\right\} \\ &= \int \left\{\xi(\hat{P}_{-k})(o) - \xi(P)(o)\right\} d\left\{P_{n,k}(o) - P(o)\right\}, \end{aligned}$$

where $\xi(P)(o) = \phi(P)(o) + \Psi(P)$, and by Lemma B.4.1, we have that

$$\sqrt{n_k} R_{k,1} = O_p\left(\|\xi(\hat{P}_{-k}) - \xi(P)\|_{L_2}\right).$$

Note that

$$\begin{aligned} &\xi(\hat{P}_{-k})(O) - \xi(P)(O) \\ &= \frac{A\delta(X)\{Y - \mu_1(X)\} + (1 - A)\{Y - \mu_0(X)\}}{\delta(X)\pi(X) + 1 - \pi(X)} - \frac{A\delta(X)\{Y - \mu_1(X)\} + (1 - A)\{Y - \mu_0(X)\}}{\delta(X)\pi(X) + 1 - \pi(X)} \\ &\quad + \frac{\delta(X)\pi(X)\mu_1(X) + \{1 - \pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)} - \frac{\delta(X)\pi(X)\mu_1(X) + \{1 - \pi(X)\}\mu_0(X)}{\delta(X)\pi(X) + 1 - \pi(X)} \\ &\quad + \frac{\delta(X)\tau(X)\{A - \pi(X)\}}{\{\delta(X)\pi(X) + 1 - \pi(X)\}^2} - \frac{\delta(X)\tau(X)\{A - \pi(X)\}}{\{\delta(X)\pi(X) + 1 - \pi(X)\}^2}, \end{aligned}$$

and we assume that $|Y|$ and $|\delta(X)|$ are bounded in probability. By the triangle and Cauchy-Schwarz inequalities, we have that

$$\|\xi(\hat{P}_{-k}) - \xi(P)\|_{L_2} \leq \hat{C}_{1,-k}\|\hat{\mu}_{0,-k}(X) - \mu_0(X)\|_{L_2} + \hat{C}_{2,-k}\|\hat{\mu}_{1,-k}(X) - \mu_1(X)\|_{L_2} + \hat{C}_{3,-k}\|\hat{\pi}_{-k}(X) - \pi(X)\|_{L_2}$$

where $\hat{C}_{1,-k}$, $\hat{C}_{2,-k}$ and $\hat{C}_{3,-k}$ are $O_p(1)$. We assume that $\|\hat{\pi}(x) - \pi(x)\|_{L_2} = o_p(n^{-1/4})$, and $\|\hat{\mu}_a - \mu_a\|_{L_2} = o_p(n^{-1/4})$ for $a = 0, 1$. Therefore, we have that $\sqrt{n_k}R_{k,1} = o_p(1)$.

By the same arguments as Section B.3, we have that $\sqrt{n_k}R_{k,2} = o_p(1)$. That is, we conclude that

$$\sqrt{n}\left\{\hat{\Psi}_{\mathrm{CF}} - \Psi(P)\right\} \to \mathcal{N}(0, E[\phi^2(P)(O)]),$$

which completes the proof.

B.5 Proof of Theorem 7.2.3

In this section, we consider a parametric policy class $\mathcal{D}(H)$ indexed by $\eta \in H$. That is, the off-policy learning task is given by the following optimization problem:

$$\begin{aligned} \eta^* = \arg\max_{\eta \in H} \quad & V(\eta), \\ \text{subject to} \quad & c(\eta) \leq 0, \end{aligned}$$

and the estimated policy is given by

$$\begin{aligned} \hat{\eta} = \arg\max_{\eta \in H} \quad & \hat{V}(\eta), \\ \text{subject to} \quad & \hat{c}(\eta) \leq 0. \end{aligned}$$

We first review a useful lemma from Shapiro [1991].

Lemma B.5.1. *Let H be a compact subset of $\mathbb{R}^k$. Let $C(H)$ denote the set of continuous real-valued functions on H, with $\mathcal{L} = C(H) \times \cdots \times C(H)$ the r-dimensional Cartesian product. Let $f(\eta) = (f_0, \ldots, f_r) \in \mathcal{L}$ be a vector of convex functions. Consider the quantity η^* defined as the solution to the following convex optimization program:*

$$\begin{aligned} \eta^* = &\arg\min_{\eta \in H} f_0(\eta), \\ &\textit{subject to } f_j(\eta) \leq 0, j = 1, \ldots, r. \end{aligned}$$

Assume that Slater's condition holds, so that there is some $\eta \in H$ for which the inequalities are satisfied and non-affine inequalities are strictly satisfied, i.e. $f_j(\eta) < 0$ if $f_j(\eta)$ is non-affine. Now consider a sequence of approximating programs, for $n = 1, 2, \ldots$:

$$\begin{aligned} \hat{\eta}_n = &\arg\min_{\eta \in H} \hat{f}_{n,0}(\eta), \\ &\textit{subject to } \hat{f}_{n,j}(\eta) \leq 0, j = 1, \ldots, r, \end{aligned}$$

with $\hat{f}_n(\eta) = \left(\hat{f}_{n,0}, \ldots, \hat{f}_{n,r}\right) \in \mathcal{L}$. *Assume that* $r(n)\left(\hat{f}_n - f\right)$ *converges in distribution to a random element* $W \in \mathcal{L}$ *for some real-valued function* $f(\eta)$. *Then*

$$r(n)\left(\hat{f}_{n,0}(\eta)(\hat{\eta}_n) - f_0(\eta^*)\right) \to L,$$

for a particular random variable L. *It follows that* $\hat{f}_{n,0}(\eta)(\hat{\eta}_n) - f_0(\eta^*) = O_p(1/r(n))$.

By Theorem 7.2.1 or 7.2.2, we have that

$$\sqrt{n}\left(\hat{V}(\eta) - V(\eta)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_V(O_i;\eta) + o_p(1),$$

and by condition (ii), we have that

$$\sqrt{n}\left(\hat{c}(\eta) - c(\eta)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_c(O_i;\eta) + o_p(1),$$

where ϕ_V and ϕ_c are the influence functions.

By condition (i) and Lemma B.5.1 with $r(n) = \sqrt{n}$, we obtain the conclusion (ii).

To prove conclusion (i), note that

$$V(\hat{\eta}) - V(\eta^*) = V(\hat{\eta}) - \hat{V}(\hat{\eta}) + \hat{V}(\hat{\eta}) - V(\eta^*),$$

where we have that $V(\hat{\eta}) - \hat{V}(\hat{\eta}) = O_p(n^{-1/2})$, and $\hat{V}(\hat{\eta}) - V(\eta^*) = O_p(n^{-1/2})$. Hence, we conclude that $V(\hat{\eta}) - V(\eta^*) = O_p(n^{-1/2})$, which completes the proof.

B.6 Proof of Theorem 7.2.4

In this section, we follow similar techniques in Li et al. [2023] and consider the off-policy learning task given by the following optimization problem:

$$d^* = \arg\max_{d\in\mathcal{D}} V(d) = \arg\max_{d\in\mathcal{D}} E[\xi(P)(O)],$$
$$\text{subject to } c(d) = E[\phi_c(P)(O)] \le 0,$$

where $\mathcal{D}$ is a Glivenko–Cantelli class, and the estimated optimal policy is given by

$$\hat{d} = \arg\max_{d\in\mathcal{D}} \hat{V}(d) = \arg\max_{d\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n}\xi(\hat{P})(O_i)$$
$$\text{subject to } \hat{c}(d) = \frac{1}{n}\sum_{i=1}^{n}\phi_c(\hat{P})(O_i) \le 0.$$

By condition (iii) of Theorems 7.2.1 or condition (ii) of Theorems 7.2.2, we have that both $\{\xi(O;d) : d \in \mathcal{D}\}$ and $\{\phi_c(O;d) : d \in \mathcal{D}\}$ are GC classes.

To simplify the notation, let we denote $\mathcal{D}_c = \{d \in \mathcal{D} : c(d) \le 0\}$, and $\mathcal{D}_{n,c} = \{d \in \mathcal{D} : \hat{c}(d) \le 0\}$. First we note that the estimation error can be expressed as

$$V(d^*) - \hat{V}(\hat{d}) = V_n^{(1)} + V_n^{(2)} + V_n^{(3)},$$

where we define

$$
\begin{aligned}
V_n^{(1)} &= \max_{d \in \mathcal{D}_c} E[\xi(P)(O)] - \max_{d \in \mathcal{D}_c} P_n \xi(P)(O), \\
V_n^{(2)} &= \max_{d \in \mathcal{D}_c} P_n \xi(P)(O) - \max_{d \in \mathcal{D}_c} P_n \xi(\hat{P})(O), \\
V_n^{(3)} &= \max_{d \in \mathcal{D}_c} P_n \xi(\hat{P})(O) - \max_{d \in \mathcal{D}_{n,c}} P_n \xi(\hat{P})(O).
\end{aligned}
$$

We analyze the three terms as follows. We have that

$$
\begin{aligned}
V_n^{(1)} &= \max_{d \in \mathcal{D}_c} E[\xi(P)(O)] - \max_{d \in \mathcal{D}_c} P_n \xi(P)(O) \\
&\leq \max_{d \in \mathcal{D}_c} |E[\xi(P)(O)] - P_n \xi(P)(O)| \\
&= o_p(1),
\end{aligned}
$$

and similarly we have that

$$
\begin{aligned}
V_n^{(2)} &= \max_{d \in \mathcal{D}_c} P_n \xi(P)(O) - \max_{d \in \mathcal{D}_c} P_n \xi(\hat{P})(O) \\
&\leq \max_{d \in \mathcal{D}_c} \left| P_n \{\xi(P)(O) - \xi(\hat{P})(O)\} \right| \\
&= o_p(1).
\end{aligned}
$$

To analyze $V_n^{(3)}$, note that for any $d \in \mathcal{D}$, we have that

$$
\begin{aligned}
&E[\phi_c(P)(O)] - P_n \phi_c(\hat{P})(O) \\
&= \{E[\phi_c(P)(O)] - P_n \phi_c(P)(O)\} + \{P_n \phi_c(P)(O) - P_n \phi_c(\hat{P})(O)\},
\end{aligned}
$$

and $E[\phi_c(P)(O)] - P_n \phi_c(P)(O)$ converges to 0 uniformly as $\{\phi_c(O; d) : d \in \mathcal{D}\}$ is a GC class, and $P_n \phi_c(P)(O) - P_n \phi_c(\hat{P})(O)$ converges to 0 uniformly by condition (ii).

Hence, $\forall \epsilon > 0$, $\exists N_1 \in \mathbb{N}$, such that for all $n > N_1$, $|E[\phi_c(P)(O)] - P_n \phi_c(\hat{P})(O)| < \epsilon$, by which we obtain that, for all $d \in \mathcal{D}_c$, i.e., $E[\phi_c(P)(O)] \leq c$, we have that $P_n \phi_c(\hat{P})(O) < c + \epsilon$. Therefore, we have that $\frac{c}{c+\epsilon} d \in \mathcal{D}_{n,c}$.

As $\xi(\hat{P})(O)$ is uniformly bounded, there exists a constant $L > 0$ such that for any d_1, d_2, we have that

$$
|\xi(\hat{P})(O; d_1) - \xi(\hat{P})(O; d_2)| \leq L \sup_{x \in \mathcal{X}} |d_1(x) - d_2(x)|.
$$

Thus, $\forall \epsilon > 0$, $\exists N_1 \in \mathbb{N}$, such that for all $n > N_1$,

$$
\begin{aligned}
V_n^{(3)} &= \max_{d \in \mathcal{D}_c} P_n \xi(\hat{P})(O) - \max_{d \in \mathcal{D}_{n,c}} P_n \xi(\hat{P})(O) \\
&\leq \max_{d \in \mathcal{D}_c} P_n \xi(\hat{P})(O) - \max_{d \in \frac{c}{c+\epsilon} \mathcal{D}_c} P_n \xi(\hat{P})(O) \\
&\leq \frac{\epsilon}{c + \epsilon} L,
\end{aligned}
$$

and similarly, we can obtain that $\exists N_2 \in \mathbb{N}$, such that for all $n > N_2$,

$$V_n^{(3)} \geq -\frac{\epsilon}{c+\epsilon} L,$$

which in combination implies that $V_n^{(3)} = o_p(1)$.

Next, we prove our result (ii) for the regret. Note that

$$V(d^*) - V(\hat{d}) = \{V(d^*) - \hat{V}(d^*)\} + \{\hat{V}(d^*) - \hat{V}(\hat{d})\} + \{\hat{V}(\hat{d}) - V(\hat{d})\}.$$

We analyze the three terms as follows. By the same argument for proving (i), we have that

$$\begin{aligned} V(d^*) - \hat{V}(d^*) &= E[\xi(P)(O; d^*)] - P_n \xi(\hat{P})(O; d^*) = o_p(1), \\ \hat{V}(\hat{d}) - V(\hat{d}) &= P_n \xi(\hat{P})(O; \hat{d}) - E[\xi(P)(O; \hat{d})] = o_p(1). \end{aligned}$$

Also by a similar argument, we have that for any $d \in \mathcal{D}$ and $\epsilon > 0$, $\exists N_2 \in \mathbb{N}$, for all $n > N_2$, $\frac{c}{c+\epsilon} d \in \mathcal{D}_{n,c}$, and

$$\begin{aligned} \hat{V}(d^*) - \hat{V}(\hat{d}) &= \hat{V}(d^*) - \hat{V}\left(\frac{c}{c+\epsilon} d^*\right) + \hat{V}\left(\frac{c}{c+\epsilon} d^*\right) - \hat{V}(\hat{d}) \\ &\leq \frac{\epsilon}{c+\epsilon} L, \end{aligned}$$

and also that for any $d \in \mathcal{D}$ and $\epsilon > 0$, $\exists N_3 \in \mathbb{N}$, for all $n > N_3$, $\frac{c}{c+\epsilon} \hat{d} \in \mathcal{D}_{n,c}$, and

$$V(d^*) - V(\hat{d}) \geq V\left(\frac{c}{c+\epsilon} \hat{d}\right) - V(\hat{d}) \geq -\frac{\epsilon}{c} L,$$

so we conclude that $V(d^*) - V(\hat{d}) = o_p(1)$, which completes the proof.

B.7 Additional simulations

In this section, we present additional simulation results.

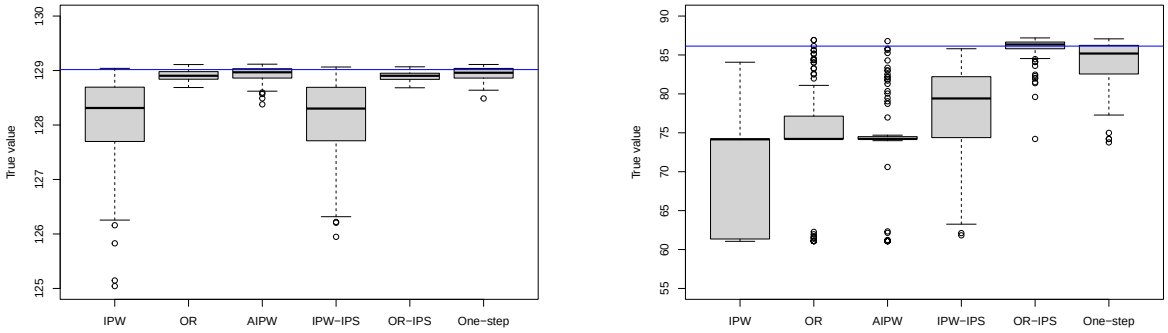
B.7.1 Incremental propensity score policy learning with sufficent overlap

We examine the performance of our proposed methods by comparison with standard policy learning methods, when sufficient overlap indeed holds. We consider the following data generating process:

$$\begin{aligned} (X_1, X_2) &\sim \text{Uniform}(0, 1), \\ (X_3, X_4) &\sim \mathcal{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix} \right\}, \\ A &\sim \text{Bernoulli}(\text{expit}(0.3 - 0.4X_1 - 0.2X_2 - 0.3X_3 + 0.1X_4)), \\ Y(0) &\sim \mathcal{N}\{20(1 + X_1 - X_2 + X_3^2 + \exp(X_2)), 20^2\}, \\ Y(1) &\sim \mathcal{N}\{20(1 + X_1 - X_2 + X_3^2 + \exp(X_2)) + 25(3 - 5X_1 + 2X_2 - 3X_3 + X_4), 20^2\}. \end{aligned}$$

We perform the vanilla direct policy search tasks without constraint. Hence, the optimal policy is simply $d^*(x) = I\{3-5X_1+2X_2-3X_3+X_4 > 0\}$. For standard methods, we consider the policy class of linear rules $\mathcal{D}_{\text{linear}} = \{d(x) = I\{(1, x_1, x_2, x_3, x_4)\beta > 0\} : \beta \in \mathbb{R}^5, \|\beta\|_2 = 1\}$. For the incremental propensity score policies, we consider the class $\mathcal{D}_{\text{IPS}} = \{d(x) = \delta(x;\beta)\pi(x)/\{\delta(x;\beta)\pi(x) + 1 - \pi(x)\} : \beta \in \mathbb{R}^5\}$, which is indexed by $\delta(x;\beta) = \exp\{(1, x_1, x_2, x_3, x_4)\beta\}$.

We estimate the outcome regression model $\mu(x)$ and the propensity score $\pi(x)$ using the generalized random forests [Athey et al., 2019] implemented in the `R` package `grf`. The unconstrained optimization problems are solved by the genetic algorithm [Sekhon and Mebane, 1998] implemented in the `R` package `rgenoud`. The sample size is $n = 2000$. We compare the true values of the estimated optimal policies using test data with sample size $N = 10^5$. The true optimal value is approximated using the test data. Simulation results of 100 Monte Carlo repetition are reported in Figure B.1(a).



(a) *Sufficient overlap.* (b) *Parametric models.*

Figure B.1 – *Performance of optimal policies under three standard methods (IPW, OR, AIPW) and our proposed three methods (IPW-IPS, OR-IPS, One-step). The blue line is the (approximate) true optimal value.*

Despite the fact that the true optimal rule is included in the standard policy class of linear rules but not in our proposed class of incremental propensity score policies, we still observe comparable performance of both classes, which exemplifies the effectiveness of our proposed methods.

B.7.2 Incremental propensity score policy learning with parametric models

We examine the performance of our proposed methods by comparison with standard policy learning methods, when using correctly specified parametric models.

The simulation setup is the same as in the main paper where the positivity assumption is violated, except that the sample size $n = 500$ is smaller and the outcome regression $\mu(s, x)$ and the propensity score $\pi(s, x)$ models are estimated by correctly specified parametric models. Simulation results of 100 Monte Carlo repetition are reported in Figure B.1(b). The standard methods IPW, OR, and AIPW have the worst

performance. The IPW-IPS estimator still has large variability, and the OR-IPS and efficient one-step estimators achieve the best performance with the highest value.

B.8 Diabetes data analysis

In this section, we provide supplementary information on our Diabetes data analysis.

The original dataset is available in the UCI Repository Diabetes 130-US hospitals for years 1999-2008 [Strack et al., 2014]. The `Fairlearn` open source project [Weerts et al., 2023] provides full dataset pre-processing script in `python` on GitHub. We follow these pre-processing steps, and provide the `R` script.

The dataset contains 101766 patients, and a detailed description of the 25 variables are available at the `Fairlearn` project. Originally, the categories of race include "African American", "Asian", "Caucasian", "Hispanic", "Other", "Unknown", and the categories of age include "30 years or younger", "$30-60$ years", "Over 60 years". We dichotomize them, so the resultant categories of race include "Caucasian" or "Non-Caucasian", and the resultant categories of age include "30 years or younger" or "Over 30 years".

The missing data are completed by multivariate imputation by chained equations, implemented in the `R` package `mice`.

CHAPTER C

APPENDIX TO PART IV

C.1 Preliminaries

C.1.1 Counting processes for Cox model

We use the counting process theory of Andersen and Gill [1982] in our theoretical framework to study the large sample properties of Cox model. We state the existing results that are used in our proof.

Let $X^{\otimes l}$ denote 1 for $l = 0$, X for $l = 1$, and XX^T for $l = 2$. Define

$$U_a^{(l)}(\beta_a, t) = \frac{1}{n_a}\sum_{i=1}^{n} I\{A_i = a\} X_i^{\otimes l} \exp(\beta_a^T X_i) Y_i(t) \text{ and } u_a^{(l)}(\beta_a, t) = \mathbb{E}\left[X^{\otimes l} \exp(\beta_a^T X) Y(t)\right],$$

where $n_a = \sum_{i=1}^{n} I\{A_i = a\}$, and define

$$E_a(\beta_a, t) = \frac{U_a^{(1)}(\beta_a, t)}{U_a^{(0)}(\beta_a, t)} \text{ and } e_a(\beta_a, t) = \frac{u_a^{(1)}(\beta_a, t)}{u_a^{(0)}(\beta_a, t)}.$$

The maximum partial likelihood estimator $\hat{\beta}_a$ for the Cox proportional hazards model solves the estimating equation

$$\mathcal{S}_{a,n}(\beta_a) = \frac{1}{n_a}\sum_{i=1}^{n} I\{A_i = a\} \int \left\{ X_i - \frac{U_1^{(1)}(\beta_a, u)}{U_1^{(0)}(\beta_a, u)} \right\} \mathrm{d}N_i(u) = 0,$$

and the cumulative baseline hazard function $\hat{\Lambda}_{0,a}$ is estimted by the Breslow estimator:

$$\hat{\Lambda}_{0,a}(t) = \int_0^t \frac{\sum_{i=1}^{n} I\{A_i = a\} \mathrm{d}N_i(u)}{\sum_{i=1}^{n} I\{A_i = a\} \exp(\hat{\beta}_a^T X_i) Y_i(u)}, a = 0, 1.$$

Under certain regularity conditions [Andersen and Gill, 1982, Conditions A – D], $\hat{\beta}_a$ and $\hat{\Lambda}_{0,a}$ converge in probability to the limits β_a^* and $\Lambda_{0,a}^*$, respectively; and we have

$$\sqrt{n_a}(\hat{\beta}_a - \beta_a^*) = \Gamma_a^{-1} \frac{1}{\sqrt{n_a}} \sum_{i=1}^{n} I\{A_i = a\} H_{a,i} + o_p(1),$$

where $\Gamma_a = \mathbb{E}[-\partial \mathcal{S}_{a,n}(\beta_a^*)/\partial \beta_a^{*T}]$ is the Fisher information matrix of β_a^*, $H_{a,i} = \int I\{A_i = a\}\{X_i - e_a(\beta_a^*, u)\}\mathrm{d}M_{a,i}(u)$ and $\mathrm{d}M_{a,i}(u) = \mathrm{d}N_i(u) - \exp(\beta_a^{*T} X_i)Y_i(u)\mathrm{d}\Lambda_{0,a}^*(u)$. Moreover, let $S^*(t \mid a, X) = \exp\{-\Lambda_{0,a}^*(t)\exp(\beta_a^{*T} X)\}$; it is shown that $\sqrt{n_a}\{\hat{S}(t \mid a, X_i) - S^*(t \mid a, X_i)\}$ converges uniformly to a mean-zero Gaussian process for all X_i.

Specifically, we consider the following expansion that we use in our proof of Theorem 9.2.2 and Corollary 9.2.5,

$$\begin{aligned}\hat{S}(t \mid a, X_i) - S^*(t \mid a, X_i) = &- S^*(t \mid a, X_i)\Lambda_{0,a}^*(t)\exp(\beta_a^{*T} X_i)X_i^T(\hat{\beta}_a - \beta_a^*) \\ &- S^*(t \mid a, X_i)\exp(\beta_a^{*T} X_i)(\hat{\Lambda}_{0,a}(t) - \Lambda_{0,a}^*(t)),\end{aligned}$$

and furthermore

$$\begin{aligned}\hat{\Lambda}_{0,a}(t) - \Lambda_{0,a}^*(t) &= \int_0^t \left\{ \frac{n_a^{-1}\sum_{i=1}^n I\{A_i = a\}\mathrm{d}N_i(u)}{U_a^{(0)}(\hat{\beta}_a, u)} - \frac{n_a^{-1}\sum_{i=1}^n I\{A_i = a\}\mathrm{d}N_i(u)}{U_a^{(0)}(\beta_a^*, u)} \right\} \\ &\quad + \int_0^t \left\{ \frac{n_a^{-1}\sum_{i=1}^n I\{A_i = a\}\mathrm{d}N_i(u)}{U_a^{(0)}(\beta_a^*, u)} - \mathrm{d}\Lambda_{0,a}^*(t) \right\} \\ &= -\left[\int_0^t \frac{U_a^{(1)}(\beta_a^*, u)}{\left\{U_a^{(0)}(\beta_a^*, u)\right\}^2} \left\{ n_a^{-1}\sum_{i=1}^n I\{A_i = a\}\mathrm{d}N_i(u) \right\} \right]^T \left(\hat{\beta}_a - \beta_a^*\right) \\ &\quad + \int_0^t \frac{n_a^{-1}\sum_{i=1}^n I\{A_i = a\}\mathrm{d}M_{a,i}(u)}{U_a^{(0)}(\beta_a^*, u)} + o_p(1) \\ &= -\left\{ \int_0^t e_a(\beta_a^*, u)\mathrm{d}\Lambda_{0,a}^*(u) \right\}^T \left(\hat{\beta}_a - \beta_a^*\right) \\ &\quad + \int_0^t \frac{n_a^{-1}\sum_{i=1}^n I\{A_i = a\}\mathrm{d}M_{a,i}(u)}{U_a^{(0)}(\beta_a^*, u)} + o_p(1).\end{aligned}$$

Combining the above two equations, we obtain

$$\begin{aligned}&\hat{S}(t \mid a, X_i) - S^*(t \mid a, X_i) \\ &= \left[-S^*(t \mid a, X_i)\Lambda_{0,a}^*(t)\exp(\beta_a^{*T} X_i)X_i^T - \left\{ \int_0^t e_a(\beta_a^*, u)\mathrm{d}\Lambda_{0,a}^*(u) \right\}^T \right] \left(\hat{\beta}_a - \beta_a^*\right) \\ &\quad + \int_0^t \frac{n_a^{-1}\sum_{i=1}^n I\{A_i = a\}\mathrm{d}M_{a,i}(u)}{U_a^{(0)}(\beta_a^*, u)} + o_p(1).\end{aligned}$$

C.1.2 Cross-fitting

To show the high-level idea of cross-fitting, we state the lemma from Kennedy et al. [2020], which is useful in our proof of Theorem 9.2.4 and Corollary 9.2.6.

Lemma C.1.1. *Consider two independent samples $\mathcal{O}_1 = (O_1, \ldots, O_n)$ and $\mathcal{O}_2 = (O_{n+1}, \ldots, O_{\tilde{n}})$, let $\hat{f}(o)$ be a function estimated from $\mathcal{O}_2$ and $\mathbb{P}_n$ the empirical measure over $\mathcal{O}_1$, then we have*

$$(\mathbb{P}_n - \mathbb{P})(\hat{f} - f) = O_{\mathbb{P}}\left(\frac{\|\hat{f} - f\|}{\sqrt{n}} \right)$$

Proof. First note that by conditioning on $\mathcal{O}_2$ we obtain

$$\mathbb{E}\left\{\mathbb{P}_n(\hat{f}-f)\,\middle|\,\mathcal{O}_2\right\} = \mathbb{E}(\hat{f}-f\,|\,\mathcal{O}_2) = \mathbb{P}(\hat{f}-f)$$

and the conditional variance is

$$var\{(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)\,|\,\mathcal{O}_2\} = var\{\mathbb{P}_n(\hat{f}-f)\,|\,\mathcal{O}_2\} = \frac{1}{n}var(\hat{f}-f\,|\,\mathcal{O}_2) \leq \|\hat{f}-f\|^2/n$$

therefore by Chebyshev's inequality we have

$$\mathbb{P}\left\{\frac{|(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)|}{\|\hat{f}-f\|^2/n} \geq t\right\} = \mathbb{E}\left[\mathbb{P}\left\{\frac{|(\mathbb{P}_n-\mathbb{P})(\hat{f}-f)|}{\|\hat{f}-f\|^2/n} \geq t\,\middle|\,\mathcal{O}_2\right\}\right] \leq \frac{1}{t^2}$$

thus for any $\epsilon > 0$ we can pick $t = 1/\sqrt{\epsilon}$ so that the probability above is no more than ϵ, which yields the result. □

C.2 Proof of Proposition 9.1.5

We first show the identification by the outcome regression formula.

$$\begin{aligned}
&V(d) = \mathbb{E}[\mathbb{E}[y(T(d))\,|\,X]] \\
&= \mathbb{E}[d(X)\mathbb{E}[y(T(1))\,|\,X] + (1-d(X))\mathbb{E}[y(T(0))\,|\,X]] \\
&= \mathbb{E}[d(X)\mathbb{E}[y(T(1))\,|\,X, I_S=1] + (1-d(X))\mathbb{E}[y(T(0))\,|\,X, I_S=1]] \\
&= \mathbb{E}[d(X)\mathbb{E}[y(T(1))\,|\,A=1, X, I_S=1] \\
&\qquad + (1-d(X))\mathbb{E}[y(T(0))\,|\,A=0, X, I_S=1]] \\
&= \mathbb{E}[d(X)\mathbb{E}[y(T)\,|\,A=1, X, I_S=1] + (1-d(X))\mathbb{E}[y(T)\,|\,A=0, X, I_S=1]] \\
&= \mathbb{E}[\mathbb{E}[y(T)\,|\,A=d(X), X, I_S=1]] \\
&= \mathbb{E}[I_T\, e(X)\mathbb{E}[y(T)\,|\,A=d(X), X, I_S=1]].
\end{aligned}$$

Similarly, we show the identification by the IPW formula.

$$\begin{aligned}
V(d) &= \mathbb{E}[\mathbb{E}[y(T)\,|\,A=d(X), X, I_S=1]] \\
&= \mathbb{E}\left[\frac{I_S}{\pi_S(X)}\mathbb{E}[y(T)\,|\,A=d(X), X, I_S=1]\right] \\
&= \mathbb{E}\left[\frac{I_S}{\pi_S(X)}\frac{I\{A=d(X)\}}{\pi_d(X)}\frac{\Delta\, y(U)}{S_C(U\,|\,A,X)}\right],
\end{aligned}$$

where the last equation follows from the standard IPTW-IPCW formula [Van der Laan and Robins, 2003].

C.3 Proof of Proposition 9.1.6

While Lee et al. [2022] derived the efficient influence function for the treatment specific survival function, here we derive the EIF for the value function $V(d) = \mathbb{E}[I_T\, e(X)\mu(d(X), X)]$.

First consider the full data $Z = (X, A, T, I_S, I_T)$, and we have the factorization as

$$p(Z) = \{p(X)\pi_S(X)p(A|X, I_S = 1)p(T|A, X, I_S = 1)\}^{I_S}\{p(X)\}^{I_T}.$$

Since $I_S I_T = 0$, the score function is $S(Z) = S(X, A, T, I_S) + I_T S(X)$. Let $V_\epsilon(d) = \mathbb{E}_\epsilon[I_T\, e(X)\mu_\epsilon(d(X), X)]$ denote the parameter of interest evaluated under the law $p_\epsilon(Z)$, where ϵ indexes a regular parametric submodel such that $p_0(Z)$ is the true data generating law. To establish that $V(d)$ is pathwise differentiable with EIF ϕ_d^F, we need to show that

$$\frac{\partial}{\partial\epsilon}V_\epsilon(d)\bigg|_{\epsilon=0} = \mathbb{E}[\phi_d^F S(Z)].$$

First, we compute

$$\frac{\partial}{\partial\epsilon}V_\epsilon(d)\bigg|_{\epsilon=0} = \mathbb{E}[I_T\, e(X)\mu(d(X), X)S(X)] + \mathbb{E}\left[\frac{\partial}{\partial\epsilon}\mu_\epsilon(d(X), X)\bigg|_{\epsilon=0}\right],$$

and further write the first term on the right hand side as

$$\begin{aligned}\mathbb{E}[I_T\, e(X)\mu(d(X), X)S(X)] &= \mathbb{E}[(I_T\, e(X)\mu(d(X), X) - V(d))S(X)] \\ &= \mathbb{E}[(I_T\, e(X)\mu(d(X), X) - V(d))S(Z)],\end{aligned}$$

and the second term as

$$\begin{aligned}&\mathbb{E}\left[\frac{\partial}{\partial\epsilon}\mu_\epsilon(d(X), X)\bigg|_{\epsilon=0}\right] \\ &= \mathbb{E}\left[d(X)\mathbb{E}[y(T)S(T \mid A, X, I_S) \mid A = 1, X, I_S = 1]\right. \\ &\qquad \left.+(1 - d(X))\mathbb{E}[y(T)S(T \mid A, X, I_S) \mid A = 0, X, I_S = 1]\right] \\ &= \mathbb{E}\left[d(X)\mathbb{E}[(y(T) - \mu(1, X))S(T \mid A, X, I_S) \mid A = 1, X, I_S = 1]\right. \\ &\qquad \left.+(1 - d(X))\mathbb{E}[(y(T) - \mu(0, X))S(T \mid A, X, I_S) \mid A = 0, X, I_S = 1]\right] \\ &= \mathbb{E}\left[d(X)\mathbb{E}\left[\frac{I_S\, A}{\pi_S(X)\pi_A(X)}(y(T) - \mu(1, X))S(T \mid A, X, I_S)\bigg|X\right]\right. \\ &\qquad \left.+(1 - d(X))\mathbb{E}\left[\frac{I_S\,(1 - A)}{\pi_S(X)(1 - \pi_A(X))}(y(T) - \mu(0, X))S(T \mid A, X, I_S)\bigg|X\right]\right] \\ &= \mathbb{E}\left[\frac{I_S}{\pi_S(X)}\left(d(X)\frac{A}{\pi_A(X)}(y(T) - \mu(1, X))\right.\right. \\ &\qquad \left.\left.+(1 - d(X))\frac{1 - A}{1 - \pi_A(X)}(y(T) - \mu(0, X))\right)S(T \mid A, X, I_S)\right] \\ &= \mathbb{E}\left[\frac{I_S}{\pi_S(X)}\frac{I\{A = d(X)\}}{\pi_d(X)}(y(T) - \mu(A, X))S(Z)\right].\end{aligned}$$

Therefore, the efficient influence function for the full data is

$$\phi_d^F = I_T\, e(X)\mu(d(X), X) + \frac{I_S}{\pi_S(X)}\frac{I\{A = d(X)\}}{\pi_d(X)}(y(T) - \mu(A, X)) - V(d).$$

Next, we consider the observed data $O = (X, A, U, \Delta, I_S, I_T)$ due to right censoring. According to Tsiatis [2006, Section 10.4], the EIF based on the observed data is given by

$$\phi_d = \frac{\Delta\, \phi_d^F}{S_C(U \mid A, X)} + \int_0^\infty \frac{L(u, A, X)}{S_C(u \mid A, X)} \mathrm{d}M_C(u \mid A, X),$$

where

$$\begin{aligned} L(u, A, X) &= \mathbb{E}[\phi_d^F \mid T \geq u, A, X] \\ &= I_T\, e(X)\mu(d(X), X) + \frac{I_S}{\pi_S(X)} \frac{I\{A = d(X)\}}{\pi_d(X)} (Q(u, A, X) - \mu(A, X)) - V(d). \end{aligned}$$

Since we have

$$\begin{aligned} \int_0^\infty \frac{\mathrm{d}M_C(u \mid A, X)}{S_C(u \mid A, X)} &= \int_0^\infty \frac{\mathrm{d}N_C(u)}{S_C(u \mid A, X)} - \int_0^U \frac{\mathrm{d}\Lambda_C(u \mid A, X)}{\exp\{\Lambda_C(u \mid A, X)\}} \\ &= 1 - \frac{\Delta}{S_C(U \mid A, X)}, \end{aligned} \tag{C.1}$$

we conclude that

$$\begin{aligned} \phi_d =& \frac{I_S}{\pi_S(X)} \frac{I\{A = d(X)\}}{\pi_d(X)} \frac{\Delta\, y(U)}{S_C(U \mid A, X)} - V(d) \\ &+ \left(I_T\, e(X) - \frac{I_S}{\pi_S(X)} \frac{I\{A = d(X)\}}{\pi_d(X)} \right) \mu(d(X), X) \\ &+ \frac{I_S}{\pi_S(X)} \frac{I\{A = d(X)\}}{\pi_d(X)} \int_0^\infty \frac{\mathrm{d}M_C(u \mid A, X)}{S_C(u \mid A, X)} Q(u, A, X). \end{aligned}$$

C.4 Proof of Theorem 9.2.2 and Corollary 9.2.5

C.4.1 Double robustness

We start with the proof of the double robustness property. We show that EIF-based estimator is consistent when either the survival outcome model or the models for the sampling score, the propensity score and the censoring process are correctly specified. Under some regularity conditions, the nuisance estimators $\hat{\mu}(a, x)$, $\hat{Q}(u, a, x)$, $\hat{\pi}_S(x)$, $\hat{\pi}_A(x)$ and $\hat{S}_C(t \mid a, x)$ converge in probability to $\mu^*(a, x)$, $Q^*(u, a, x)$, $\pi_S^*(x)$, $\pi_A^*(x)$ and $S_C^*(t \mid a, x)$, respectively. It suffices to show that $\mathbb{E}[V^*(d)] = V(d)$, where

$$\begin{aligned} V^*(d) =& I_T\, e(X)\mu^*(A = d(X), X) \\ &+ \frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} \left\{ \frac{\Delta\, y(U)}{S_C^*(U \mid A, X)} - \mu^*(A, X) \right. \\ &\left. + \int_0^\infty \frac{\mathrm{d}M_C^*(u \mid A, X)}{S_C^*(u \mid A, X)} Q^*(u, A, X) \right\} \\ =& (I) + (II) + (III). \end{aligned}$$

First, consider the case when the survival outcome model is correct, thus we have

$$(I) = \mathbb{E}[I_T\, e(X)\mu^*(A = d(X), X)] = V(d)$$

By Equation C.1, we obtain

$$
\begin{aligned}
&(II) + (III) \\
&= \frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} \left\{ y(T) - \mu^*(A, X) - \int_0^\infty \frac{\mathrm{d}M_C^*(u \mid A, X)}{S_C^*(u \mid A, X)} (y(T) - Q^*(u, A, X)) \right\}.
\end{aligned}
$$

In this case, we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} (y(T) - \mu^*(A, X)) \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} (y(T) - \mu^*(A, X)) \,\middle|\, X \right] \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} (y(T) - \mu^*(A, X)) \,\middle|\, A, X, I_S = 1 \right] \,\middle|\, X \right] \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} \mathbb{E}[(y(T) - \mu^*(A, X)) \mid A, X, I_S = 1] \,\middle|\, X \right] \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} (\mathbb{E}[y(T) \mid A, X, I_S = 1] - \mu^*(A, X)) \,\middle|\, X \right] \right] = 0.
\end{aligned}
$$

Also define $\mathrm{d}\tilde{M}_C(u \mid A, X) = \mathrm{d}\tilde{N}_C(u) - I\{C \geq u\}\mathrm{d}\Lambda_C(u \mid A, X)$ where $\tilde{N}_C(u) = I\{C \leq u\}$, so we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} \int_0^\infty \frac{\mathrm{d}M_C^*(u \mid A, X)}{S_C^*(u \mid A, X)} (y(T) - Q^*(u, A, X)) \right] \\
&= \mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} \int_0^\infty \frac{\mathrm{d}\tilde{M}_C(u \mid A, X)}{S_C^*(u \mid A, X)} I\{T \geq u\} (y(T) - Q^*(u, A, X)) \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} \int_0^\infty \frac{\mathrm{d}\tilde{M}_C^(u \mid A, X)}{S_C^*(u \mid A, X)} I\{T \geq u\} (y(T) - Q^*(u, A, X)) \,\middle|\, X \right] \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} \int_0^\infty \frac{\mathrm{d}\tilde{M}_C(u \mid A, X)}{S_C^*(u \mid A, X)} I\{T \geq u\} \right.\right.\right. \\
&\qquad \left.\left.\left. (y(T) - Q^*(u, A, X)) \,\middle|\, A, X, C, I_S = 1 \right] \,\middle|\, X \right] \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} \int_0^\infty \frac{\mathrm{d}\tilde{M}_C(u \mid A, X)}{S_C^*(u \mid A, X)} \mathbb{E}\left[I\{T \geq u\} \right.\right.\right. \\
&\qquad \left.\left.\left. (y(T) - Q^*(u, A, X)) \,\middle|\, A, X, C, I_S = 1 \right] \,\middle|\, X \right] \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A = d(X)\}}{\pi_d^*(X)} \int_0^\infty \frac{\mathrm{d}\tilde{M}_C(u \mid A, X)}{S_C^*(u \mid A, X)} \right.\right. \\
&\qquad \left.\left. (\mathbb{E}[I\{T \geq u\} y(T) \mid A, X, I_S = 1] \right.\right. \\
&\qquad \left.\left. -\mathbb{E}[I\{T \geq u\} \mid A, X, I_S = 1] Q^*(u, A, X)) \,\middle|\, X \right] \right] = 0.
\end{aligned}
$$

Next, consider the case when the models for the sampling score, the propensity score and the censoring process are correctly specified. Rearranging the terms of $V^*(d)$,

we obtain

$$\begin{aligned} V^*(d) =& \frac{I_S}{\pi_S^*(X)} \frac{I\{A=d(X)\}}{\pi_d^*(X)} \frac{\Delta\, y(U)}{S_C^*(U \mid A, X)} \\ &+ \left(I_T\, e(X) - \frac{I_S}{\pi_S^*(X)} \right) \mu^*(A = d(X), X) \\ &+ \frac{I_S}{\pi_S^*(X)} \frac{I\{A=d(X)\}}{\pi_d^*(X)} \int_0^\infty \frac{\mathrm{d}M_C^*(u \mid A, X)}{S_C^*(u \mid A, X)} Q^*(u, A, X) \\ =& (I) + (II) + (III). \end{aligned}$$

In this case, we have

$$(I) = \mathbb{E}\left[\frac{I_S}{\pi_S^*(X)} \frac{I\{A=d(X)\}}{\pi_d^*(X)} \frac{\Delta\, y(U)}{S_C^*(U \mid A, X)} \right] = V(d),$$

$$\begin{aligned} (II) &= \mathbb{E}\left[\left(I_T\, e(X) - \frac{I_S}{\pi_S^*(X)} \right) \mu^*(A = d(X), X) \right] \\ &= \mathbb{E}\left[\mathbb{E}\left[I_T\, e(X) - \frac{I_S}{\pi_S^*(X)} \middle| X \right] \mu^*(A = d(X), X) \right] = 0, \end{aligned}$$

and (III) is a stochastic integral with respect to the martingale $M_C^*(u \mid A, X)$, thus equals 0 as well, which completes the double robustness property.

C.4.2 Asymptotic properties

To establish the asymptotic results, we need some regularity conditions such that the nuisance estimators $\mu(a, x; \hat{\beta}_a, \hat{\Lambda}_{0,a})$, $Q(u, a, x; \hat{\beta}_a, \hat{\Lambda}_{0,a})$, $\pi_S(x; \hat{\lambda})$, $\pi_A(x; \hat{\theta})$ and $S_C(u \mid a, x; \hat{\alpha}_a, \hat{\Lambda}_{C0,a})$ converge in probability to $\mu(a, x; \beta_a^*, \Lambda_{0,a}^*)$, $Q(u, a, x; \beta_a^*, \Lambda_{0,a}^*)$, $\pi_S(x; \lambda^*)$, $\pi_A(x; \theta^*)$ and
$S_C(t \mid a, x; \alpha_a^*, \Lambda_{C0,a}^*)$, respectively.

Condition 4. *We assume the following conditions hold:*
(C1) X *is bounded almost surely.*
(C2) The equation $\mathbb{E}\left[\left\{ A - \frac{\exp(\theta^T X)}{1+\exp(\theta^T X)} \right\} X \right] = 0$ *has a unique solution* θ^*.
(C3) For $a = 0, 1$, *the equation*

$$\mathbb{E}\left[\int_0^L \left(X_i - \frac{\mathbb{E}[Y_i(u) \exp(\beta_a^T X) X]}{\mathbb{E}[Y_i(u) \exp(\beta_a^T X)]} \right) \times \mathrm{d}N_i(u) \right] = 0,$$

has a unique solution β_a^*, *where* $L > u$ *is a pre-specified time point such that* $Pr(U_i > L) > 0$. *Moreover, let*

$$\Lambda_{0,a}^*(u) = \mathbb{E}\left[\int_0^u \frac{\mathrm{d}N_i(u)}{\mathbb{E}[Y_i(u) \exp(\beta_a^{*T} X_i)]} \right],$$

and assume $\Lambda_{0,a}^*(L) < \infty$.
(C4) For $a = 0, 1$, *the equation*

$$\mathbb{E}\left[\int_0^L \left(X_i - \frac{\mathbb{E}[Y_i(u) \exp(\alpha_a^T X) X]}{\mathbb{E}[Y_i(u) \exp(\alpha_a^T X)]} \right) \times \mathrm{d}N_i(u) \right] = 0,$$

has a unique solution α_a^*. *Moreover, let*

$$\Lambda^*_{C0,a}(u) = \mathbb{E}\left[\int_0^u \frac{\mathrm{d}N_i(u)}{\mathbb{E}[Y_i(u)\exp(\alpha_a^{*T}X_i)]}\right],$$

and assume $\Lambda^*_{C0,a}(L) < \infty$.
(C5) *The estimating equation for the sampling score model* $\pi_S(X;\lambda)$ *has a unique solution* λ^*, *and achieves root-*n *rate of convergence.*

Under Condition 4, we have the following asymptotic representations:

$$\sqrt{n}(\hat\theta - \theta^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_{\theta i} + o_p(1), \qquad \sqrt{n}(\hat\lambda - \lambda^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_{\lambda i} + o_p(1),$$
$$\sqrt{n}(\hat\beta_a - \beta_a^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_{\beta_a i} + o_p(1), \quad \sqrt{n}(\hat\alpha_a - \alpha_a^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_{\alpha_a i} + o_p(1), \quad \text{for } a = 0, 1.$$

We focus on the estimation of survival functions by our proposed method:

$$\begin{aligned}\hat S(t;\eta) =& \frac{1}{N}\sum_{i=1}^N \Bigg[I_{T,i}\, e(X_i)\hat S(t \mid A = d_\eta(X_i), X_i) \\ &+ \frac{I_{S,i} I\{A_i = d_\eta(X_i)\}}{\hat\pi_S(X_i)\hat\pi_d(X_i)} \Bigg\{ \frac{\Delta_i\, Y_i(t)}{\hat S_C(t \mid A_i, X_i)} - \hat S(t \mid A_i, X_i) \\ &+ \int_0^\infty \frac{\hat S(t \mid A_i, X_i)\mathrm{d}\hat M_C(u \mid A_i, X_i)}{\hat S(u \mid A_i, X_i)\hat S_C(u \mid A_i, X_i)} \Bigg\} \Bigg],\end{aligned}$$

and for the ease of notation, define

$$\hat J(t,a,x) = \frac{\Delta_i\, Y_i(t)}{\hat S_C(t \mid a, x)} - \hat S(t \mid a, x) + \int_0^\infty \frac{\hat S(t \mid a, x)\mathrm{d}\hat M_C(u \mid a, x)}{\hat S(u \mid a, x)\hat S_C(u \mid a, x)},$$
$$J^*(t,a,x) = \frac{\Delta_i\, Y_i(t)}{S^*_C(t \mid a, x)} - S^*(t \mid a, x) + \int_0^\infty \frac{S^*(t \mid a, x)\mathrm{d}M^*_C(u \mid a, x)}{S^*(u \mid a, x)S^*_C(u \mid a, x)}.$$

Our proof has three main parts as follows.

PART 1. By the double robustness property shown in Section C.4.1, we have, by the strong law of large numbers and uniform consistency, that $\hat S(t;\eta) = S(t;\eta) + o_p(1)$, which proves (i) of Theorem 9.2.2. Moreover, define

$$S^*_N(t;\eta) = \frac{1}{N}\sum_{i=1}^N \left[I_{T,i}\, e(X_i) S^*(t \mid A = d_\eta(X_i), X_i) + \frac{I_{S,i} I\{A_i = d_\eta(X_i)\}}{\pi^*_S(X_i)\pi^*_d(X_i)} J^*(t, A_i, X_i)\right],$$

and by applying the Taylor expansion and the counting processes result in Section C.1.1, we obtain

$$\begin{aligned}\hat S(t;\eta) =& S^*_n(t;\eta) + H^T_\lambda(\hat\lambda - \lambda^*) + H^T_\theta(\hat\theta - \theta^*) + H^T_{\beta_0}(\hat\beta_0 - \beta_0^*) + H^T_{\beta_1}(\hat\beta_1 - \beta_1^*) \\ &+ H^T_{\alpha_0}(\hat\alpha_0 - \alpha_0^*) + H^T_{\alpha_1}(\hat\alpha_1 - \alpha_1^*) + R_S + o_p(N^{-1/2}),\end{aligned}$$

where

$$H_\lambda = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\hat{S}(t;\eta)}{\partial\lambda^*}, H_\theta = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\hat{S}(t;\eta)}{\partial\theta^*},$$

$$H_{\beta_a} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\left\{I_{T,i}\,e(X_i)(-1)^{a+1}G(t,a,X_i) + \frac{I_{S,i}I\{A_i=a\}}{\pi_S^*(X_i)\pi_d^*(X_i)}\left(\int_0^\infty \frac{G(t,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S^*(u\,|\,a,X_i)S_C^*(u\,|\,a,X_i)}\right.\right.$$
$$\left.\left. -G(t,a,X_i) - \int_0^\infty \frac{G(u,a,X_i)S^*(t\,|\,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S^{*2}(u\,|\,a,X_i)S_C^*(u\,|\,a,X_i)}\right)\right\},$$

$$H_{\alpha_a} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{I_{S,i}I\{A_i=a\}}{\pi_S^*(X_i)\pi_d^*(X_i)}\left\{\frac{-\Delta_i Y_i(t)}{S_C^*(t\,|\,a,X_i)}G_C(t,a,X_i)\right.$$
$$\left. - \int_0^\infty \frac{G_C(u,a,X_i)S^*(t\,|\,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^{*2}(u\,|\,a,X_i)S^*(u\,|\,a,X_i)} + \tilde{G}_C(t,a,X_i)\right\},$$

$$\begin{aligned}R_S = \frac{1}{N}\sum_{i=1}^{N}\sum_{a=0,1}&\left\{I_{T,i}\,e(X_i)(-1)^{a+1}H(t,a,X_i)\right.\\
&+\frac{I_{S,i}I\{A_i=a\}}{\pi_S^*(X_i)\pi_d^*(X_i)}\left(\int_0^\infty \frac{H(t,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^*(u\,|\,a,X_i)S^*(u\,|\,a,X_i)} - H(t,a,X_i)\right.\\
&-\int_0^\infty \frac{H(u,a,X_i)S^*(t\,|\,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^*(u\,|\,a,X_i)S^{*2}(u\,|\,a,X_i)} - \frac{\Delta_i Y_i(t)}{S_C^*(t|a,X_i)}H_C(t,a,X_i)\\
&\left.\left.-\int_0^\infty \frac{H_C(u,a,X_i)S^*(t\,|\,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^{*2}(u\,|\,a,X_i)S^*(u\,|\,a,X_i)} - \tilde{H}_C(t,a,X_i)\right)\right\}\\
=\frac{1}{N}\sum_{i=1}^{N}&\phi_{Rs,i},\end{aligned}$$

with

$$G(t,a,x) = -S^*(t\,|\,a,x)\Lambda_{0,a}^*(t)x^T + S^*(t\,|\,a,x)\exp(\beta_a^{*T}x)\left\{\int_0^t e_a(\beta_a^*,u)\mathrm{d}\Lambda_{0,a}^*(u)\right\}^T,$$

$$H(t,a,x) = -S^*(t\,|\,a,x)\exp(\beta_a^{*T}x)\int_0^t \frac{n_a^{-1}\sum_{i=1}^n I\{A_i=a\}\mathrm{d}M_{a,i}(u)}{U_a^{(0)}(\beta_a^*,u)},$$

$$G_C(t,a,x) = -S^*(t\,|\,a,x)\Lambda_{0,a}^*(t)x^T + S^*(t\,|\,a,x)\exp(\beta_a^{*T}x)\left\{\int_0^t e_a(\beta_a^*,u)\mathrm{d}\Lambda_{0,a}^*(u)\right\}^T,$$

$$H_C(t,a,x) = -S^*(t\,|\,a,x)\exp(\beta_a^{*T}x)\int_0^t \frac{n_a^{-1}\sum_{i=1}^n I\{A_i=a\}\mathrm{d}M_{a,i}(u)}{U_a^{(0)}(\beta_a^*,u)},$$

$$\tilde{G}_C(t,a,x) = \int_0^{U_i}\frac{S^*(t\,|\,a,x)\mathrm{d}\Lambda_C^*(u\,|\,a,x)}{S_C^*(u\,|\,a,x)S^*(u\,|\,a,x)}x^T + \left\{\int_0^t \frac{S^*(t\,|\,a,x)e_a(\beta_a^*,u)\mathrm{d}\Lambda_{0,a}^*(u)}{S_C^*(u\,|\,a,x)S^*(u\,|\,a,x)}\right\}^T,$$

$$\tilde{H}_C(t,a,x) = \int_0^t \frac{S^*(t\,|\,a,x)n_a^{-1}\sum_{i=1}^n I\{A_i=a\}\mathrm{d}M_{a,i}(u)}{S_C^*(u\,|\,a,x)S^*(u\,|\,a,x)U_a^{(0)}(\beta_a^*,u)}.$$

Thus, we have

$$\sqrt{N}\left\{\hat{S}(t;\eta) - S(t;\eta)\right\} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\xi_{1,i}(t;\eta) + \xi_{2,i}(t;\eta)) + o_p(1), \tag{C.2}$$

where

$$\xi_{1,i}(t;\eta) = S_n^*(t;\eta) - S(t;\eta),$$
$$\xi_{2,i}(t;\eta) = H_\lambda^T \phi_{\lambda^*,i} + H_\theta^T \phi_{\theta^*,i} + \sum_{a=0,1} H_{\beta_a}^T \phi_{\beta_0^*,i} + \sum_{a=0,1} H_{\alpha_a}^T \phi_{\alpha_a^*,i} + H_{\alpha_1}^T + \phi_{Rs,i},$$

and $\xi_{1,i}(t;\eta), \xi_{2,i}(t;\eta)$ are independent mean-zero processes. Therefore, we obtain that $\sqrt{N}\left\{\hat{S}(t;\eta) - S(t;\eta)\right\}$ converges weakly to a mean-zero Gaussian process, which proves (ii) of Theorem 9.2.2.

PART 2. We show that $N^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$. Recall that

$$\hat{\eta} = \arg\max_\eta \hat{S}(t;\eta) \text{ and } \eta^* = \arg\max_\eta S(t;\eta).$$

By Assumption 9.2.1 (i), $S(t;\eta)$ is twice continuously differentiable at a neighborhood of η^*; in Step 1, we show that $\hat{S}(t;\eta) = S(t;\eta) + o_p(1), \forall \eta$; since $\hat{\eta}$ maximizes $\hat{S}(t;\eta)$, we have that $\hat{S}(t;\hat{\eta}) \geq \sup_\eta \hat{S}(t;\eta)$, thus by the Argmax theorem, we have $\hat{\eta} \xrightarrow{p} \eta^*$ as $N \to \infty$.

In order to establish the $N^{-1/3}$ rate of convergence of $\hat{\eta}$, we apply Theorem 14.4 (Rate of convergence) of Kosorok [2008], and need to find the suitable rate that satisfies three conditions below.

Condition 1 For every η in a neighborhood of η^* such that $\|\eta - \eta^*\|_2 < \delta$, by Assumption 9.2.1 (i), we apply the second-order Taylor expansion,

$$\begin{aligned} S(t;\eta) - S(t;\eta^*) &= S'(\eta^*)\|\eta - \eta^*\|_2 + \frac{1}{2}S''(\eta^*)\|\eta - \eta^*\|_2^2 + o(\|\eta - \eta^*\|_2^2) \\ &= \frac{1}{2}S''(\eta^*)\|\eta - \eta^*\|_2^2 + o(\|\eta - \eta^*\|_2^2), \end{aligned}$$

and as $S''(\eta^*) < 0$, there exists $c_0 = -\frac{1}{2}S''(\eta^*) > 0$ such that $S(t;\eta) - S(t;\eta^*) \leq -c_0\|\eta - \eta^*\|_2^2$.

Condition 2 For all N large enough and sufficiently small δ, we consider the centered process $\hat{S} - S$, and have that

$$\begin{aligned} &\mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{S}(t;\eta) - S(t;\eta) - \left\{\hat{S}(t;\eta^*) - S(t;\eta^*)\right\}\right|\right] \\ &= \mathbb{E}\Bigg[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\Big|\hat{S}(t;\eta) - S_n^*(t;\eta) + S_n^*(t;\eta) - S(t;\eta) \\ &\qquad - \left\{\hat{S}(t;\eta^*) - S_n^*(t;\eta^*) + S_n^*(t;\eta^*) - S(t;\eta^*)\right\}\Big|\Bigg] \\ &\leq \mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{S}(t;\eta) - S_n^*(t;\eta) - \left\{\hat{S}(t;\eta^*) - S_n^*(t;\eta^*)\right\}\right|\right] \qquad (I) \\ &\quad + \mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\left|S_n^*(t;\eta) - S(t;\eta) - \{S_n^*(t;\eta^*) - S(t;\eta^*)\}\right|\right], \qquad (II) \end{aligned}$$

and we bound (I) and (II) respectively as follows.

Condition 2.1 To bound (II), we need the useful facts that

$$I\{A = d_\eta(X)\} - I\{A = d_{\eta^*}(X)\} = (2A-1)(d_\eta(X) - d_{\eta^*}(X)),$$

$$S^*(t \,|\, d_\eta(X_i), X_i) - S^*(t \,|\, d_{\eta^*}(X_i), X_i) = (S^*(t \,|\, 1, X_i) - S^*(t \,|\, 0, X_i))(d_\eta(X_i) - d_{\eta^*}(X_i)),$$

and obtain

$$S_n^*(t;\eta) - S_n^*(t;\eta^*) = \frac{1}{N}\sum_{i=1}^{N}(d_\eta(X_i) - d_{\eta^*}(X_i)) \times \left\{ I_{T,i}\, e(X_i)(S^*(t \,|\, 1, X_i) - S^*(t \,|\, 0, X_i)) + \frac{(2A_i - 1)I_{S,i}}{\pi_S^*(X_i)\pi_d^*(X_i)} J^*(t, A_i, X_i) \right\}.$$

Define a class of functions

$$\mathcal{F}_\eta^1 = \left\{ (d_\eta(x) - d_{\eta^*}(x)) \left(I_T\, e(x)(S^*(t \,|\, 1, x) - S^*(t \,|\, 0, x)) + \frac{(2a-1)I_S}{\pi_a^*(x)\pi_S^*(x)} J^*(t, a, x) \right) : \|\eta - \eta^*\|_2 < \delta \right\},$$

and let $M_1 = \sup \left| I_T\, e(x)(S^*(t \,|\, 1, x) - S^*(t \,|\, 0, x)) + \frac{(2a-1)I_S}{\pi_a^*(x)\pi_S^*(x)} J^*(t, a, x) \right|$. By Assumption 9.1.1, 9.1.3 and Condition 4, we have that $M_1 < \infty$.

When $\|\eta - \eta^*\|_2 < \delta$, by Condition 4 *(C1)*, there exists a constant $0 < k_0 < \infty$ such that $|(1, x^T)(\eta - \eta^*)| < k_0\delta$; furthermore, we show that $|d_\eta(x) - d_{\eta^*}(x)| = |I\{(1, x^T)\eta > 0\} - I\{(1, x^T)\eta^* > 0\}| \le I\{-k_0\delta \le (1, x^T)\eta^* \le k_0\delta\}$, by considering the three cases:

— when $-k_0\delta \le (1, x^T)\eta^* \le k_0\delta$, we have $|d_\eta(x) - d_{\eta^*}(x)| \le 1 = I\{-k_0\delta \le (1, x^T)\eta^* \le k_0\delta\}$;

— when $(1, x^T)\eta^* > k_0\delta > 0$, we have $(1, x^T)\eta = (1, x^T)(\eta - \eta^*) + (1, x^T)\eta^* > 0$, so $|d_\eta(x) - d_{\eta^*}(x)| = 0 = I\{-k_0\delta \le (1, x^T)\eta^* \le k_0\delta\}$;

— when $(1, x^T)\eta^* < -k_0\delta < 0$, we have $(1, x^T)\eta = (1, x^T)(\eta - \eta^*) + (1, x^T)\eta^* < 0$, so $|d_\eta(x) - d_{\eta^*}(x)| = 0 = I\{-k_0\delta \le (1, x^T)\eta^* \le k_0\delta\}$.

Thus we can define the envelope of $\mathcal{F}_\eta^1$ as $F_1 = M_1 I\{-k_0\delta \le (1, x^T)\eta^* \le k_0\delta\}$. By Assumption 9.2.1 *(ii)*, there exists a constant $0 < k_1 < \infty$ such that

$$\|F_1\|_{P,2} \le M_1 \sqrt{Pr(-k_0\delta \le (1, x^T)\eta^* \le k_0\delta)} \le M_1 \sqrt{2k_0k_1}\delta^{1/2}.$$

By Lemma 9.6 and Lemma 9.9 of Kosorok [2008], we have that $\mathcal{F}_\eta^1$, a class of indicator functions, is a Vapnik-Cervonenkis (VC) class with bounded bracketing entropy $J_{[]}^*(1, \mathcal{F}_\eta^1) < \infty$.

Since we have the fact that

$$\mathbb{G}_N \mathcal{F}_\eta^1 = N^{-1/2} \sum_{i=1}^{N} \left\{ \mathcal{F}_\eta^1 - \mathbb{E}[\mathcal{F}_\eta^1] \right\} = \sqrt{N} \left(S_n^*(t;\eta) - S_n^*(t;\eta^*) - \{S(t;\eta) - S(t;\eta^*)\} \right),$$

By Theorem 11.2 of Kosorok [2008], we obtain that there exists a constant $0 < c_1 < \infty$,

$$(II) = \mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta} |\mathbb{G}_N \mathcal{F}_\eta^1|\right] \leq c_1 J_{[]}^*(1, \mathcal{F}_\eta^1)\|F_1\|_{P,2} \leq c_1 J_{[]}^*(1, \mathcal{F}_\eta^1) M_1 \sqrt{2k_0 k_1}\delta^{1/2} = \tilde{c}_1 \delta^{1/2},$$

so we conclude that $(II) \leq \tilde{c}_1 \delta^{1/2}$ where $\tilde{c}_1 > 0$ is a finite constant.

Condition 2.2 To bound (I), first we have

$$\begin{aligned}
&\hat{S}(t;\eta) - S_n^*(t;\eta) - \{\hat{S}(t;\eta^*) - S_n^*(t;\eta^*)\} = \hat{S}(t;\eta) - \hat{S}(t;\eta^*) - \{S_n^*(t;\eta) - S_n^*(t;\eta^*)\} \\
&= \frac{1}{N}\sum_{i=1}^N (d_\eta(X_i) - d_{\eta*}(X_i)) \Big[I_{T,i}\, e(X_i)\{\hat{S}(t|1,X_i) - \hat{S}(t|0,X_i) - (S^*(t|1,X_i) - S^*(t|0,X_i))\} \\
&\quad + \frac{(2A_i-1)I_{S,i}}{\hat{\pi}_{A_i}(X_i)\hat{\pi}_S(X_i)}\hat{J}(t,A_i,X_i) - \frac{(2A_i-1)I_{S,i}}{\pi_{A_i}^*(X_i)\pi_S^*(X_i)} J^*(t,A_i,X_i)\Big],
\end{aligned}$$

and then apply the Taylor expansion and counting processes result in Section C.1.1,

$$\begin{aligned}
&\hat{S}(t;\eta) - S_n^*(t;\eta) - \{\hat{S}(t;\eta^*) - S_n^*(t;\eta^*)\} \\
&= \frac{1}{N}\sum_{i=1}^N (d_\eta(X_i) - d_{\eta*}(X_i)) \times \Big\{ D_\lambda(\hat{\lambda} - \lambda^*) + D_\theta(\hat{\theta} - \theta^*) + D_{\beta_0}(\hat{\beta}_0 - \beta_0^*) \\
&\quad + D_{\beta_1}(\hat{\beta}_1 - \beta_1^*) + D_{\alpha_0}(\hat{\alpha}_0 - \alpha_0^*) + D_{\alpha_1}(\hat{\alpha}_1 - \alpha_1^*) + R_{S,i}\Big\} + o_p(N^{-1/2}),
\end{aligned} \tag{C.3}$$

where

$$D_\lambda = -\frac{(2A_i-1)I_{S,i}}{\pi_{A_i}^*(X_i)\pi_S^{*2}(X_i)} J^*(t,A_i,X_i)\left(\frac{\partial \pi_S^*(X_i)}{\partial \lambda}\right)^T,$$

$$D_\theta = -\frac{I_{S,i}}{\pi_{A_i}^{*2}(X_i)\pi_S^*(X_i)} J^*(t,A_i,X_i)\left(\frac{\partial \pi_A^*(X_i)}{\partial \theta}\right)^T,$$

$$\begin{aligned}
D_{\beta_a} =& I_{T,i}\, e(X_i)(-1)^{a+1} G(t,a,X_i) + \frac{(2A_i-1)I\{A_i=a\}I_{S,i}}{\pi_{A_i}^*(X_i)\pi_S^*(X_i)}\left(\int_0^\infty \frac{G(t,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^*(u\,|\,a,X_i)S^*(u\,|\,a,X_i)}\right. \\
&\left. - G(t,a,X_i) - \int_0^\infty \frac{G(u,a,X_i)S^*(t\,|\,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^*(u\,|\,a,X_i)S^{*2}(u\,|\,a,X_i)}\right),
\end{aligned}$$

$$\begin{aligned}
D_{\alpha_a} =& \frac{(2A_i-1)I\{A_i=a\}I_{S,i}}{\pi_{A_i}^*(X_i)\pi_S^*(X_i)}\left\{-\frac{\Delta_i Y_i(t)}{S_C^*(t\,|\,a,X_i)}G_C(t,a,X_i)\right. \\
&\left. - \int_0^\infty \frac{G_C(u,a,X_i)S^*(t\,|\,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^{*2}(u\,|\,a,X_i)S^*(u\,|\,a,X_i)} + \tilde{G}_C(t,a,X_i)\right\},
\end{aligned}$$

$$\begin{aligned}
R_{S,i} = \sum_{a=0,1}\Bigg[& I_{T,i}\, e(X_i)(-1)^{a+1} H(t,a,X_i) + \frac{(2A_i-1)I\{A_i=a\}I_{S,i}}{\pi_{A_i}^*(X_i)\pi_S^*(X_i)}\Bigg(\int_0^\infty \frac{H(t,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^*(u\,|\,a,X_i)S^*(u\,|\,a,X_i)} \\
&- H(t,a,X_i) - \int_0^\infty \frac{H(u,a,X_i)S^*(t\,|\,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^*(u\,|\,a,X_i)S^{*2}(u\,|\,a,X_i)} \\
&- \frac{\Delta_i Y_i(t)}{S_C^*(t\,|\,a,X_i)}H_C(t,a,X_i) - \int_0^\infty \frac{H_C(u,a,X_i)S^*(t\,|\,a,X_i)\mathrm{d}M_C^*(u\,|\,a,X_i)}{S_C^{*2}(u\,|\,a,X_i)S^*(u\,|\,a,X_i)} - \tilde{H}_C(t,a,X_i)\Bigg)\Bigg].
\end{aligned}$$

Similarly, we define the following classes of functions:

$$\mathcal{F}_\eta^2 = \left\{ (d_\eta(x) - d_{\eta^*}(x)) \frac{(2a-1)I_{S,i}}{\pi_a^*(x)\pi_S^{*2}(x)} J^*(t,a,x) \left(\frac{\partial \pi_S^*(x)}{\partial \lambda} \right)^T : \|\eta - \eta^*\|_2 < \delta \right\},$$

$$\mathcal{F}_\eta^3 = \left\{ (d_\eta(x) - d_{\eta^*}(x)) \frac{-I_{S,i}}{\pi_a^{*2}(x)\pi_S^*(x)} J^*(t,a,x) \left(\frac{\partial \pi_A^*(x)}{\partial \theta} \right)^T : \|\eta - \eta^*\|_2 < \delta \right\},$$

$$\mathcal{F}_\eta^4 = \left\{ (d_\eta(x) - d_{\eta^*}(x)) \left[I_T\, e(x)(-1)^{a+1} G(t,a,x) + \frac{(2a-1)I_S}{\pi_a^*(x)\pi_S^*(x)} \right.\right.$$
$$\times \left(\int_0^\infty \frac{G(t,a,x)\mathrm{d}M_C^*(u \mid a,x)}{S_C^*(u \mid a,x)S^*(u \mid a,x)} - G(t,a,x) \right.$$
$$\left.\left. - \int_0^\infty \frac{G(u,a,x)S^*(t \mid a,x)\mathrm{d}M_C^*(u \mid a,x)}{S_C^*(u \mid a,x)S^{*2}(u \mid a,x)} \right) \right] : \|\eta - \eta^*\|_2 < \delta \right\},$$

$$\mathcal{F}_\eta^5 = \left\{ (d_\eta(x) - d_{\eta^*}(x)) \left[I_T\, e(x)(-1)^{a+1} G(t,a,x) + \frac{(2a-1)I_S}{\pi_a^*(x)\pi_S^*(x)} \right.\right.$$
$$\times \left(\int_0^\infty \frac{G(t,a,x)\mathrm{d}M_C^*(u \mid a,x)}{S_C^*(u \mid a,x)S^*(u \mid a,x)} - G(t,a,x) \right.$$
$$\left.\left. - \int_0^\infty \frac{G(u,a,x)S^*(t \mid a,x)\mathrm{d}M_C^*(u \mid a,x)}{S_C^*(u \mid a,x)S^{*2}(u \mid a,x)} \right) \right] : \|\eta - \eta^*\|_2 < \delta \right\},$$

$$\mathcal{F}_\eta^6 = \left\{ (d_\eta(x) - d_{\eta^*}(x)) \left[\frac{(2a-1)I_S}{\pi_a^*(x)\pi_S^*(x)} \left\{ -\frac{\Delta Y(t)}{S_C^*(t \mid a,x)} G_C(t,a,x) \right.\right.\right.$$
$$\left.\left.\left. - \int_0^\infty \frac{G_C(u,a,x)S^*(t \mid a,x)\mathrm{d}M_C^*(u \mid a,x)}{S_C^{*2}(u \mid a,x)S^*(u \mid a,x)} + \tilde{G}_C(t,a,x) \right\} \right] : \|\eta - \eta^*\|_2 < \delta \right\},$$

$$\mathcal{F}_\eta^7 = \left\{ (d_\eta(x) - d_{\eta^*}(x)) \left[\frac{(2a-1)I_S}{\pi_a^*(x)\pi_S^*(x)} \left\{ -\frac{\Delta Y(t)}{S_C^*(t \mid a,x)} G_C(t,a,x) \right.\right.\right.$$
$$\left.\left.\left. - \int_0^\infty \frac{G_C(u,a,x)S^*(t \mid a,x)\mathrm{d}M_C^*(u \mid a,x)}{S_C^{*2}(u \mid a,x)S^*(u \mid a,x)} + \tilde{G}_C(t,a,x) \right\} \right] : \|\eta - \eta^*\|_2 < \delta \right\},$$

$$\mathcal{F}_\eta^8 = \left\{ (d_\eta(x) - d_{\eta^*}(x)) \left[\sum_{a=0,1} \left[I_T\, e(x)^{a+1} H(t,a,x) + \frac{(2a-1)I_S}{\pi_a^*(x)\pi_S^*(x)} \right.\right.\right.$$
$$\times \left(\int_0^\infty \frac{H(t,a,x)\mathrm{d}M_C^*(u \mid a,x)}{S_C^*(u \mid a,x)S^*(u \mid a,x)} - H(t,a,x) \right.$$
$$- \int_0^\infty \frac{H(u,a,x)S^*(t \mid a,x)\mathrm{d}M_C^*(u \mid a,x)}{S_C^*(u \mid a,x)S^{*2}(u \mid a,x)} - \frac{\Delta Y(t)}{S_C^*(t \mid a,x)} H_C(t,a,x)$$
$$\left.\left.\left. - \int_0^\infty \frac{H_C(u,a,x)S^*(t \mid a,x)\mathrm{d}M_C^*(u \mid a,x)}{S_C^{*2}(u \mid a,x)S^*(u \mid a,x)} - \tilde{H}_C(t,a,x) \right) \right] \right] : \|\eta - \eta^*\|_2 < \delta \right\}.$$

Let

$$M_2 = \sup \left| \frac{(2a-1)}{\pi_a^*(x)} J^*(t,a,x) \left(\frac{\partial \pi_S^*(x)}{\partial \lambda} \right)^T \right|,$$

where $M_2 \in \mathbb{R}^+$ and the supremum is taken over all the coordinates; and $M_3, \ldots, M_8$ are defined accordingly for $\mathcal{F}_\eta^3, \ldots, \mathcal{F}_\eta^8$. By Assumption 9.1.1, 9.1.3 and Condition 4, we have that $M_2, \ldots, M_8 < \infty$.

Using the same technique as in **Condition 2.1**, we define the envelop of $\mathcal{F}_\eta^j$ as $F_j = M_j I\{-k_0\delta \leq (1, x^T)\eta^* \leq k_0\delta\}$ for $j = 2, \ldots, 8$, and obtain that

$$\|F_j\|_{P,2} \leq \tilde{M}_j \delta^{1/2} < \infty, \quad j = 2, \ldots, 8,$$

where $\tilde{M}_2, \ldots, \tilde{M}_8$ are some finite constants, and that $\mathcal{F}_\eta^j$ is a VC class with bounded bracketing entropy $J_{[]}^*(1, \mathcal{F}_\eta^j) < \infty$, for $j = 2, \ldots, 8$. By Theorem 11.2 of Kosorok [2008], we obtain

$$\mathbb{E}\left[\sup_{\|\eta - \eta^*\|_2 < \delta} \left| \mathbb{G}_N \mathcal{F}_\eta^j \right| \right] \leq c_j J_{[]}^*(1, \mathcal{F}_\eta^j) \|F_j\|_{P,2}, \quad j = 2, \ldots, 8,$$

where $c_2, \ldots, c_8$ are some finite constants. That is, we have

$$\mathbb{E}\left[\sup_{\|\eta - \eta^*\|_2 < \delta} \left| \mathbb{G}_N \mathcal{F}_\eta^8 \right| \right] \leq \tilde{c}_8 \delta^{1/2},$$

and furthermore by Theorem 2.14.5 of van der Vaart and Wellner [1996], we obtain

$$\begin{aligned} \left\{ \mathbb{E}\left[\sup_{\|\eta - \eta^*\|_2 < \delta} \|\mathbb{G}_n \mathcal{F}_\eta^j\|_2^2 \right] \right\}^{1/2} &\leq l_j \left\{ \mathbb{E}\left[\sup_{\|\eta - \eta^*\|_2 < \delta} |\mathbb{G}_n \mathcal{F}_\eta^j| \right] + \|F_j\|_{P,2} \right\} \\ &\leq l_j \{ c_j J_{[]}^*(1, \mathcal{F}_\eta^j) + 1 \} \|F_j\|_{P,2} \\ &\leq \tilde{c}_j \delta^{1/2}, \qquad j = 2, \ldots, 7, \end{aligned}$$

where $l_2, \ldots, l_7$ and $\tilde{c}_2, \ldots, \tilde{c}_7$ are some finite constants.

By Equation (C.3), we have that

$$
\begin{aligned}
(I) &= \mathbb{E}\left[N^{1/2}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{S}(t;\eta)-S_N^*(t;\eta)-\{\hat{S}(t;\eta^*)-S_N^*(t;\eta^*)\}\right|\right] \\
&\leq \mathbb{E}\Bigg[\sup_{\|\eta-\eta^*\|_2<\delta}\Big\{|\mathbb{G}_n\mathcal{F}_\eta^2(\hat{\lambda}-\lambda^*)|+|\mathbb{G}_n\mathcal{F}_\eta^3(\hat{\theta}-\theta^*)|+|\mathbb{G}_n\mathcal{F}_\eta^4(\hat{\beta}_0-\beta_0^*)|+|\mathbb{G}_n\mathcal{F}_\eta^5(\hat{\beta}_1-\beta_1^*)| \\
&\qquad +|\mathbb{G}_n\mathcal{F}_\eta^6(\hat{\alpha}_0-\alpha_0^*)|+|\mathbb{G}_n\mathcal{F}_\eta^7(\hat{\alpha}_1-\alpha_1^*)|+|\mathbb{G}_n\mathcal{F}_\eta^8|\Big\}+o_p(1)\Bigg] \\
&\leq N^{-1/2}\Bigg\{\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^2\cdot N^{1/2}(\hat{\lambda}-\lambda^*)|\right]+\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^3\cdot N^{1/2}(\hat{\theta}-\theta^*)|\right] \\
&\qquad +\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^4\cdot N^{1/2}(\hat{\beta}_0-\beta_0^*)|\right]+\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^5\cdot N^{1/2}(\hat{\beta}_1-\beta_1^*)|\right] \\
&\qquad +\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^6\cdot N^{1/2}(\hat{\alpha}_0-\alpha_0^*)|\right]+\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^7\cdot N^{1/2}(\hat{\alpha}_1-\alpha_1^*)|\right]\Bigg\} \\
&\quad +\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}\left|\mathbb{G}_N\mathcal{F}_\eta^8\right|\right]+o_p(1),
\end{aligned}
$$

and then by the Cauchy-Schwarz inequality, we obtain

$$
\begin{aligned}
(I) \leq\ & N^{-1/2}\left\{\mathbb{E}[N\|\hat{\lambda}-\lambda^*\|_2^2]\right\}^{1/2}\left\{\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_N\mathcal{F}_\eta^2\|_2^2\right]\right\}^{1/2} \\
&+N^{-1/2}\left\{\mathbb{E}[N\|\hat{\theta}-\theta^*\|_2^2]\right\}^{1/2}\left\{\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_N\mathcal{F}_\eta^3\|_2^2\right]\right\}^{1/2} \\
&+N^{-1/2}\left\{\mathbb{E}[N\|\hat{\beta}_0-\beta_0^*\|_2^2]\right\}^{1/2}\left\{\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_N\mathcal{F}_\eta^4\|_2^2\right]\right\}^{1/2} \\
&+N^{-1/2}\left\{\mathbb{E}[N\|\hat{\beta}_1-\beta_1^*\|_2^2]\right\}^{1/2}\left\{\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_N\mathcal{F}_\eta^5\|_2^2\right]\right\}^{1/2} \\
&+N^{-1/2}\left\{\mathbb{E}[N\|\hat{\alpha}_0-\alpha_0^*\|_2^2]\right\}^{1/2}\left\{\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_N\mathcal{F}_\eta^6\|_2^2\right]\right\}^{1/2} \\
&+N^{-1/2}\left\{\mathbb{E}[N\|\hat{\alpha}_1-\alpha_1^*\|_2^2]\right\}^{1/2}\left\{\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}\|\mathbb{G}_N\mathcal{F}_\eta^7\|_2^2\right]\right\}^{1/2} \\
&+\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}\left|\mathbb{G}_N\mathcal{F}_\eta^8\right|\right].
\end{aligned}
$$

Let $M_\lambda = \left\{\mathbb{E}[N\|\hat{\lambda}-\lambda^*\|_2^2]\right\}^{1/2}$, and $M_\theta, M_{\beta_0}, M_{\beta_1}, M_{\alpha_0}, M_{\alpha_1}$ are defined accordingly. By Condition 4, we have that $M_\lambda, M_\theta, M_{\beta_0}, M_{\beta_1}, M_{\alpha_0}, M_{\alpha_1} < \infty$, and therefore

$$
(I) \leq N^{-1/2}(M_\lambda\tilde{c}_2+M_\theta\tilde{c}_3+M_{\beta_0}\tilde{c}_4+M_{\beta_1}\tilde{c}_5+M_{\alpha_0}\tilde{c}_6+M_{\alpha_1}\tilde{c}_7)\delta^{1/2}+\tilde{c}_8\delta^{1/2}.
$$

In summary, we obtain that, let $N \to \infty$, the centered process satisfies

$$\begin{aligned}&\mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{S}(t;\eta)-S(t;\eta)-\{\hat{S}(t;\eta^*)-S(t;\eta^*)\}\right|\right]\\&\leq (I)+(II)\leq(\tilde{c}_1+\tilde{c}_8)\delta^{1/2}.\end{aligned}\tag{C.4}$$

Let $\phi_N(\delta)=\delta^{1/2}$ and $\alpha=\frac{3}{2}<2$, thus we have $\frac{\phi_n(\delta)}{\delta^\alpha}=\delta^{-1}$ is decreasing, and α does not depend on N. That is, the second condition holds.

Condition 3 By the facts that $\hat{\eta}\xrightarrow{p}\eta^*$ as $N\to\infty$, and that $\hat{S}(t;\hat{\eta})\geq\sup_\eta\hat{S}(t;\eta)$, we choose $r_N=N^{1/3}$ such that $r_N^2\phi_N(r_N^{-1})=N^{2/3}\phi_N(N^{-1/3})=N^{1/2}$. The third condition holds.

In the end, the three conditions are satisfied with $r_N=N^{1/3}$; thus we conclude that $N^{1/3}\|\hat{\eta}-\eta^*\|_2=O_p(1)$, which completes the proof of (iii) of Theorem 9.2.2.

PART 3. We characterize the asymptotic distribution of $\hat{S}(t;\hat{\eta})$. Since we have

$$\sqrt{N}\{\hat{S}(t;\hat{\eta})-S(t;\eta^*)\}=\sqrt{N}\{\hat{S}(t;\hat{\eta})-\hat{S}(t;\eta^*)\}+\sqrt{N}\{\hat{S}(t;\eta^*)-S(t;\eta^*)\},$$

we study the two terms in two steps.

Step 3.1 To establish $\sqrt{N}\{\hat{S}(t;\hat{\eta})-\hat{S}(t;\eta^*)\}=o_p(1)$, it suffices to show that $\sqrt{N}\{S(t;\hat{\eta})-S(t;\eta^*)\}=o_p(1)$ and $\sqrt{N}(\hat{S}(t;\hat{\eta})-\hat{S}(t;\eta^*)-\{S(t;\hat{\eta})-S(t;\eta^*)\})=o_p(1)$.

First, as $N^{1/3}\|\hat{\eta}-\eta^*\|_2=O_p(1)$, we take the second-order Taylor expansion

$$\begin{aligned}\sqrt{N}\{S(t;\hat{\eta})-S(t;\eta^*)\}&=\sqrt{N}\left\{S'(\eta^*)\|\hat{\eta}-\eta^*\|_2+\frac{1}{2}S''(\eta^*)\|\hat{\eta}-\eta^*\|_2^2+o_p(\|\hat{\eta}-\eta^*\|_2^2)\right\}\\&=\sqrt{N}\left\{\frac{1}{2}S''(\eta^*)\|\hat{\eta}-\eta^*\|_2^2+o_p(\|\hat{\eta}-\eta^*\|_2^2)\right\}\\&=\sqrt{N}\left\{\frac{1}{2}S''(\eta^*)O_p(N^{-2/3})+o_p(N^{-2/3})\right\}=o_p(1).\end{aligned}$$

Next, we follow the result (C.4) obtained in **PART 2**. As $N^{1/3}\|\hat{\eta}-\eta^*\|_2=O_p(1)$, there exists $\tilde{\delta}=c_9N^{-1/3}$, where $c_9<\infty$ is a finite constant, such that $\|\hat{\eta}-\eta^*\|_2\leq\tilde{\delta}$. Therefore we have

$$\begin{aligned}&\sqrt{N}(\hat{S}(t;\hat{\eta})-\hat{S}(t;\eta^*)-\{S(t;\hat{\eta})-S(t;\eta^*)\})\\&\leq\mathbb{E}\left[\sqrt{N}\sup_{\|\hat{\eta}-\eta^*\|_2<\tilde{\delta}}\left|\hat{S}(t;\hat{\eta})-S(t;\hat{\eta})-\{\hat{S}(t;\eta^*)-S(t;\eta^*)\}\right|\right]\\&\leq(\tilde{c}_1+\tilde{c}_8)\tilde{\delta}^{1/2}=(\tilde{c}_1+\tilde{c}_8)\sqrt{c_9}N^{-1/6}=o_p(1),\end{aligned}$$

which yields the result.

Step 3.2 To derive the asymptotic distribution of $\sqrt{n}\{\hat{S}(t;\eta^*)-S(t;\eta^*)\}$, we follow the result (C.2) obtained in **PART 1** and have that

$$\sqrt{N}\left\{\hat{S}(t;\eta^*)-S(t;\eta^*)\right\}\xrightarrow{D}\mathcal{N}(0,\sigma_{t,1}^2),$$

where $\sigma_{t,1}^2 = \mathbb{E}[(\xi_{1,i}(t;\eta^*) + \xi_{2,i}(t;\eta^*))^2]$. Therefore we obtain in the end

$$\begin{aligned}\sqrt{N}\{\hat{S}(t;\hat{\eta}) - S(t;\eta^*)\} &= \sqrt{N}\{\hat{S}(t;\hat{\eta}) - \hat{S}(t;\eta^*)\} + \sqrt{N}\{\hat{S}(t;\eta^*) - S(t;\eta^*)\} \\ &= o_p(1) + \sqrt{N}\{\hat{S}(t;\eta^*) - S(t;\eta^*)\} \\ &\xrightarrow{D} \mathcal{N}(0, \sigma_{t,1}^2),\end{aligned}$$

which completes the proof.

For Corollary 9.2.5 where we consider RMST, the proof can follow the same steps as before, and is thus omitted here.

C.5 Proof of Theorem 9.2.4 and Corollary 9.2.6

Our proof has three main parts below.

PART 1. Recall that the cross-fitting technique, at a high level as exemplified in Lemma C.1.1, uses sample splitting to avoid bias due to over-fitting. For simplicity, consider that the datasets $\mathcal{O}_s$ and $\mathcal{O}_t$ are randomly split into 2 folds with equal size respectively such that $\mathcal{O}_s = \mathcal{O}_{s,1} \cup \mathcal{O}_{s,2}, \mathcal{O}_t = \mathcal{O}_{t,1} \cup \mathcal{O}_{t,2}$. The extension to K-folds as described in Algorithm 1 is straightforward. Here the subscript CF is omitted to simplify the notation. Define $\mathcal{I}_1 = \mathcal{O}_{s,1} \cup \mathcal{O}_{t,1}, \mathcal{I}_2 = \mathcal{O}_{s,2} \cup \mathcal{O}_{t,2}$, and $N_1 = |\mathcal{I}_1|, N_2 = |\mathcal{I}_2|$. The cross-fitted estimator for the value function under the ITR d_η is

$$\hat{V}(\eta) = \frac{N_1}{N}\hat{V}^{\mathcal{I}_1}(\eta) + \frac{N_2}{N}\hat{V}^{\mathcal{I}_2}(\eta),$$

where

$$\begin{aligned}\hat{V}^{\mathcal{I}_1}(\eta) =& \frac{1}{N_1}\sum_{\mathcal{I}_1}\Bigg\{ I_{T,i}\, e(X_i)\hat{\mu}(d_\eta(X_i), X_i) + \frac{I_{S,i}}{\hat{\pi}_S(X_i)}\frac{I\{A_i = d_\eta(X_i)\}}{\hat{\pi}_d(X_i)} \\ &\times \left(\frac{\Delta_i\, y(U_i)}{\hat{S}_C(U_i \mid A_i, X_i)} - \hat{\mu}(A_i, X_i) + \int_0^\infty \frac{\mathrm{d}\hat{M}_C(u \mid A_i, X_i)}{\hat{S}_C(u \mid A_i, X_i)}\hat{Q}(u, A_i, X_i) \right) \Bigg\},\end{aligned}$$

and the nuisance parameters are estimated from $\mathcal{I}_2$. $\hat{V}^{\mathcal{I}_2}(\eta)$ is defined accordingly.

In this step, we show that

$$\hat{V}(\eta) - V_N(\eta) = o_p(N^{-1/2}),$$

and essentially it suffices to prove that

$$\hat{V}^{\mathcal{I}_1}(\eta) - V_N^{\mathcal{I}_1}(\eta) = o_p(N^{-1/2}),$$

where

$$\begin{aligned}V_N(\eta) =& \frac{1}{N}\sum_{i=1}^{N}\Bigg\{ I_{T,i}\, e(X_i)\mu(d_\eta(X_i), X_i) + \frac{I_{S,i}}{\pi_S(X_i)}\frac{I\{A_i = d_\eta(X_i)\}}{\pi_d(X_i)} \\ &\times \left(\frac{\Delta_i\, y(U_i)}{S_C(U_i \mid A_i, X_i)} - \mu(A_i, X_i) + \int_0^\infty \frac{\mathrm{d}M_C(u \mid A_i, X_i)}{S_C(u \mid A_i, X_i)}Q(u, A_i, X_i) \right) \Bigg\},\end{aligned}$$

and $V_N^{\mathcal{I}_1}(\eta)$ is defined accordingly.

First, we have the following decomposition

$$
\begin{aligned}
&\hat{V}^{\mathcal{I}_1}(\eta) - V_N^{\mathcal{I}_1}(\eta) \\
&= \frac{1}{N_1}\sum_{\mathcal{I}_1}\Bigg\{ I_{T,i}\, e(X_i)(\hat{\mu}(d_\eta(X_i), X_i) - \mu(d_\eta(X_i), X_i)) \\
&\quad + I_{S,i}\left(\frac{1}{\pi_S(X_i)} - \frac{1}{\hat{\pi}_S(X_i)}\right)\frac{I\{A_i = d_\eta(X_i)\}}{\pi_d(X_i)} K(A_i, X_i) \\
&\quad + \frac{I_{S,i} I\{A_i = d_\eta(X_i)\}}{\pi_S(X_i)}\left(\frac{1}{\pi_d(X_i)} - \frac{1}{\hat{\pi}_d(X_i)}\right) K(A_i, X_i) \\
&\quad + \frac{I_{S,i}}{\pi_S(X_i)}\frac{I\{A_i = d_\eta(X_i)\}}{\pi_d(X_i)}(\hat{K}(A_i, X_i) - K(A_i, X_i)) \\
&\quad + I_{S,i} I\{A_i = d_\eta(X_i)\}\left(\frac{1}{\pi_S(X_i)} - \frac{1}{\hat{\pi}_S(X_i)}\right)\left(\frac{1}{\pi_d(X_i)} - \frac{1}{\hat{\pi}_d(X_i)}\right) K(A_i, X_i) \\
&\quad + \frac{I_{S,i} I\{A_i = d_\eta(X_i)\}}{\pi_d(X_i)}\left(\frac{1}{\pi_S(X_i)} - \frac{1}{\hat{\pi}_S(X_i)}\right)(\hat{K}(A_i, X_i) - K(A_i, X_i)) \\
&\quad + \frac{I_{S,i} I\{A_i = d_\eta(X_i)\}}{\pi_S(X_i)}\left(\frac{1}{\pi_d(X_i)} - \frac{1}{\hat{\pi}_d(X_i)}\right)(\hat{K}(A_i, X_i) - K(A_i, X_i)) \\
&\quad + I_{S,i} I\{A_i = d_\eta(X_i)\}\left(\frac{1}{\pi_S(X_i)} - \frac{1}{\hat{\pi}_S(X_i)}\right)\left(\frac{1}{\pi_d(X_i)} - \frac{1}{\hat{\pi}_d(X_i)}\right)(\hat{K}(A_i, X_i) - K(A_i, X_i))\Bigg\},
\end{aligned}
\tag{C.5}
$$

where

$$
\hat{K}(A_i, X_i) = \frac{\Delta_i\, y(U_i)}{\hat{S}_C(U_i \mid A_i, X_i)} - \hat{\mu}(A_i, X_i) + \int_0^\infty \frac{\mathrm{d}\hat{M}_C(u \mid A_i, X_i)}{\hat{S}_C(u \mid A_i, X_i)}\hat{Q}(u, A_i, X_i),
$$

$$
K(A_i, X_i) = \frac{\Delta_i\, y(U_i)}{S_C(U_i \mid A_i, X_i)} - \mu(A_i, X_i) + \int_0^\infty \frac{\mathrm{d}M_C(u \mid A_i, X_i)}{S_C(u \mid A_i, X_i)} Q(u, A_i, X_i).
$$

In summary, the decomposition (C.5) consists of two types of terms: four mean-zero terms and four product terms. For the mean-zero terms, we utilize the method introduced in Section C.1.2; since

$$
\mathbb{E}[I_{T,i}\, e(X_i)(\hat{\mu}(d_\eta(X_i), X_i) - \mu(d_\eta(X_i), X_i))] = 0,
$$

by applying Lemma C.1.1, we obtain

$$
\frac{1}{N_1}\sum_{\mathcal{I}_1} I_{T,i}\, e(X_i)(\hat{\mu}(d_\eta(X_i), X_i) - \mu(d_\eta(X_i), X_i)) = o_p(N^{-1/2}).
$$

Similarly we have

$$
\mathbb{E}\left[I_{S,i}\left(\frac{1}{\pi_S(X_i)} - \frac{1}{\hat{\pi}_S(X_i)}\right)\frac{I\{A_i = d_\eta(X_i)\}}{\pi_d(X_i)} K(A_i, X_i)\right] = 0,
$$

so we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{N_1}\sum_{\mathcal{I}_1} I_{S,i}\left(\frac{1}{\pi_S(X_i)}-\frac{1}{\hat{\pi}_S(X_i)}\right)\frac{I\{A_i=d_\eta(X_i)\}}{\pi_d(X_i)}K(A_i,X_i)\right)^2\right]\\
&=\mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{N_1}\sum_{\mathcal{I}_1} I_{S,i}\left(\frac{1}{\pi_S(X_i)}-\frac{1}{\hat{\pi}_S(X_i)}\right)\frac{I\{A_i=d_\eta(X_i)\}}{\pi_d(X_i)}K(A_i,X_i)\right)^2\middle|\mathcal{I}_2\right]\right]\\
&=\mathbb{E}\left[var\left[\frac{1}{N_1}\sum_{\mathcal{I}_1} I_{S,i}\left(\frac{1}{\pi_S(X_i)}-\frac{1}{\hat{\pi}_S(X_i)}\right)\frac{I\{A_i=d_\eta(X_i)\}}{\pi_d(X_i)}K(A_i,X_i)\middle|\mathcal{I}_2\right]\right]\\
&=\frac{1}{N_1}\mathbb{E}\left[var\left[I_{S,i}\left(\frac{1}{\pi_S(X_i)}-\frac{1}{\hat{\pi}_S(X_i)}\right)\frac{I\{A_i=d_\eta(X_i)\}}{\pi_d(X_i)}K(A_i,X_i)\middle|I_2\right]\right]\\
&\leq\frac{O_p(1)}{N_1}=o_p(\frac{1}{N}).
\end{aligned}
$$

We also have

$$
\mathbb{E}\left[\frac{I_{S,i}I\{A_i=d_\eta(X_i)\}}{\pi_S(X_i)}\left(\frac{1}{\pi_d(X_i)}-\frac{1}{\hat{\pi}_d(X_i)}\right)K(A_i,X_i)\right]=0,
$$

$$
\mathbb{E}\left[\frac{I_{S,i}}{\pi_S(X_i)}\frac{I\{A_i=d_\eta(X_i)\}}{\pi_d(X_i)}(\hat{K}(A_i,X_i)-K(A_i,X_i))\right]=0,
$$

and using the same technique, we conclude that these two mean-zero terms are $o_p(N^{-1/2})$ as well.

The product terms can be handled simply by the Cauchy-Schwarz inequality and the rate of convergence conditions in Assumption 9.2.3. Additionally we have the

decomposition as follows

$$
\begin{aligned}
&\frac{1}{N_1}\sum_{\mathcal{I}_1}(\hat{K}(A_i,X_i)-K(A_i,X_i))\\
&=\frac{1}{N_1}\sum_{\mathcal{I}_1}\Bigg\{-(\hat{\mu}(A_i,X_i)-\mu(A_i,X_i))+\frac{1-\Delta_i}{S_C(U_i\,|\,A_i,X_i)}(\hat{Q}(U_i\,|\,A_i,X_i)-Q(U_i\,|\,A_i,X_i))\\
&\quad-\int_0^{U_i}\frac{\lambda_C(u\,|\,A_i,X_i)}{S_C(u\,|\,A_i,X_i)}(\hat{Q}(U_i\,|\,A_i,X_i)-Q(U_i\,|\,A_i,X_i))\mathrm{d}u\\
&\quad+(1-\Delta_i)\left(\frac{1}{\hat{S}_C(U_i\,|\,A_i,X_i)}-\frac{1}{S_C(U_i\,|\,A_i,X_i)}\right)Q(U_i\,|\,A_i,X_i)\\
&\quad+\left(\frac{1}{\hat{S}_C(U_i\,|\,A_i,X_i)}-\frac{1}{S_C(U_i\,|\,A_i,X_i)}\right)\Delta_i\,y(U_i)\\
&\quad-\int_0^{U_i}\left(\frac{\hat{\lambda}_C(u\,|\,A_i,X_i)}{\hat{S}_C(u\,|\,A_i,X_i)}-\frac{\lambda_C(u\,|\,A_i,X_i)}{S_C(u\,|\,A_i,X_i)}\right)Q(U_i\,|\,A_i,X_i)\mathrm{d}u\\
&\quad+(1-\Delta_i)\left(\frac{1}{\hat{S}_C(U_i\,|\,A_i,X_i)}-\frac{1}{S_C(U_i\,|\,A_i,X_i)}\right)(\hat{Q}(U_i\,|\,A_i,X_i)-Q(U_i\,|\,A_i,X_i))\\
&\quad-\int_0^{U_i}\left(\frac{\hat{\lambda}_C(u\,|\,A_i,X_i)}{\hat{S}_C(u\,|\,A_i,X_i)}-\frac{\lambda_C(u\,|\,A_i,X_i)}{S_C(u\,|\,A_i,X_i)}\right)(\hat{Q}(U_i\,|\,A_i,X_i)-Q(U_i\,|\,A_i,X_i))\mathrm{d}u,
\end{aligned}
$$

and similarly we have three mean-zero terms which are $o_p(N^{-1/2})$ by the same technique in Section C.1.2 and the facts that

$$\mathbb{E}[\hat{\mu}(A_i,X_i)-\mu(A_i,X_i)]=0,$$

$$
\begin{aligned}
\mathbb{E}\Bigg[&\frac{1-\Delta_i}{S_C(U_i\,|\,A_i,X_i)}(\hat{Q}(U_i\,|\,A_i,X_i)-Q(U_i\,|\,A_i,X_i))\\
&-\int_0^{U_i}\frac{\lambda_C(u\,|\,A_i,X_i)}{S_C(u\,|\,A_i,X_i)}(\hat{Q}(u\,|\,A_i,X_i)-Q(u\,|\,A_i,X_i))\mathrm{d}u\Bigg]=0,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\Bigg[&(1-\Delta_i)\left(\frac{1}{\hat{S}_C(U_i\,|\,A_i,X_i)}-\frac{1}{S_C(U_i\,|\,A_i,X_i)}\right)Q(U_i\,|\,A_i,X_i)\\
&+\left(\frac{1}{\hat{S}_C(U_i\,|\,A_i,X_i)}-\frac{1}{S_C(U_i\,|\,A_i,X_i)}\right)\Delta_i\,y(U_i)\\
&-\int_0^{U_i}\left(\frac{\hat{\lambda}_C(u\,|\,A_i,X_i)}{\hat{S}_C(u\,|\,A_i,X_i)}-\frac{\lambda_C(u\,|\,A_i,X_i)}{S_C(u\,|\,A_i,X_i)}\right)Q(U_i\,|\,A_i,X_i)\mathrm{d}u\Bigg]=0,
\end{aligned}
$$

and we can bound the two product terms as well

$$
\begin{aligned}
&\frac{1}{N_1}\sum_{\mathcal{I}_1}\left[(1-\Delta_i)\left(\frac{1}{\hat{S}_C(U_i\,|\,A_i,X_i)}-\frac{1}{S_C(U_i\,|\,A_i,X_i)}\right)(\hat{Q}(U_i\,|\,A_i,X_i)-Q(U_i\,|\,A_i,X_i))\right.\\
&\quad\left.-\int_0^{U_i}\left(\frac{\hat{\lambda}_C(u\,|\,A_i,X_i)}{\hat{S}_C(u\,|\,A_i,X_i)}-\frac{\lambda_C(u\,|\,A_i,X_i)}{S_C(u\,|\,A_i,X_i)}\right)(\hat{Q}(U_i\,|\,A_i,X_i)-Q(U_i\,|\,A_i,X_i))\mathrm{d}u\right]\\
&\leq\left[\frac{1}{N_1}\sum_{\mathcal{I}_1}(1-\Delta_i)\left(\frac{1}{\hat{S}_C(U_i\,|\,A_i,X_i)}-\frac{1}{S_C(U_i\,|\,A_i,X_i)}\right)^2\right]^{1/2}\\
&\quad\times\left[\frac{1}{N_1}\sum_{\mathcal{I}_1}(1-\Delta_i)(\hat{Q}(U_i\,|\,A_i,X_i)-Q(U_i\,|\,A_i,X_i))^2\right]^{1/2}\\
&\quad-\int_0^{U_i}\left[\frac{1}{N_1}\sum_{\mathcal{I}_1}\left(\frac{\hat{\lambda}_C(u\,|\,A_i,X_i)}{\hat{S}_C(u\,|\,A_i,X_i)}-\frac{\lambda_C(u\,|\,A_i,X_i)}{S_C(u\,|\,A_i,X_i)}\right)^2\right]^{1/2}\\
&\quad\times\left[\frac{1}{N_1}\sum_{\mathcal{I}_1}(\hat{Q}(U_i\,|\,A_i,X_i)-Q(U_i\,|\,A_i,X_i))^2\right]^{1/2}\mathrm{d}u\\
&=o_p(N^{-1/2}),
\end{aligned}
$$

which proves that $\frac{1}{N_1}\sum_{\mathcal{I}_1}(\hat{K}(A_i,X_i)-K(A_i,X_i))=o_p(N^{-1/2})$.

Therefore, we conclude that the four product terms in (C.5) are $o_p(N^{-1/2})$ as well, which completes the proof of (i) in Theorem 9.2.4.

PART 2: We show that $N^{1/3}\|\hat{\eta}-\eta^*\|_2=O_p(1)$.

By Assumption 9.2.1 (i), $V(\eta)$ is twice continuously differentiable at a neighborhood of η^*; in **PART 1**, we show that $\hat{V}(\eta)=V(\eta)+o_p(1),\forall\eta$; since $\hat{\eta}$ maximizes $\hat{V}(\eta)$, we have that $\hat{V}(\hat{\eta})\geq\sup_\eta\hat{V}(\eta)$, thus by the Argmax theorem, we have $\hat{\eta}\xrightarrow{p}\eta^*$ as $N\to\infty$.

In order to establish the $N^{-1/3}$ rate of convergence of $\hat{\eta}$, we apply Theorem 14.4 (Rate of convergence) of Kosorok [2008], and need to find the suitable rate that satisfies three conditions below.

Condition 1 For every η in a neighborhood of η^* such that $\|\eta-\eta^*\|_2<\delta$, by Assumption 9.2.1 (i), we apply the second-order Taylor expansion,

$$
\begin{aligned}
V(\eta)-V(\eta^*)&=V'(\eta^*)\|\eta-\eta^*\|_2+\frac{1}{2}V''(\eta^*)\|\eta-\eta^*\|_2^2+o(\|\eta-\eta^*\|_2^2)\\
&=\frac{1}{2}V''(\eta^*)\|\eta-\eta^*\|_2^2+o(\|\eta-\eta^*\|_2^2),
\end{aligned}
$$

and as $V''(\eta^*)<0$, there exists $c_{10}=-\frac{1}{2}V''(\eta^*)>0$ such that $V(\eta)-V(\eta^*)\leq-c_{10}\|\eta-\eta^*\|_2^2$.

Condition 2 For all N large enough and sufficiently small δ, we consider the centered

process $\hat{V} - V$, and have that

$$\begin{aligned}
&\mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{V}(\eta)-V(\eta)-\{\hat{V}(\eta^*)-V(\eta^*)\}\right|\right] \\
&=\mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{V}(\eta)-V_n(\eta)+V_n(\eta)-V(\eta)-\{\hat{V}(\eta^*)-V_n(\eta^*)+V_n(\eta^*)-V(\eta^*)\}\right|\right] \\
&\le\mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{V}(\eta)-V_n(\eta)-\{\hat{V}(\eta^*)-V_n(\eta^*)\}\right|\right] \qquad (I) \\
&\quad+\mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\left|V_n(\eta)-V(\eta)-\{V_n(\eta^*)-V(\eta^*)\}\right|\right] \qquad (II)
\end{aligned}$$

It follows from the result in **PART 1** that $(I) = o_p(1)$. To bound (II), we have

$$\begin{aligned}
&V_n(\eta)-V_n(\eta^*) \\
&=\frac{1}{N}\sum_{i=1}^{N}(d_\eta(X_i)-d_{\eta^*}(X_i))\times\left(I_{T,i}\,e(X_i)(\mu(1,X_i)-\mu(0,X_i))+\frac{(2A_i-1)I_{S,i}}{\pi_{A_i}(X_i)\pi_S(X_i)}K(A_i,X_i)\right).
\end{aligned}$$

Define a class of functions

$$\mathcal{F}_\eta^9=\left\{(d_\eta(x)-d_{\eta^*}(x))\times\left(I_T\,e(x)(\mu(1,x)-\mu(0,x))+\frac{(2a-1)I_S}{\pi_a(x)\pi_S(x)}K(a,x)\right):\|\eta-\eta^*\|_2<\delta\right\},$$

and let $M_9=\sup\left|I_T\,e(x)(\mu(1,x)-\mu(0,x))+\frac{(2a-1)I_S}{\pi_a(x)\pi_S(x)}K(a,x)\right|$. By Assumption 9.1.1, 9.1.3 and Condition 4, we have that $M_9<\infty$. Using the same technique as in Section C.4.2 **Condition 2.1**, we define the envelop of $\mathcal{F}_\eta^9$ as $F_9=M_9I\{-k_0\delta\le(1,x^T)\eta^*\le k_0\delta\}$, and obtain that $\|F_9\|_{P,2}\le\tilde{M}_9\delta^{1/2}<\infty$, where $\tilde{M}_9$ is a finite constant, and that $\mathcal{F}_\eta^9$ is a VC class with bounded entropy $J_{[]}^*(1,\mathcal{F}_\eta^9)<\infty$. By Theorem 11.2 of Kosorok [2008], we obtain

$$\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}\left|\mathbb{G}_N\mathcal{F}_\eta^9\right|\right]\le\tilde{c}_9\delta^{1/2},$$

where $\tilde{c}_9$ is a finite constant. Therefore, we obtain

$$\begin{aligned}
(II)&=\mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}|V_N(\eta)-V(\eta)-\{V_N(\eta^*)-V(\eta^*)\}|\right] \\
&=\mathbb{E}\left[\sup_{\|\eta-\eta^*\|_2<\delta}|\mathbb{G}_n\mathcal{F}_\eta^9|\right]\le\tilde{c}_9\delta^{1/2}.
\end{aligned}$$

In summary, we obtain that the centered process satisfies

$$\begin{aligned}
&\mathbb{E}\left[\sqrt{N}\sup_{\|\eta-\eta^*\|_2<\delta}\left|\hat{S}(t;\eta)-S(t;\eta)-\{\hat{S}(t;\eta^*)-S(t;\eta^*)\}\right|\right] \\
&\le(I)+(II)\le\tilde{c}_9\delta^{1/2}.
\end{aligned} \qquad (C.6)$$

Let $\phi_N(\delta)=\delta^{1/2}$ and $\alpha=\frac{3}{2}<2$, thus we have $\frac{\phi_n(\delta)}{\delta^\alpha}=\delta^{-1}$ is decreasing, and α does not depend on N. That is, the second condition holds.

Condition 3 By the facts that $\hat{\eta} \xrightarrow{p} \eta^*$ as $N \to \infty$, and that $\hat{S}(t;\hat{\eta}) \geq \sup_\eta \hat{S}(t;\eta)$, we choose $r_N = N^{1/3}$ such that $r_N^2 \phi_N(r_N^{-1}) = N^{2/3}\phi_N(N^{-1/3}) = N^{1/2}$. The third condition holds.

In the end, the three conditions are satisfied with $r_N = N^{1/3}$; thus we conclude that $N^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$, which completes the proof of (ii) in Theorem 9.2.4.

PART 3: We characterize the asymptotic distribution of $\hat{V}(\hat{\eta})$. Since we have

$$\sqrt{N}\{\hat{V}(\hat{\eta}) - V(\eta^*)\} = \sqrt{N}\{\hat{V}(\hat{\eta}) - \hat{V}(\eta^*)\} + \sqrt{N}\{\hat{V}(\eta^*) - V(t;\eta^*)\},$$

we study the two terms in two steps.

Step 3.1 To establish $\sqrt{N}\{\hat{V}(\hat{\eta}) - \hat{V}(\eta^*)\} = o_p(1)$, it suffices to show that $\sqrt{N}\{V(\hat{\eta}) - V(\eta^*)\} = o_p(1)$ and $\sqrt{N}(\hat{V}(\hat{\eta}) - \hat{V}(\eta^*) - \{V(\hat{\eta}) - V(\eta^*)\}) = o_p(1)$.

First, as $N^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$, we take the second-order Taylor expansion

$$\begin{aligned}
\sqrt{N}\{V(\hat{\eta}) - V(\eta^*)\} &= \sqrt{N}\left\{V'(\eta^*)\|\hat{\eta} - \eta^*\|_2 + \frac{1}{2}V''(\eta^*)\|\hat{\eta} - \eta^*\|_2^2 + o_p(\|\hat{\eta} - \eta^*\|_2^2)\right\} \\
&= \sqrt{N}\left\{\frac{1}{2}V''(\eta^*)\|\hat{\eta} - \eta^*\|_2^2 + o_p(\|\hat{\eta} - \eta^*\|_2^2)\right\} \\
&= \sqrt{N}\left\{\frac{1}{2}V''(\eta^*)O_p(N^{-2/3}) + o_p(N^{-2/3})\right\} = o_p(1).
\end{aligned}$$

Next, we follow the result (C.6) obtained in **PART 2**. As $N^{1/3}\|\hat{\eta} - \eta^*\|_2 = O_p(1)$, there exists $\tilde{\delta}_2 = c_{11}N^{-1/3}$, where $c_{11} < \infty$ is a finite constant, such that $\|\hat{\eta} - \eta^*\|_2 \leq \tilde{\delta}_2$. Therefore we have

$$\begin{aligned}
&\sqrt{N}(\hat{V}(\hat{\eta}) - \hat{V}(\eta^*) - \{V(\hat{\eta}) - V(\eta^*)\}) \\
&\leq \mathbb{E}\left[\sqrt{N}\sup_{\|\hat{\eta}-\eta^*\|_2 < \tilde{\delta}_2}\left|\hat{V}(\hat{\eta}) - V(\hat{\eta}) - \{\hat{V}(\eta^*) - V(\eta^*)\}\right|\right] \\
&\leq \tilde{c}_9\tilde{\delta}^{1/2} = \tilde{c}_9\sqrt{c_{11}}N^{-1/6} = o_p(1),
\end{aligned}$$

which yields the result.

Step 3.2 To derive the asymptotic distribution of $\sqrt{N}\{\hat{V}(\eta^*) - V(\eta^*)\}$, we follow the result obtained in **PART 1** that $\hat{V}(\eta^*) = V_N(\eta^*) + o_p(N^{-1/2})$, and thus

$$\sqrt{N}\left\{\hat{V}(\eta^*) - V(\eta^*)\right\} \xrightarrow{D} \mathcal{N}(0, \sigma_2^2),$$

where $\sigma_2^2 = \mathbb{E}[\phi_{d_{\eta^*}}^2]$ is the semiparametric efficiency bound.

Therefore we obtain in the end

$$\begin{aligned}
\sqrt{N}\{\hat{V}(\hat{\eta}) - v(\eta^*)\} &= \sqrt{N}\{\hat{V}(\hat{\eta}) - \hat{V}(\eta^*)\} + \sqrt{N}\{\hat{V}(\eta^*) - V(\eta^*)\} \\
&= o_p(1) + \sqrt{N}\{\hat{V}(\eta^*) - V(\eta^*)\} \\
&\xrightarrow{D} \mathcal{N}(0, \sigma_2^2),
\end{aligned}$$

which completes the proof of Theorem 9.2.4 and Corollary 9.2.6.

C.6 Proof of Theorem 9.2.7 and Theorem 9.2.8

When the source and target populations have the same distributions, both $\hat{V}_{DR}(\eta)$ and $\hat{V}_{CF}(\eta)$ converge to $V(\eta)$. The asymptotic variance of $\hat{V}_{DR}(\eta)$ is

$$\begin{aligned}\sigma^2_{DR} &= \mathbb{E}\left[\frac{I_S}{\mathbb{P}(I_S=1)}\left(\mu(d(X),X)+\frac{I\{A=d(X)\}}{\pi_d(X)}K(A,X)-V(\eta)\right)^2\right]\\ &= \mathbb{E}\left[\frac{I_S}{\mathbb{P}(I_S=1)}\left(\mu^2(d(X),X)+\frac{I\{A=d(X)\}}{\pi_d^2(X)}K^2(A,X)-V^2(\eta)\right.\right.\\ &\quad +\frac{2I\{A=d(X)\}}{\pi_d(X)}K(A,X)\mu(d(X),X)-2\mu(d(X),X)V(\eta)\\ &\quad \left.\left.-\frac{2I\{A=d(X)\}}{\pi_d(X)}K(A,X)V(\eta)\right)\right],\end{aligned}$$

while the asymptotic variance of $\hat{V}_{CF}(\eta)$ is

$$\begin{aligned}\sigma^2_{CF} &= \mathbb{E}\left[\left(I_T\, e(X)\mu(d(X),X)+\frac{I_S\, I\{A=d(X)\}}{\pi_S(X)\pi_d(X)}K(A,X)-V(\eta)\right)^2\right]\\ &= \mathbb{E}\left[\left(I_T\, e^2(X)\mu^2(d(X),X)+\frac{I_S\, I\{A=d(X)\}}{\pi_S^2(X)\pi_d^2(X)}K^2(A,X)-V^2(\eta)\right.\right.\\ &\quad \left.\left.-2I_T\, e^2(X)\mu(d(X),X)V(\eta)-\frac{2I_S\, I\{A=d(X)\}}{\pi_S(X)\pi_d(X)}K(A,X)V(\eta)\right)\right],\end{aligned}$$

where

$$K(A,X)=\frac{\Delta\, y(U)}{S_C(U\,|\,A,X)}-\mu(A,X)+\int_0^\infty \frac{\mathrm{d}M_C(u\,|\,A,X)}{S_C(u\,|\,A,X)}Q(u,A,X).$$

Since we have that

$$\mathbb{E}\left[\frac{I_S}{\mathbb{P}(I_S=1)}\frac{2I\{A=d(X)\}}{\pi_d(X)}K(A,X)\mu(d(X),X)\right]=0,$$

and for

$$B\in\left\{\mu^2(d(X),X),\frac{I\{A=d(X)\}}{\pi_d^2(X)}K^2(A,X),\mu(d(X),X)V(\eta),\frac{I\{A=d(X)\}}{\pi_d^2(X)}K(A,X)V(\eta)\right\},$$

we also have that

$$\mathbb{E}\left[\frac{I_S}{\mathbb{P}(I_S=1)}B\right]=\mathbb{E}[I_T\, e(X)B]=\mathbb{E}\left[\frac{I_S}{\pi_S(X)}B\right],$$

we conclude that $\sigma^2_{DR}=\sigma^2_{CF}$.

By the law of iterated expectations, the value function $V_d=\mathbb{E}[y(T(d))]=\mathbb{E}_X[\mathbb{E}[y(T(d))\,|\,X]]$. When there is no restriction on the class of ITRs, the true optimal ITR is

$$\begin{aligned}d^{**}(X)&=\arg\max_d V_d=\arg\max_d \mathbb{E}_X[\mathbb{E}[y(T(d))\,|\,X]]\\ &=I\{\mathbb{E}[y(T(1))\,|\,X]>\mathbb{E}[y(T(0))\,|\,X]\}.\end{aligned}$$

That is, the optimal ITR does not depend on the covariate distributions, but only the bilp function which is the same in both the source and target populations by Assumption 9.1.2. Thus both the maximizers of $\hat{V}_{DR}(\eta)$ and $\hat{V}_{CF}(\eta)$ converge to the true population parameter η^{**}. However, $\hat{V}_{DR}(\eta)$ is biased since the expectation $\mathbb{E}_X$ is taken with respect to the source population.

C.7 Additional simulations

We first investigate the performance of the cross-fitted ACW estimator with different sample sizes $(N, m) = (5\times10^4, 2000), (1\times10^5, 4000), (2\times10^5, 8000), (4\times10^5, 16000), (6\times10^5, 24000), (8\times10^5, 32000)$. Figure C.1 and Table C.1 report the results from 200 Monte Carlo replications. The variance is computed using the EIF.

C.8 Details of real data analysis

There are around 0.5% and 1.6% missing values in the RCT and OS data, respectively. We use the `mice` function in the `R` package `mice` [Van Buuren and Groothuis-Oudshoorn, 2011] to impute the missing values.

Motivated by the clinical practice and existing work in the medical literature, we consider ITRs that depend on the following five variables:

— AGE, SEX and Sequential Organ Failure Assessment (SOFA) score: these three baseline variables are well related to mortality in ICUs, so we consider them as important risk factors.

— Acute Kidney Injury Network (AKIN) score: Jaber et al. [2018] observed that the infusion of sodium bicarbonate improved survival outcomes and mortality rate in critically ill patients with severe metabolic acidemia and acute kidney injury. In the observational data, the AKIN score was not recorded, so we computed the score using serum creatinine measurement [Závada et al., 2010].

— SEPSIS: we consider the presence of sepsis as a risk factor because it is the main condition associated with severe acidemia at the arrival in ICU. The effect of sodium bicarbonate infusion on patients with acidema and acute kidney injury was also observed in septic patients [Zhang et al., 2018b].

Figure C.1 – *Boxplot of estimated value by ACW estimator with different sample sizes.*

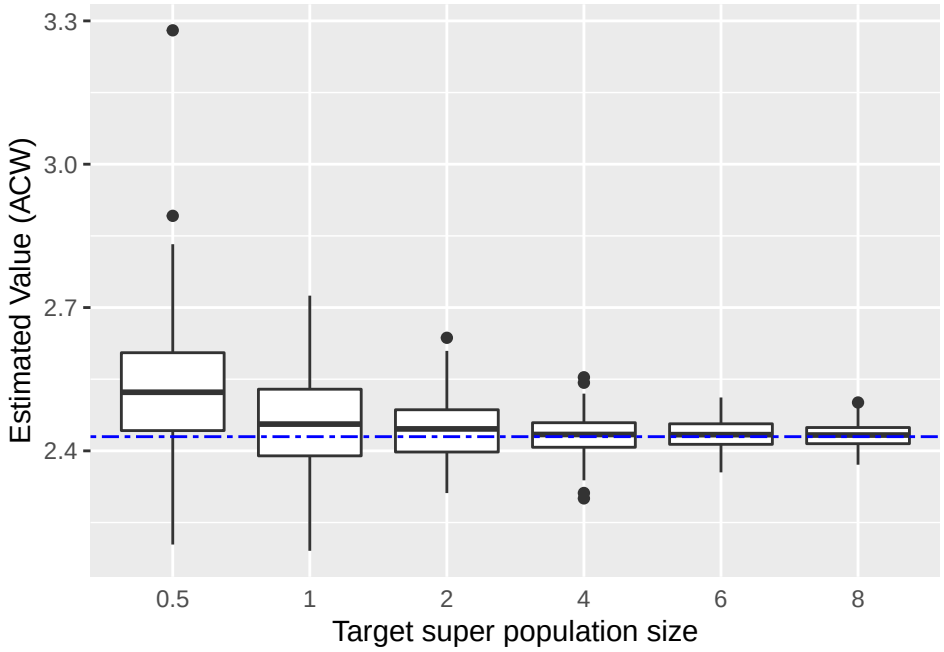


Table C.1 – *Numeric results of the ACW estimator. Bias is the empirical bias of point estimates; SD is the empirical standard deviation of point estimates; SE is the average of standard error estimates; CP is the empirical coverage probability of the 95% Wald confidence intervals.*

$n; m(\times 10^3)$	$\sim 780; 2$	$\sim 1560; 4$	$\sim 3120; 8$	$\sim 6240; 16$	$\sim 9360; 24$	$\sim 12480; 32$
Bias	0.1041	0.0253	0.0134	0.0046	0.0031	0.0030
SD	0.1394	0.0985	0.0635	0.0419	0.0317	0.0267
SE	0.1611	0.0942	0.0627	0.0417	0.0330	0.0284
CP(%)	97.5	93.5	96.0	94.5	97.5	97.0