Topics in Causal Inference and Privacy

Kormos, M.

DOI
10.4233/uuid:ca2b9c6e-b8fd-4dc5-b1d2-94959a583aae

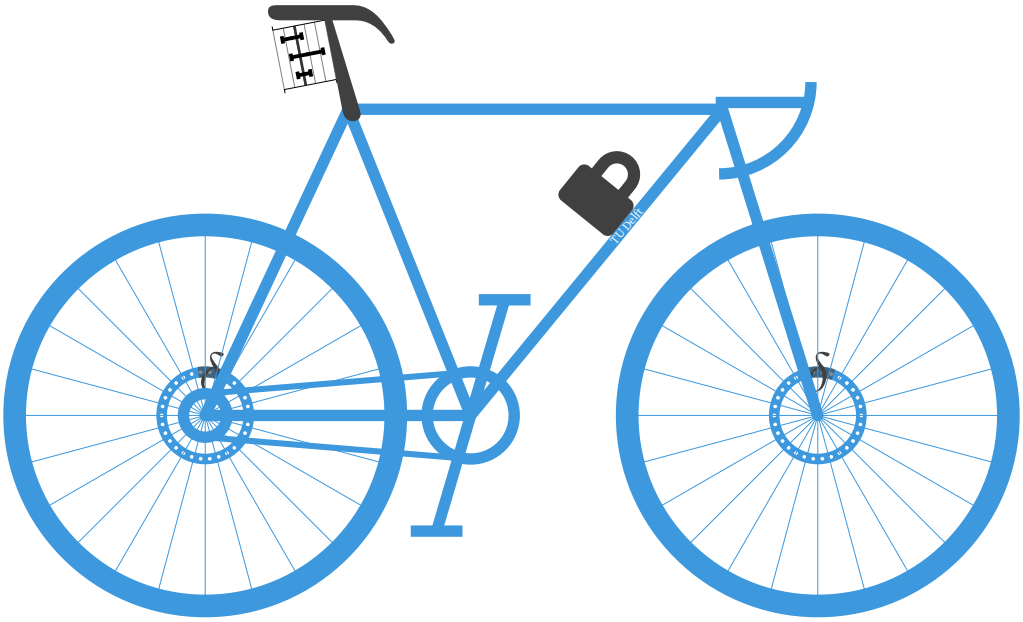
Publication date
2025

Document Version
Final published version

Citation (APA)
Kormos, M. (2025). *Topics in Causal Inference and Privacy*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:ca2b9c6e-b8fd-4dc5-b1d2-94959a583aae

Topics in Causal Inference and Privacy

Máté Kormos



Topics in Causal Inference and Privacy

Máté Kormos

Topics in Causal Inference and Privacy

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of Rector Magnificus
prof. dr. ir. T.H.J.J. van der Hagen,
Chair of the Board for Doctorates,
to be defended publicly on
Thursday 5 June 2025 at 10:00 o'clock

by

Máté KORMOS
Master of Science in Econometrics, Vrije Universiteit Amsterdam,
the Netherlands,
born in Budapest, Hungary

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. A. W. van der Vaart,	Delft University of Technology, promotor
Dr. S. L. van der Pas,	Vrije Universiteit Amsterdam, copromotor

Independent members:

Prof. dr. A. J. Cabo,	Delft University of Technology
Dr. N. van Geloven,	Leiden UMC
Prof. dr. ir. G. Jongbloed,	Delft University of Technology
Dr. ir. R. A. J. Post,	Erasmus MC
Prof. dr. M. van de Wiel,	Amsterdam UMC

Contents

Summary

Causal inference underlies many scientific inquires, aiming to infer the effect of some treatment, such as an educational program, on an outcome of interest, such as job satisfaction. When randomised experiments — arguably, the most credible methods for causal inference — are of limited availability, causal inference may be hindered by self-selection into the treatment. For instance, people participating in the educational program may be more motivated to begin with, and thus, more likely to be satisfied with their jobs. In this example, *motivation* is a so-called confounder and the failure to account for it amounts to what is called confounding bias.

Chapter 1 contributes to causal inference by deriving theoretical properties of a method for confounding-bias adjustment. Specifically, we study a causal effect estimator called *caliper matching*. Informally, an estimator is a function of an n-large sample of individuals in a population, aiming to recover the true population-level causal effect. Caliper matching achieves this aim by comparing individuals whose confounders are closer than a threshold called caliper. More precisely, instead of the confounders themselves, the comparison happens with respect to the conditional probability of participating in the treatment given the confounders (propensity score). A probability, the propensity score is a scalar, which makes comparison easier compared to the multi-dimensional confounders. The caliper determines how similar of individuals we compare. With a large caliper we compare dissimilar individuals; with a small caliper, we may have trouble finding anyone for comparison. We propose a caliper value balancing these two aspects in the asymptotic regime where the sample size n tends to infinity. We show that by comparing similar individuals under this suitable caliper value, caliper matching eliminates confounder bias, provided all conceivable confounders are observed. We also show that caliper matching is asymptotically normal: the uncertainty in the caliper

matching estimator's capacity for approximating the population-level causal effect is characterised by the normal distribution when n tends to infinity. We derive similar results for the case when the propensity score is unknown and needs to be estimated.

Chapter 2 contributes to causal inference by estimating the causal effects of two hip fracture treatments, using data from a randomised experiment and from a setting impaired by confounding bias (observational data). It is an application of a method to combine experimental and observational data. The combined data set clearly has a larger sample size, which, in turn, results in smaller variability (improved precision) of the estimator.

Privacy is a basic human need.[1] Curiosity, on the other hand, is a basic human trait. In Chapter 3, we aim to reconcile privacy and curiosity by devising inference procedures that preserve the privacy of individuals. The inference procedures consist in the construction of estimators and the description of their large-sample behaviour with a focus on asymptotic normality results. The privacy preservation consists in deliberately injecting noise into the data of individuals, and only releasing the noisy, privatised data for inference purposes. Inference remains feasible because the noise is injected in a controlled manner via what is called a local privacy mechanism, where 'local' is understood as individual-level. Indeed, one of our contributions is to devise a local privacy mechanism which enables the inference of *any* population-level quantity (called parameter; e.g. causal effect) of the non-noisy data from the noisy, privatised data. This is achieved by leaving the data intact with a small probability α and setting it to pure noise with probability $1 - \alpha$.

In particular, we restrict our attention to parameters which exhibit a so-called rate double robust property. It often happens that for the estimation of a parameter of interest θ_0, we need to estimate two other auxiliary (or nuisance) parameters. Then θ_0 is said to possess the rate double robust property if its asymptotic bias is characterised by the *product* of the estimation errors of the two nuisance parameters. This is desirable as a large error in one nuisance parameter can be offset by a small error in the other. Another one of our con-

[1] As is also argued for by a certain Czech–French novelist.

tributions is to extend the class of double robust parameters. These parameters include the causal effects of Chapters 1 and 2, but many other parameters of interest too.

Finally, we contribute to privacy-preserving inference by making a direct connection between privacy-agnostic and privacy-preserving estimators of (double robust) parameters. This connection enables the translation of privacy-agnostic estimators to the privacy-preserving setting.

Samenvatting[2]

Causale inferentie ligt ten grondslag aan veel wetenschappelijke onderzoeken, waarbij het doel is om het effect van een bepaalde behandeling, zoals een onderwijsprogramma, op een resultaat van belang, zoals werktevredenheid, af te leiden. Wanneer gerandomiseerde experimenten — de meest betrouwbare methoden voor causale inferentie — beperkt beschikbaar zijn, kan causale inferentie belemmerd worden door zelfselectie voor de behandeling. Zo kunnen mensen die deelnemen aan het onderwijsprogramma om te beginnen gemotiveerder zijn en dus meer kans hebben om tevreden te zijn met hun baan. In dit voorbeeld is motivatie een zogenaamde confounder en hier geen rekening mee houden komt neer op een zogenaamde confounding bias.

Chapter 1 draagt bij aan causale inferentie door theoretische eigenschappen af te leiden van een methode voor correctie van confounding-bias. Specifiek bestuderen we een causaal effect schatter, genaamd caliper matching. Een schatter is, informeel, een functie van een n-grote steekproef van individuen in een populatie, met als doel het ware causale effect op populatieniveau te bepalen. Caliper matching bereikt dit doel door individuen te vergelijken waarvan de confounders dichter bij elkaar liggen dan een drempel die caliper wordt genoemd. Preciezer gezegd, in plaats van de confounders zelf, gebeurt de vergelijking met betrekking tot de voorwaardelijke kans op deelname aan de behandeling (propensity score). Omdat de propensity score een waarschijnlijkheid is, is het een scalair, waardoor vergelijking eenvoudiger is in vergelijking met de multidimensionale confounders. De caliper bepaalt hoe vergelijkbaar individuen zijn die we vergelijken. Met een te grote caliper vergelijken we ongelijke individuen; met een te kleine caliper kunnen we moeite hebben om iemand te vinden om te vergelijken. We stellen een waarde voor

[2]Dank je wel, Jeffrey.

die deze twee aspecten in evenwicht brengt in het asymptotische regime waar de steekproefgrootte n naar oneindig gaat. We laten zien dat door vergelijkbare individuen te vergelijken onder deze geschikte waarde, caliper matching confounder bias elimineert, op voorwaarde dat alle denkbare confounders worden waargenomen. We laten ook zien dat caliper matching asymptotisch normaal is: de onzekerheid in het vermogen van de caliper matching schatter om het causale effect op populatieniveau te benaderen wordt gekarakteriseerd door de normale verdeling wanneer n naar oneindig gaat. We leiden vergelijkbare resultaten af voor het geval dat de propensity score onbekend is en geschat moet worden.

Chapter 2 draagt bij aan causale inferentie door de causale effecten van twee heupfractuurbehandelingen te schatten met behulp van gegevens uit gerandomiseerde experimenten en uit situaties waar sprake is van confounding bias (observationele gegevens). Het is een toepassing van een methode om experimentele en observationele gegevens te combineren. De gecombineerde dataset heeft duidelijk een grotere steekproefomvang, wat op zijn beurt resulteert in een kleinere variabiliteit (verbeterde precisie) van de schatter.

Privacy is een menselijke basisbehoefte.[3] Nieuwsgierigheid daarentegen is een basale menselijke eigenschap. In Chapter 3 proberen we privacy en nieuwsgierigheid met elkaar te verzoenen door inferentieprocedures te bedenken die de privacy van individuen beschermen. De inferentieprocedures bestaan uit de constructie van schatters en de beschrijving van hun gedrag bij grote steekproeven met een focus op asymptotische normaliteitsresultaten. De privacybescherming bestaat uit het opzettelijk injecteren van ruis in de gegevens van individuen en het alleen vrijgeven van de ruis bevattende, geprivatiseerde gegevens voor inferentiedoeleinden. Inferentie blijft haalbaar, omdat de ruis op een gecontroleerde manier wordt geïnjecteerd via wat een lokaal privacymechanisme wordt genoemd, waarbij 'lokaal' wordt opgevat als op individueel niveau. Een van onze bijdragen is het ontwikkelen van een lokaal privacymechanisme dat het mogelijk maakt om alle grootheden op populatieniveau (parameters genoemd; bijvoorbeeld causale effecten) van de gegevens zonder

[3] Zoals ook wordt betoogd door een zekere Tsjechisch–Franse romanschrijver.

ruis af te leiden uit de geprivatiseerde gegevens met ruis. Dit wordt bereikt door de gegevens intact te laten met een kleine waarschijnlijkheid van α en ze op pure ruis te zetten met een waarschijnlijkheid van $1 - \alpha$.

In het bijzonder beperken we onze aandacht tot parameters die een zogenaamde rate double robust eigenschap vertonen. Het komt vaak voor dat we voor de schatting van een parameter van belang θ_0 twee andere hulpparameters (of hinderparameters) moeten schatten. Dan heeft θ_0 de eigenschap rate double robust te zijn als de asymptotische bias wordt gekarakteriseerd door het *product* van de schattingsfouten van de twee hinderparameters. Dit is wenselijk, omdat een grote fout in de ene kan worden gecompenseerd door een kleine fout in de andere. Een andere bijdrage is het uitbreiden van de klasse van dubbel robuuste parameters. Deze parameters omvatten de causale effecten van Chapters 1 and 2, maar ook veel andere interessante parameters.

Tot slot dragen we bij aan privacy-beschermende inferentie door een direct verband te leggen tussen privacy-agnostische en privacy-beschermende schatters van (dubbel robuuste) parameters. Dit verband maakt de vertaling van privacy-agnostische schatters naar de privacy-beschermende omgeving mogelijk.

Acknowledgements

First of all, I would like to express my gratitude towards my supervisors Aad and Stéphanie. I am grateful for the freedom I was given, and indeed the trust I was given to handle this freedom. I am grateful for your help and for correcting me when I was on the wrong track. I am likewise grateful for all the conference opportunities I could live with, be that a *ragù* in Firenze or a *pastel de Belém* in Lisboa, which are at least as delicious as a *boterham* (or, according to some, even more so), a hike in the mountains close to Fréjus and in the beautiful forests around Lunteren, or a less beautiful covid in the otherwise rather pleasant city of Montréal. As Márai Sándor phrases it, essential things between people are never settled by means of words but by means of behaviour and acts. *Dank jullie zeer.*

I would like to thank the Bigstatistics Group of Amsterdam UMC for facilitating a regular change of environment for me in the form of in principle flexible but in practice arguably less so office space and of conversations and PhD activities.

A PhD thesis is, supposedly, an imprint of independent research. Nevertheless, the path leading up to the first letter of this thesis has been surrounded by influential and inspiring teachers: Reiff Ádám, Lieli Robert, Sergey Lychagin, Mátyás László, Charles Bos and Florian Wagener; thank you. Special thanks to Szarvas Beatrix, who, sadly, cannot read this anymore. *Köszönöm*, Olivér és Benedek.

It is with the greatest pleasure to acknowledge the importance of people who made my stay in the Netherlands a stay at home. Hong, thank you for everything. David, the only thing *sweeter* than a tiramisù — with or without Bailey's — being had in your company is your company itself. It is only fair to admit that notwithstanding our best efforts with Jeffrey, never have I witnessed a drunk performance as amusing as the sober one of yours. *Danke schön, Süßer.*

Dan, no acknowledgement can be complete without a recogniton of your impeccable sense of style of hats and keenness to always guide us onto the path towards *the* right restaurant. 谢谢. Dennis: subdued, I think you are also right sometimes... Jeffrey, everything that could have been written here was already said over countably infinite sessions in *Brouwhuis*, with consumption of questionable quantity in relation to the recommended intake for adults. I am positive that we can agree on the most *vanzelfsprekende* fact that the extent of your fine appreciation for *zuur bier* is only matched by our enthusiasm for the charm of certain Dutch cities. Your presence has been essential during my stay in Delft. *Hartelijk bedankt.*

−− .− −.−− −.−− −−− ..− .−. ..−. .−.. .. −−. .... −
. −−.− ..− .− .−..
−.−− −−− ..− .. −. −−. .−. .− −.−. .,
... .. −−. −. .. ..−. .. −.−. .− −. −.−. . .
.−− .... . .−. . ...− . .−. − .... . .−− .... .. −−...

Ádám, a true Hungarian *huszár*, my companion in the 'conquest' of the Netherlands. Rarely do I fail a course and rarely do friendships start in a bathroom. Yet, TI made both of them a reality — the former a true bore, the latter a true blessing. I dare say there is justice in life: the number of pizzas we had together well exceeds the number of days I spent studying macroeconomics (both years combined). Add to that the countless calls we have had since covid, keeping our spirits at least as high as the highest point in the Netherlands.[4] *Köszönöm.*

I would like to thank my friends in Hungary, whose presence was nearly not as remote as their physical distance from the Netherlands. That said, it has always been a pleasure meeting in person when I returned. Kata, I hope you find the thesis, or at least the cover, kitschy enough. Zsombi, luckily, the quality of Hungarian 'beers' has always been offset by the quality of our conversations. Anna, I could write here a less ironical sentence, but that would be rather unbecoming of you. Ágó, it can only be the faintest hope that this thesis lives up to the legacy of *föci fakt*. *Köszönöm nektek*.

[4] Vaalserberg towers at 322.4 metres (Wikipedia, 2024).

Hera, there is nothing quite like a lunch of, as someone at times of summit of their British consciousness would venture to exclaim, *I'm afraid*, inordinate portion at the St John's buttery to make semiparametrics almost exactly what may only be described as pure joy. Thank you very much for sharing your warm-heartedness with me, and for the time spent together in the land of left-inclined but (mostly) right-minded — at least as far as scones are concerned — people.

Finally, I would like to thank my family.

Köszönöm szépen a sok csillaghegyi finomságot, kezdve a tejfölszaftos fasírttól a bejglin át a háztetőig, amelyek nélkül ínségesebben telt volna ez a négy év.

Anyu, köszönöm szépen a sok kuglófot, amely a repülőutakat követően jelentősen feljavította a Hollandiában fellelhető ételek minőségét, illetve a rétesházi ebédeket, amelyek nem csak a finom ételek miatt emlékezetesek.

Papa, ebben a disszertációban éppúgy jelen van a sok tanulással együtt töltött óra, mint a sok sütemény. Köszönöm szépen.

Mama, álljon e köszönet hossza éles ellentétben a megköszönendő dolgok sokaságával. Köszönök szépen mindent, teljes szívemből.

Apu, elvitathatatlanok az érdemeid a jelen disszertáció utolsó betűjéig vezető út megvalósulásában. Hálás vagyok, és szeretettel köszönök mindent.

Zalán, remélem, ha egyszer el akarod, akkor el is fogod tudni olvasni ezt a disszertációt (és akkor már nem eszel ketchupot).

Kevés embert szólíthatok meg igazán a neve említése nélkül. Köszönöm.

Introduction

Suppose that we wish to infer the effect of a medical treatment on the well-being of individuals in a population. This research aim falls into the realm of causal inference, which is the main underlying topic connecting the three chapters of this thesis. Suppose also that the possibility of conducting an experiment — arguably, the most credible method for causal inference — is limited, e.g. for monetary or ethical reasons. In this case, we have to rely on observational data — data *not* from an experiment — to infer the effect of the treatment. This is the topic of the first and second chapter: the first chapter studies the case when experiments are completely unavailable, while the second chapter combines experimental and observational data. In addition, suppose that we also wish to protect the privacy of individuals. How to perform (causal) inference in a privacy-preserving manner is the content of the third chapter.

Chapter 1 and Chapter 3 contain novel theoretical results, while Chapter 2 is an application of existent methods to a concrete, medical research question. Each chapter is independent and self-contained. Accordingly, the rest of the Introduction is devoted to a high-level, nontechnical description of the content of the chapters; technical descriptions are in the corresponding chapters. Even so, we first need to be more precise about concepts regarding causality and inference.

One of the simplest ways to formalise the concept of *causality* is the potential outcome framework of Neyman (1924) and Rubin (1974). Consider our example, where we observe whether an individual takes the medical treatment ($D = 1$) or not ($D = 0$), and the well-being of the individual. We introduce the corresponding potential outcomes Y^1 and Y^0. The potential outcomes Y^1 and Y^0 are the well-being of the individual if they take and, respectively, do not take the treatment. The notion of a causal effect of the treatment is then captured by some contrast between Y^1 and Y^0, and the aim of causal inference

is to learn about this contrast. Perhaps the most commonly chosen contrast is $Y^1 - Y^0$, the difference in well-being when the individual takes the treatment, compared to when they do not take it. The average $\mathbb{E}\left[Y^1 - Y^0\right]$ of $Y^1 - Y^0$ across the individuals in the whole population under consideration is the average treatment effect; if the average is across individuals who took the treatment ($\mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D = 1\right]$), it is the average treatment effect on the treated. In the thesis, we adopt this potential outcome framework to describe causality, and our primary aim pertaining to causality is to infer average treatment effects of a treatment with two possible values $D = 0$ or $D = 1$.

By the *inference* of some unknown quantity θ_0 in a population, called the true parameter, we mean the construction of estimators and the description of their behaviour. For example, θ_0 may be $\mathbb{E}\left[Y^1 - Y^0\right]$ or $\mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D = 1\right]$. Informally, an estimator $\hat{\theta}_n$ of θ_0 is any function of the observed n data points, that is, an n-large sample drawn from the underlying population. In our well-being example, one may take the difference between the sample averages of well-being among people who took the medical treatment versus people who did not as an estimator of the average treatment effect. The behaviour of estimators we are interested in is their capacity for approaching the true parameter we wish to infer: the properties of $\hat{\theta}_n - \theta_0$. In particular, we are interested in their behaviour in the limit where the number of observed data points approaches infinity ($n \to \infty$). Intuitively, the more we observe of the world, the more accurate our estimators are ought to be in approaching the truth. We also quantify the uncertainty in the behaviour of $\hat{\theta}_n - \theta_0$, which arises from the fact that the sample is a *random* collection of individuals from the population. Specifically, we construct random intervals, called confidence intervals, which with high probability under repeated sampling, contain the true parameter θ_0.

We are now familiar with the minimal background needed to describe the content of all three chapters.

1. Chapter 1

Chapter 1 is concerned with causal inference when only observational and no experimental data are available. Take our medical treatment example, and for concreteness, let the treatment be a low-sugar diet. The diet is recommended

to all patients meeting some criteria, but, clearly, cannot be enforced. It may well happen that patients who decide to follow the diet lead a more health-conscious lifestyle to begin with, hence their well-being is, in general, better, relative to patients who decide not to follow the diet. Then simply comparing the well-being of treated versus nontreated patients leads to a bias. This is an example of what is called confounding bias: health consciousness confounds the casual relationship between the diet and well-being. To prevent confounding bias, we should compare patients with the same health consciousness.

More generally, to prevent confounding bias, we collect all conceivable confounders, referring to the thus collected variables as covariates, and compare the outcome of individuals with the same value of covariates. Even if we manage to collect all confounders, finding individuals with *exactly* the same value of covariates is, in general, infeasible, because the covariates may be continuous, e.g. average heart rate in the previous years. Therefore, we have to content ourselves with the comparison of individuals with *similar* covariate values. One approach to this is to compare an individual to all individuals whose covariate are closer to that of the individual than a threshold called caliper. This is called *caliper matching*, and is the content of Chapter 1. Therein, we study the properties of caliper matching, answering the question of how large the caliper should be, and deriving the limiting behaviour of caliper matching as the number of individuals in the observational sample tends to infinity. In turn, we use this to construct confidence intervals for the average treatment effect and the average treatment effect on the treated.

2. Chapter 2

Chapter 2 applies causal inference methods when experimental as well as observational data are available. Specifically, we compare the efficacy of two hip fracture treatments. To combine data from both experimental and observational sources is desirable to improve the quality of the comparison: by including more observations in the sample, by increasing n, the capacity of the estimator of the difference between the two treatments for approaching the true difference is enhanced. Such combination, however, should only happen if the

experimental and observational data share some similarities. In Chapter 2, we test combinability, only to find no evidence to the contrary. Hence, we combine the two data sources for a better comparison of the two treatments, and tackle some practical problems arising from missing data-measurements.

3. Chapter 3

Chapter 3 is dedicated to privacy-preserving inference. To preserve the privacy of individuals, noise is deliberately injected into their data. Then, instead of revealing the original data for inference, only the noisy version of their data is revealed, thus guaranteeing privacy. This 'noising process' is performed via what is called a local privacy mechanism. A local privacy mechanism injects noise into the data of every single individual — hence the name 'local' — in a controlled manner. It is because of this controlled noising, that inference remains feasible. Injecting too little noise would mean no privacy, while injecting too much noise would preclude inference. In Chapter 3, we propose a privacy mechanism balancing these two extremes, which lends the noisy, privatised data for inference.

In particular, we focus on the inference of parameters θ_0 which exhibit a so-termed rate double robust property. It often happens that for the estimation of θ_0, we need to estimate other parameters which are not of primary interest; these are called auxiliary or nuisance parameters. A parameter θ_0 exhibits a rate double robust property if the large-sample bias $\hat{\theta}_0 - \theta_0$ of an estimator $\hat{\theta}_n$ of θ_0 is characterised by the *product* of the estimation errors of two nuisance parameters. This property is considered desirable, because a large error in the estimation of one nuisance parameter can be offset by a smaller estimation error of the other, due to the product structure. In Chapter 3, we extend the class of parameters that were previously known to have the rate double robust property. This includes the average treatment effects in Chapters 1 and 2, but reaches well beyond that. To infer rate double robust parameters, we consider their estimation from samples which are privatised by a suitable local privacy mechanism. In this endeavour, we make a connection between nonprivate and private estimation theory.

4. Summary

Most briefly, our contributions in each chapter may be summarised as follows.

- Chapter 1 (theory): the derivation of the large-sample properties of the caliper matching estimator of average treatment effects.
- Chapter 2 (application): the comparison of two hip fracture treatments by combining experimental and observational data sources.
- Chapter 3 (theory): the extension of rate double robust parameter classes and the development of their privacy-preserving inference.

Finally, the Conclusion reviews the three chapters, highlighting their social and practical relevance.

Chapter 1

Caliper Matching

Abstract

Caliper matching is used to estimate causal effects of a binary treatment from observational data by comparing matched treated and control units. Units are matched when their propensity scores, the conditional probability of receiving treatment given pretreatment covariates, are within a certain distance called caliper. So far, theoretical results on caliper matching are lacking, leaving practitioners with ad-hoc caliper choices and inference procedures. We bridge this gap by proposing a caliper that balances the quality and the number of matches. We prove that the resulting estimator of the average treatment effect, and average treatment effect on the treated, is asymptotically unbiased and normal at parametric rate. We describe the conditions under which semiparametric efficiency is obtainable, and show that when the parametric propensity score is estimated, the variance is increased for both estimands. Finally, we construct asymptotic confidence intervals for the two estimands.

1.1. Introduction

Matching is applied in empirical studies to estimate the causal effect of a binary treatment from observational data. The estimate is the mean difference in the outcome of interest of matched treated and control units. Matches may be formed in various ways. We consider matching on the propensity score, the conditional probability of receiving treatment given the observed pretreatment covariates (Rosenbaum and Rubin, 1983). Specifically, we consider caliper matching, where a treated and a control unit are matched if their propensity scores are within a certain distance called *caliper* (Cochran and Rubin, 1973; Dehejia and Wahba, 1998). Caliper matching is applied in empirical research such as labour (Dehejia and Wahba, 2002; Huber et al., 2015b) and health economics (Erhardt, 2017; Salmasi and Pieroni, 2015; Keng and Sheu, 2013), policy evaluation (Bannor et al., 2020; Patel-Campillo and García, 2022), business and finance (Shen and Chang, 2009; Heese et al., 2017) as well as healthcare (Capogrossi and You, 2017; Cho, 2018; Vecchio et al., 2018; Izudi et al., 2019; Wang et al., 2020; Brenna, 2021; Krishnamoorthy and Rehman, 2022). Nonetheless, no rigorous results have been established on the choice of the caliper and the limiting distribution of the estimator.

Our contribution is a theory driven caliper choice, the derivation of the asymptotic distribution of the caliper matching estimator based on propensity scores, and the construction of asymptotic confidence intervals. We consider the estimation of the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT). We show that when the order of the caliper decreases at the right speed as the sample size n increases, the estimators of both estimands are asymptotically unbiased and normal at $\sqrt{n}$-rate, even when the parametric propensity score is estimated. In the rest of this section, we situate our contribution in the literature.

Matching has attracted much attention in the literature, with the idea of comparing similar units dating back to at least Densen et al. (1952) (Cochran, 1953). Cochran and Rubin (1973) review then-available matching methods applicable to observational studies. The reader is referred to Rubin (2006) for a collection of historical results and to Stuart (2010) for a comprehensive sur-

vey. Abadie and Imbens (2006) present a key result closely related to ours. They study nearest neighbor matching, where the M closest units in terms of covariates are matched to a given unit. They show that nearest neighbor matching on covariates is asymptotically normal, but unbiased only when we match on a scalar variable, such as the propensity score. Providing the identification results for unbiasedness, the foundations of propensity score matching is laid down by Rosenbaum and Rubin (1983). Abadie and Imbens (2016) derive some asymptotic properties of nearest neighbor matching on the estimated parametric propensity score. They discretise the maximum likelihood estimator of the propensity score parameter and show that the resulting matching estimator converges to a normal distribution as, first, the sample size increases and, *then*, the discretisation gets finer. Since their approach changes the estimator, this asymptotic result is not equivalent to the asymptotic normality of nearest neighbor matching on the estimated parametric propensity score. In contrast, we do not change the estimator, nor do we appeal to discretisation arguments and double limits. Employing sample-splitting to estimate the propensity score, we establish the asymptotic normality of caliper matching on the estimated parametric propensity score as the sample size increases. Consequently, we are able to construct confidence intervals for ATE and ATT, centred at the caliper matching estimator based on the estimated propensity scores, which get more reliable as the sample size increases.

The first mention of caliper matching appears to be in Cochran and Rubin (1973). Therein, it is analysed for a few specific models and is compared with other matching methods, such as nearest neighbor. The caliper is chosen based on the variances of the outcome in the treatment and control group. Rosenbaum and Rubin (1985) seem to be the first to consider caliper matching involving the propensity score as well as the covariates. They assume a logistic model for the propensity score, and match on the logit of the propensity score, that is, a linear function of the covariates. They choose the caliper based on the variances of the logit in the treatment and control group. The caliper choices of Cochran and Rubin (1973) and Rosenbaum and Rubin (1985) may lead to a large enough number of matches to reduce the variance of the caliper matching estimator. However, they do not make the bias of the caliper matching estima-

tor converge to zero — unless the caliper is used in combination with nearest neighbor matching; see next paragraph —, for that the caliper needs to shrink with the sample size as we show in our present work.

Some authors, including Rosenbaum and Rubin (1985), use the term caliper matching to refer to nearest neighbor matching with a caliper restriction: the M nearest units are to be matched, but only if they are within the caliper. Others, for instance Dehejia and Wahba (1998), use the term to mean that all units within the caliper are matched, even though they may be differently weighted.[1] We adopt the latter approach with uniform weights, sometimes also called radius matching (Huber et al., 2015a), because of its simplicity. Caliper matching can then be regarded as a kernel matching method with rectangular kernel and the bandwidth equal to the caliper. As such, the seminal work of Heckman et al. (1998), establishing the asymptotic normality of the kernel matching estimator of ATT even for nonparametrically estimated propensity score — with bandwidth choice further investigated by Frölich (2005) —, is closely related to our work. However, their results do not apply to caliper matching because they require the kernel to be Lipschitz continuous. The rectangular kernel fails to be so, prohibiting the asymptotic linear expansion of the kernel matching estimator, which is key to their argument. The work of Lee (2018) is similar in spirit. It extends Heckman et al. (1998) to a richer set of estimands beyond average effects using kernel matching methods, but also assuming a smooth kernel, excluding the rectangular one of caliper matching.

We overcome the nonsmoothness of the rectangular kernel by employing empirical process theory in Alexander (1987) and Van der Vaart and Wellner (1996). Writing the number of matches in terms of empirical measures enables us to characterise the asymptotic behaviour of caliper matching using ratio and tail bounds for empirical measures and processes. Furthermore, we can establish the efficiency properties of caliper matching. More efficient estimators have smaller variance and thus yield narrower confidence intervals.

[1] The two interpretations coincide when M is taken to be, for example, n in the caliper restriction case. As M is usually set to a constant independent of n, it is reasonable to distinguish the two interpretations.

The efficiency of caliper matching depends on the estimand, the observed sample, the regression of the outcome on the covariates, and the knowledge of the propensity score.

First, we consider the case when the propensity score is known. We prove that if we only observe the propensity scores in our sample but not the covariates, or the regression of the outcome on the covariates only depends on the covariates through the propensity score, then the limiting variance of the caliper matching estimator of (i) ATE reaches the semiparametric lower bound; (ii) ATT reaches the semiparametric lower bound for *unknown* propensity score (Hahn, 1998). The latter is not the best possible result as the lower bound for ATT, unlike ATE, is smaller when the propensity score is known (Hahn, 1998). Yet, we show that caliper matching is more efficient than nearest neighbor matching on the propensity scores studied by Abadie and Imbens (2006, 2016), yielding narrower confidence intervals for ATE as well as ATT — regardless of whether we observe the covariates in the sample or whether the outcome regression depends on the covariates or the propensity scores.

Second, if the propensity score is unknown, but we assume and estimate a parametric specification such as the logit or probit model, then the limiting variance of the caliper matching estimator of both estimands is in general larger compared to when the propensity score is known. Consequently, it remains unclear whether the caliper or the nearest neighbor matching (Abadie and Imbens, 2016) on the estimated propensity scores is more efficient.

Our assumptions include the usual common support for the propensity score, and smoothness conditions for the conditional moments of the outcome and for (the density of) the propensity score. We verify our assumptions for a logit or probit model for the propensity score and for smooth, potentially nonlinear and heteroskedastic, regression of the outcome on the covariates with a well-behaved density on a compact support.

The rest of the chapter is organised as follows. In Section 1.2, we introduce the conceptual framework and the caliper matching estimator. Section 1.3 contains our contributions, the caliper choice and the asymptotic properties of the estimator. Section 1.4 concludes.

1.2. Preliminaries

1.2.1. Framework

We adopt the potential outcome framework of Neyman (1924) and Rubin (1974) with no interference between the units (stable unit-treatment value assumption, Rosenbaum and Rubin (1983)). Let D be the treatment indicator with value one corresponding to treatment and zero to control. The real-valued Y^1, Y^0 are the potential outcomes under treatment and control, respectively. We observe exactly one of Y^1 and Y^0, depending on D, so that the observed outcome is $Y = DY^1 + (1-D)Y^0$. The estimands of interest, ATE and ATT, are defined respectively as

$$\tau \coloneqq \mathbb{E}\left[Y^1 - Y^0\right], \quad \tau_{\mathrm{t}} \coloneqq \mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D = 1\right].$$

To identify ATE and ATT from observational data, we assume that the observed pretreatment covariates X, taking values in $\mathcal{X} \subset \mathbb{R}^K$, account for all the systematic differences between treated and control units. Formally, the potential outcomes are assumed to be independent of the treatment participation given the covariates, which is a standard assumption of causal inference (Rubin, 1974).

Assumption 1.1 (Unconfoundedness). *$Y^0 \perp\!\!\!\perp D \mid X$ and $Y^1 \perp\!\!\!\perp D \mid X$.*

Let $\pi(x) \coloneqq \mathbb{P}(D = 1 \mid X = x)$ be the propensity score with conditional distribution function $F_d(p) \coloneqq \mathbb{P}(\pi(X) \leq p \mid D = d)$. The F_d are assumed to satisfy Assumption 1.2.

Assumption 1.2 (Propensity Score Distribution). *(i) F_0, F_1 admit densities f_0, f_1, respectively.*

(ii) f_0, f_1 have the same compact support $[\underline{p}, \bar{p}]$, $0 < \underline{p} < \bar{p} < 1$.

(iii) f_0, f_1 are bounded away from zero on their support.

(iv) f_0, f_1 are continuous on their support.

Assumption 1.2 imposes the same requirements on the propensity score distribution as Abadie and Imbens (2016), except that it also requires the densities

f_0, f_1 to be strictly positive. This requirement ensures that the quantile functions F_d^{-1} have bounded derivatives, which we use for the caliper choice. It also plays a role in the proof of the asymptotic normality by ensuring that ratio bounds for empirical processes apply.[2]

Assumption 1.2 implies that if there is a unit with propensity score in some region of $[0, 1]$, then there is a positive probability of finding a unit from the opposite treatment group therein. This ensures that treated and control units can be compared in terms of their propensity scores. In combination with Assumption 1.1, this yields the identification of the estimands from observed variables, by comparing treated and control units with the same propensity scores (Rosenbaum and Rubin, 1983):

$$\tau = \mathbb{E}\left[\mathbb{E}\left[Y \mid D = 1, \pi(X)\right] - \mathbb{E}\left[Y \mid D = 0, \pi(X)\right]\right], \tag{1.1}$$

$$\tau_{\mathrm{t}} = \mathbb{E}\left[\mathbb{E}\left[Y \mid D = 1, \pi(X)\right] - \mathbb{E}\left[Y \mid D = 0, \pi(X)\right] \mid D = 1\right]. \tag{1.2}$$

1.2.2. Caliper Matching Estimator

We wish to construct estimators based on identification formulae (1.1) and (1.2) from an independently and identically distributed (i.i.d.) sample from the distribution of (Y, D, X), denoted by $((Y_i, D_i, X_i))_{i\in[n]}$, with $[n] \coloneqq \{1, 2, \ldots, n\}$. This would necessitate finding sample units with the same value of the propensity score, which is infeasible for continuously distributed propensity scores. Rather, matching estimators look for units with *similar* propensity scores. The caliper matching estimator explicitly controls the extent of similarity with the caliper δ, whose choice is discussed later on in Section 1.3.

Suppose for now that the propensity score is known. Given $\delta > 0$, the caliper matching estimator constructs the match set $\mathcal{J}(i) \coloneqq \{j \in [n] : D_j \neq D_i, |\pi(X_j) - \pi(X_i)| \leq \delta\}$ of unit $i \in [n]$. Next, it estimates the missing potential outcome of the unit with the mean outcome of units in the match set. Averaging out the difference between the (estimated) potential outcomes then gives the estimate of the causal effect. Let $M_i \coloneqq |\mathcal{J}(i)|$ be the number of matches of

[2] The strict positivity of f_d implies that $\inf_{p\in[\underline{p},\bar{p}]} \int_{p-\delta}^{p+\delta} f_d(\tilde{p})\mathrm{d}\tilde{p} \gtrsim \delta > 0$, so that the denominator in the ratios of empirical to true measures is bounded away from zero, keeping the ratios finite.

unit $i \in [n]$, and write $N_0 \coloneqq \sum_{i\in[n]}(1-D_i)$, $N_1 \coloneqq \sum_{i\in[n]} D_i$ for the number of control and treated units, respectively. The estimators of ATE and ATT are defined respectively as

$$\begin{aligned}\hat{\tau}_\pi \coloneqq \frac{1}{n}\sum_{i\in[n]}\Bigg[& D_i\left(Y_i - \frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j\right) \\ & + (1-D_i)\left(\frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j - Y_i\right)\Bigg] \mathbb{1}_{M_i>0}, \\ \hat{\tau}_{\mathrm{t},\pi} \coloneqq \frac{1}{N_1}\sum_{i\in[n]} & D_i\left(Y_i - \frac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j\right)\mathbb{1}_{M_i>0}.\end{aligned}$$

The indicator $\mathbb{1}_{M_i>0}$, being one if unit i has matches and zero if not, ensures that only units that have matches are included in the estimate.

In practice, the propensity score is usually unknown. Often, it is assumed to follow a smooth parametric model, such as logit or probit. Following Abadie and Imbens (2016), we also make this assumption.

Assumption 1.3 (Smooth Parametric Propensity Score)**.** *(i) The propensity score is* $\mathbb{P}(D=1 \mid X) = \pi(X,\theta_0)$ *for a parametric model* $\{\pi(\cdot,\theta) : \theta \in \Theta \subset \mathbb{R}^K\}$ *with* θ_0 *in the interior of* Θ*.*

(ii) $\theta \mapsto \pi(x,\theta)$ *is differentiable in the neighbourhood of* θ_0 *for all* $x \in \mathcal{X}$*.*

(iii) The derivative in Assumption 1.3(ii) is bounded uniformly in $x \in \mathcal{X}$ *in the neighbourhood of* θ_0*.*

The caliper matching estimator is then defined by a plug-in rule. Let $\mathcal{J}_\theta(i) \coloneqq \{j \in [n] : D_j \neq D_i, |\pi(X_j,\theta) - \pi(X_i,\theta)| \leq \delta\}$ be the match set and $M_i(\theta) \coloneqq |\mathcal{J}_\theta(i)|$ its cardinality for some $\theta \in \Theta$. For an estimator $\hat{\theta}$ of θ_0, the matching

estimators of ATE and ATT are, respectively,

$$\begin{aligned}\hat{\tau}_{\hat{\pi}} :=& \frac{1}{n} \sum_{i\in[n]} \left[D_i \left(Y_i - \frac{1}{M_i(\hat{\theta})} \sum_{j\in\mathcal{J}_{\hat{\theta}}(i)} Y_j \right) \right. \\ & \left. + (1 - D_i) \left(\frac{1}{M_i(\hat{\theta})} \sum_{j\in\mathcal{J}_{\hat{\theta}}(i)} Y_j - Y_i \right) \right] \mathbb{1}_{M_i(\hat{\theta})>0}, \\ \hat{\tau}_{\mathfrak{t},\hat{\pi}} :=& \frac{1}{N_1} \sum_{i\in[n]} D_i \left(Y_i - \frac{1}{M_i(\hat{\theta})} \sum_{j\in\mathcal{J}_{\hat{\theta}}(i)} Y_j \right) \mathbb{1}_{M_i(\hat{\theta})>0}.\end{aligned}$$

1.3. Asymptotics

In this section, we state our main results: the caliper choice (Section 1.3.1), the asymptotic normality of the caliper matching estimators of ATE and ATT for known (Section 1.3.2) and estimated (Section 1.3.2) propensity scores, and the variance estimation (Section 1.3.4).

1.3.1. Caliper Choice

A smaller caliper means that the propensity scores of matched treated and control units are closer, so the match quality is better. At the same time, a smaller caliper leads to fewer matches. Hence, the caliper controls directly the quality and, indirectly, the number of matches, which, in turn, govern the properties of the matching estimator. The match quality determines the bias: comparing dissimilar units threatens the identification of estimands in (1.1) and (1.2). The number of matches determines the bias — by excluding units with no matches — as well as the variance of the estimator: since the estimator involves averages over the match set, a small match set gives large variance.

Thus, the right caliper choice must balance the quality and the number of matches. As the sample size increases, we expect that under Assumption 1.2, we can find both treated and control units in every region of $[\underline{p}, \bar{p}]$ with increasing probability. It is then reasonable to aim for finding matches for each unit

in the large sample limit. If we were to set the caliper to

$$\overline{\Delta}_n := \max_{i\in[n]} \min_{j\in[n]:D_j\neq D_i} |\pi(X_i) - \pi(X_j)|,$$

the largest closest distance between treated and control units, we would have at least one match for each unit. The order of $\mathbb{E}\overline{\Delta}_n$ can be concisely described in terms of the sample size, relying on the results of Shorack and Wellner (2009) on spacings (all proofs are presented in Sections 1.B and 1.C).

Proposition 1.1 (Order of Expected Largest Closest Distance)**.** *Under Assumption 1.2, there exist constants* $0 < n_0, c < \infty$ *such that* $\mathbb{E}\overline{\Delta}_n \leq c\frac{\log n}{n}$ *for all* $n \geq n_0$.

This suggests that the caliper choices, for $n \geq 2$,

$$\delta := \delta_n := s\frac{\log n}{n} \quad \text{or} \quad \delta := \delta_n := \underline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1} \tag{1.3}$$

for any constant $0 < s \leq 2/\log(2)$ (so that $0 < \delta_n \leq 1$ for $n \geq 2$) are asymptotically of the same order and large enough to guarantee matches for each unit, although the data-dependent choice $\delta_n = \underline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$ can better accommodate smaller samples thus it is generally preferred. Indeed, Proposition 1.2 shows that, in fact, the implied number of matches is of the order $\log n$.

Proposition 1.2 (Number of Matches)**.** *Let the caliper satisfy* (1.3)*. If Assumption 1.2 holds, then there exist constants* $0 < c_l, c_u < \infty$ *such that*

$$c_l(1 + o_P(1)) \log n \leq \min_{i\in[n]} M_i \leq \max_{i\in[n]} M_i \leq c_u(1 + o_P(1)) \log n$$

as $n \to \infty$*. Thus,* $\mathbb{P}\left(\min_{i\in[n]} M_i \geq 1\right) \to 1$ *as* $n \to \infty$.

1.3.2. Known Propensity Score

Assume for now that the propensity score $x \mapsto \pi(x)$ is known. We derive the asymptotic distribution of caliper matching in this setting, and show that the caliper choice (1.3) not only leads to a number of matches increasing in the sample size, but also to the asymptotic unbiasedness of the matching estimator.

In the following, we make a series of assumptions amounting to the asymptotic normality of caliper matching, and we prove that, for instance, the models of Example 1.1 satisfy these assumptions. Popular models, including the

logit and probit for the propensity score and smooth heteroskedastic outcome regressions, are all covered by Example 1.1 as long as the covariates admit a well-behaved density.[3] The condition of having $K \geq 2$ continuously distributed covariates with nonzero propensity score parameters is not restrictive; for if we had only one, then matching on the propensity score and matching on the covariate would be akin.[4] Regarding other conditions of Example 1.1, $\nu_d \perp\!\!\!\perp D \mid X$ implies Assumption 1.1, while differentiability of $x \mapsto \mathbb{E}\left[\nu_d^2 \,\middle|\, X = x\right]$ allows for smooth heteroskedastic models.

Example 1.1 (Admissible Models)**.** *Let $g : \mathbb{R} \to [0,1]$ be a strictly increasing function that is twice continuously differentiable on $\mathbb{R}$, with first derivative g' satisfying $\sup_{t\in\mathbb{R}} g'(t) < \infty$. The $(K \geq 2)$-dimensional covariates have density Ψ, which is bounded away from zero on the compact support $\mathcal{X}$ and continuously differentiable. The propensity score and the potential outcomes satisfy*

$$\begin{aligned} \pi(x) &= g(\theta_0^\intercal x) \\ Y^d &= m_d(X) + \nu_d, \quad \mathbb{E}\left[\nu_d \,\middle|\, X\right] = 0, \quad d \in \{0,1\}, \end{aligned}$$

where θ_0 is in the interior of $\Theta \subset \mathbb{R}^K$, and it has at least two nonzero coordinates, Θ is bounded, and the m_d are continuously differentiable. For all $d \in \{0,1\}$, $\nu_d \perp\!\!\!\perp D \mid X$, the $x \mapsto \mathbb{E}\left[\nu_d^r \,\middle|\, X = x\right]$, $r \in \{2,4\}$, are continuously differentiable on $\mathcal{X}$, and $\inf_{x\in\mathcal{X}} \mathbb{E}\left[\nu_d^2 \,\middle|\, X = x\right] > 0$.

We can rewrite $\hat{\tau}_\pi, \hat{\tau}_{\mathrm{t},\pi}$ as weighted averages of the outcome variable Y as follows:

$$\begin{aligned} \hat{\tau}_\pi &= \frac{1}{n} \sum_{i\in[n]} (2D_i - 1)(\mathbb{1}_{M_i>0} + w_i) Y_i, \\ \hat{\tau}_{\mathrm{t},\pi} &= \frac{1}{N_1} \sum_{i\in[n]} (\mathbb{1}_{M_i>0} D_i - (1 - D_i) w_i) Y_i, \quad w_i \coloneqq \sum_{j\in\mathcal{J}(i)} \frac{1}{M_j}, \end{aligned}$$

[3] For simplicity of exposition, we assume throughout the chapter that X does not include an intercept. The intercept can be accommodated by redefining the distributional assumptions on X to refer to the nonintercept coordinates of X.

[4] Replacing the propensity score with the scalar covariate in Assumptions 1.2, 1.4 and 1.5 would yield a version of Propositions 1.1 and 1.2 and Theorems 1.1 and 1.2 with the propensity score replaced with the covariate.

where $M_j = 0$ only if $\mathcal{J}(i)$ is empty, in which case the sum in w_i is taken to be zero.[5]

Let $\mu^d(p) \coloneqq \mathbb{E}[Y \mid D = d, \pi(X) = p]$ be the regression function and $\varepsilon \coloneqq Y - \mu^D(\pi(X))$ be the corresponding disturbance term with conditional variance

$$\sigma_d^2(p) \coloneqq \mathbb{V}[\varepsilon \mid D = d, \pi(X) = p] = \mathbb{V}[Y \mid D = d, \pi(X) = p], \quad d \in \{0, 1\}.$$

When we apply caliper matching to imitate (1.1) and (1.2), we make two approximations. First, we compare the outcome Y, rather than the regression $\mu^D(\pi(X))$, of the units. The error we make in doing so is ε. Second, we compare units with similar, rather than the same, propensity scores. Therefore, some assumptions must be imposed on the magnitude of ε and the smoothness of μ^d. The magnitude of ε cannot be too large, but also, for convenience, not too small either to avoid degenerate limits. Assumptions 1.4 and 1.5 are the same as Assumption 4 in Abadie and Imbens (2006), adapted to matching on the propensity score $\pi(X)$, rather than on the covariates X.

Assumption 1.4 (Disturbance Term). *(i)* $\inf_{d\in\{0,1\}, p\in[\underline{p},\bar{p}]} \sigma_d^2(p) > 0$ *and* $\sup_{d\in\{0,1\}, p\in[\underline{p},\bar{p}]} \sigma_d^2(p) < \infty$.

(ii) $\sup_{d\in\{0,1\}, p\in[\underline{p},\bar{p}]} \mathbb{E}\left[\varepsilon^4 \mid D = d, \pi(X) = p\right] < \infty$.

Assumption 1.5 (Lipschitz Regression Functions). *The μ^d are Lipschitz continuous: there exists a constant $0 < L_\mu < \infty$ such that $|\mu^d(p) - \mu^d(p')| \leq L_\mu |p - p'|$ for all $p, p' \in [\underline{p}, \bar{p}]$ for all $d \in \{0, 1\}$.*

Lipschitz continuity guarantees that when the propensity scores $\pi(X_i)$ and $\pi(X_j)$ are close, which we control with δ_n, then so are $\mu^d(\pi(X_i))$ and $\mu^d(\pi(X_j))$. This is in agreement with identification formulae (1.1) and (1.2), leading to asymptotic unbiasedness. Similarly to Abadie and Imbens (2006), write the ATE estimator as

$$\hat{\tau}_\pi = \overline{\tau(\pi(X))} + E + B, \tag{1.4}$$

[5] This follows from the symmetry of caliper matching: $j \in \mathcal{J}(i)$ if and only if $i \in \mathcal{J}(j)$.

where we defined the terms

$$\overline{\tau(\pi(X))} \coloneqq \frac{1}{n}\sum_{i\in[n]} \tau(\pi(X_i)), \quad \tau(\pi(X_i)) \coloneqq \mu^1(\pi(X_i)) - \mu^0(\pi(X_i)), \tag{1.5}$$

$$E \coloneqq \frac{1}{n}\sum_{i\in[n]} E_i, \quad E_i \coloneqq (2D_i - 1)(\mathbb{1}_{M_i>0} + w_i)\varepsilon_i, \tag{1.6}$$

$$B \coloneqq \frac{1}{n}\sum_{i\in[n]} B_i, \tag{1.7}$$

$$\begin{aligned} B_i \coloneqq{}& (2D_i - 1)\frac{\mathbb{1}_{M_i>0}}{M_i}\sum_{j\in\mathcal{J}(i)} (\mu^{1-D_i}(\pi(X_i)) - \mu^{1-D_i}(\pi(X_j))) \\ &+ (2D_i - 1)(\mathbb{1}_{M_i>0} - 1)(\mu^{1-D_i}(\pi(X_i)) - \mu^{D_i}(\pi(X_i))). \end{aligned} \tag{1.8}$$

The first term $\overline{\tau(\pi(X))}$ has mean τ and the second term E has mean zero. After centering at τ, the first two terms shall be shown to be asymptotically jointly normal and independent at $\sqrt{n}$-rate. The third term B has two sources of bias. The first term in (1.8) is the bias stemming from imperfect matches. If matches were exact, this term would be zero. By Assumption 1.5, the magnitude of this term is δ_n, hence it tends to zero even when multiplied with $\sqrt{n}$. The second term in (1.8) is due to discarding unmatched units, which may happen for the caliper choice $\delta_n = s\frac{\log n}{n}$, unlike for the data-dependent choice $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$. This leads to a bias because we introduce an artificial sample selection based on δ_n. If every unit had at least one match, as is the case for the data-dependent caliper choice, this term would be zero. But, as shown in Proposition 1.2, this happens in the large sample limit, giving the asymptotic normality of the ATE estimator $\hat{\tau}_\pi$.

Theorem 1.1 (Asymptotic Normality for Known Propensity Score (ATE))**.** *Suppose that $x \mapsto \pi(x)$ is known and the caliper δ_n satisfies (1.3). If Assumptions 1.1, 1.2, 1.4 and 1.5 all hold, then*

$$\sqrt{n}(\hat{\tau}_\pi - \tau) \rightsquigarrow \mathcal{N}(0, V) \quad \text{as } n \to \infty,$$

where $V \coloneqq V_\tau + V_{\sigma,\pi}$, $V_\tau \coloneqq \mathbb{E}\left[(\tau(\pi(X)) - \tau)^2\right]$, $V_{\sigma,\pi} \coloneqq \mathbb{E}\left[\frac{\sigma_0^2(\pi(X))}{1-\pi(X)} + \frac{\sigma_1^2(\pi(X))}{\pi(X)}\right]$.

Abadie and Imbens (2006) prove that nearest neighbor matching is asymptotically unbiased only when we match on a scalar covariate. Caliper matching

is very much alike. If we were to match on the K-dimensional covariates, similar arguments show that, under regularity conditions, the bias of $\sqrt{n}(\hat{\tau}_\pi - \tau)$ would be of the order $\sqrt{n}(\delta_n + \mathbb{1}_{\{\exists i \in [n]: M_i = 0\}})$ and the number of matches would be of the order $n\delta_n^K$. It would then be impossible to have sufficiently good match quality and enough matches at the same time for $K \geq 2$, so the bias B would not vanish. Therefore, it is crucial that we match on the scalar propensity score. When we do so, the ATT estimator $\hat{\tau}_{\mathrm{t},\pi}$ is also asymptotically normal.

Theorem 1.2 (Asymptotic Normality for Known Propensity Score (ATT))**.** *Suppose that $x \mapsto \pi(x)$ is known and the caliper δ_n satisfies (1.3). Let $p_1 \coloneqq \mathbb{E}\pi(X)$. If Assumptions 1.1, 1.2, 1.4 and 1.5 all hold, then*

$$\sqrt{n}(\hat{\tau}_{\mathrm{t},\pi} - \tau_{\mathrm{t}}) \rightsquigarrow \mathcal{N}(0, V_{\mathrm{t}}) \quad \text{as } n \to \infty,$$

where $V_{\mathrm{t}} \coloneqq V_{\tau_{\mathrm{t}}} + V_{\mathrm{t},\sigma,\pi}$ with $V_{\tau_{\mathrm{t}}} \coloneqq \frac{1}{p_1^2}\mathbb{E}\left[\pi(X)(\tau(\pi(X)) - \tau_{\mathrm{t}})^2\right]$ and

$$V_{\mathrm{t},\sigma,\pi} \coloneqq \frac{1}{p_1^2}\mathbb{E}\left[\frac{\pi(X)^2\sigma_0^2(\pi(X))}{1-\pi(X)} + \pi(X)\sigma_1^2(\pi(X))\right].$$

To examine the efficiency of $\hat{\tau}_\pi$ and $\hat{\tau}_{\mathrm{t},\pi}$, let

$$\mu_{\mathcal{X}}^d(x) \coloneqq \mathbb{E}\left[Y \mid D = d, X = x\right] \text{ and } \sigma_{\mathcal{X},d}^2(x) \coloneqq \mathbb{V}\left[Y \mid D = d, X = x\right]$$

for $d \in \{0,1\}$. The semiparametric efficiency bound of ATE is

$$V_{\mathrm{eff}} \coloneqq \mathbb{E}\left[(\mu_{\mathcal{X}}^1(X) - \mu_{\mathcal{X}}^0(X) - \tau)^2 + \frac{\sigma_{\mathcal{X},0}^2(X)}{1-\pi(X)} + \frac{\sigma_{\mathcal{X},1}^2(X)}{\pi(X)}\right], \tag{1.9}$$

irrespective of whether or not the propensity scores are known (Hahn (1998, Theorems 1 and 2)). The semiparametric efficiency bound of ATT is

$$V_{\mathrm{t},\mathrm{eff},\pi} \coloneqq$$
$$\frac{1}{p_1^2}\mathbb{E}\left[(\mu_{\mathcal{X}}^1(X) - \mu_{\mathcal{X}}^0(X) - \tau_{\mathrm{t}})^2\pi(X)^2 + \frac{\pi(X)^2\sigma_{\mathcal{X},0}^2(X)}{1-\pi(X)} + \pi(X)\sigma_{\mathcal{X},1}^2(X)\right]$$

if the propensity scores are known, and

$$V_{\mathrm{t},\mathrm{eff}} \coloneqq$$
$$\frac{1}{p_1^2}\mathbb{E}\left[(\mu_{\mathcal{X}}^1(X) - \mu_{\mathcal{X}}^0(X) - \tau_{\mathrm{t}})^2\pi(X) + \frac{\pi(X)^2\sigma_{\mathcal{X},0}^2(X)}{1-\pi(X)} + \pi(X)\sigma_{\mathcal{X},1}^2(X)\right]$$

if the propensity scores are unknown (Hahn (1998, Theorems 1 and 2)). The limiting variance V of $\hat{\tau}_\pi$ resembles the efficiency bound V_{eff}, except that V involves moments of the outcome conditional on the propensity score $\pi(X)$, rather than on the covariates X as in V_{eff}. Hence, if we were to observe only $\pi(X)$ in our sample, instead of X, $\hat{\tau}_\pi$ would be semiparametrically efficient, reaching V_{eff}. It is also immediate from Theorem 1.1 and (1.9), that if we had $\mu^d_{\mathcal{X}}(X) = \mu^d(\pi(X))$ and $\sigma^2_{\mathcal{X},d}(X) = \sigma^2_d(\pi(X))$ for all $d \in \{0,1\}$ — so that the conditional moments of the outcome given the covariates only depended on the propensity score —, then too, the ATE estimator $\hat{\tau}_\pi$ would be semiparametrically efficient. In truth, a more precise result in Proposition 1.3 holds.

Proposition 1.3 (Semiparametric Efficiency)**.** *Suppose Assumption 1.1 holds. Then $V_{\text{eff}} \leq V$ and $V_{\text{t,eff}} \leq V_{\text{t}}$ with equality in both cases if and only if*

$$\mu^D_{\mathcal{X}}(X) = \mu^D(\pi(X)) \quad \textit{almost surely}. \tag{1.10}$$

Suppose that (1.10) in Proposition 1.3 holds. Even then, in contrast to the ATE estimator $\hat{\tau}_\pi$, the ATT estimator $\hat{\tau}_{\text{t},\pi}$ only reaches $V_{\text{t,eff}}$, the semiparametric efficiency bound for *unknown* propensity scores, which is larger than the bound $V_{\text{t,eff},\pi}$ for known propensity scores. The difference between them, under (1.10), is

$$V_{\text{t,eff}} - V_{\text{t,eff},\pi} = \frac{1}{p_1^2}\mathbb{E}\left[\pi(X)(1-\pi(X))(\tau(\pi(X)) - \tau_{\text{t}})^2\right] \geq 0. \tag{1.11}$$

As $\pi(X)(1-\pi(X)) \leq 1/2$, the difference is bounded by $\frac{1}{2p_1^2}\mathbb{E}\left[(\tau(\pi(X)) - \tau_{\text{t}})^2\right]$. Thus, the more homogeneous the treatment effects are across $\pi(X)$ (equivalently, under (1.10), across X) and the treatment groups D, the smaller the difference is.

The efficiency loss (1.11) is not specific to caliper matching. In fact, the limiting variance of the ATT estimator in Theorem 1.2 is lower than that of the nearest neighbor matching estimator in Abadie and Imbens (2016, Proposition 1). The difference is

$$\frac{1}{2Mp_1^2}\mathbb{E}\left[\sigma_0^2(\pi(X))\pi(X)\left(2 + \frac{\pi(X)}{1-\pi(X)}\right)\right] \geq 0, \tag{1.12}$$

where the *constant* M is the number of nearest neighbors to match. This shows that the efficiency gain (1.12) of caliper matching is smaller for larger M. However, there is no proof that letting M to infinity closes the gap as the results of Abadie and Imbens (2016) are contingent on a fixed M. In contrast, with the caliper choice of Theorem 1.2, the number of matches for caliper matching goes to infinity by Proposition 1.2, thereby cutting variance. Unless $p \mapsto \sigma_0^2(p)$ decreases rapidly around one, which is ruled out by Assumption 1.4(i), (1.12) is larger when the propensity score tends to be close to one. In that case, we gain even more by using caliper instead of nearest neighbor matching, although then $V_{\mathrm{t},\sigma,\pi}$, and thus V_{t}, increases too.

We close the case for the known propensity score by verifying the assumptions of Theorems 1.1 and 1.2 for the models of Example 1.1.

Proposition 1.4 (Admissible Models (Known Propensity Score))**.** *The family of models described in Example 1.1 satisfies all Assumptions 1.1, 1.2, 1.4 and 1.5.*

1.3.3. Estimated Propensity Score

Suppose that the propensity score $\pi(\cdot, \theta_0)$ of Assumption 1.3 is estimated. A reasonable estimator of θ_0 will converge to θ_0. We then expect that if local versions of Assumptions 1.2, 1.4 and 1.5 hold in the neighbourhood of θ_0, then the caliper matching estimators on the estimated propensity scores will also be asymptotically normal, provided they are smooth enough in θ.

To this end, we require the conditional distribution

$$F_{d,\theta}(p) \coloneqq \mathbb{P}_{\theta_0}\left(\pi(X,\theta) \leq p \mid D = d\right)$$

to resemble that of the true propensity score, but only locally. Extending Assumption 1.2, we need that the densities $f_{0,\theta}, f_{1,\theta}$ are not only continuous but differentiable, and that they depend smoothly on θ. For some arbitrary fixed constant $\epsilon > 0$, let $\mathrm{Nb}(\theta_0, \epsilon) \coloneqq \{\theta \in \Theta : \|\theta - \theta_0\| < \epsilon\}$ denote a neighbourhood of θ_0, and further let

$$\mathcal{S}_{\theta_0,\epsilon} \coloneqq \left\{(\theta, p) : p \in [\underline{p}_\theta, \bar{p}_\theta], \theta \in \mathrm{Nb}(\theta_0, \epsilon)\right\}.$$

Assumption 1.6 (Distribution of the Parametric Propensity Score)**.** *(i) $F_{0,\theta}$, $F_{1,\theta}$ admit densities $f_{0,\theta}, f_{1,\theta}$, respectively, for all $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$.*

(ii) $f_{0,\theta}, f_{1,\theta}$ *have the same support* $[\underline{p}_\theta, \bar{p}_\theta]$ *with* $0 < \underline{p}_\theta < \bar{p}_\theta < 1$ *for all* $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$.

(iii) $f_{0,\theta}, f_{1,\theta}$ *are bounded away from zero:* $\inf_{\theta \in \mathrm{Nb}(\theta_0,\epsilon)} \inf_{p \in [\underline{p}_\theta, \bar{p}_\theta]} f_{d,\theta}(p) > 0$ *for all* $d \in \{0, 1\}$.

(iv) $(\theta, p) \mapsto f_{d,\theta}(p)$ *is continuously differentiable on* $\mathcal{S}_{\theta_0,\epsilon}$ *for all* $d \in \{0, 1\}$.

Next, we decompose the outcome in a way that depends on the propensity score parameter θ. Rather than the continuity of Assumption 1.5, we need that the regression function $\mu^d(\theta, p) := \mathbb{E}[Y \mid D = d, \pi(X, \theta) = p]$ is continuously differentiable, also in θ. In combination with Assumption 1.3, Assumption 1.7 implies that $\theta \mapsto \mu^d(\theta, \pi(x, \theta))$ can be approximated in the neighbourhood of θ_0 with an error of the order $\|\theta - \theta_0\|$. Specifically, they imply that the derivative of $\theta \mapsto \mu^d(\theta, \pi(x, \theta))$ exists for all $(\tilde{\theta}, x) \in \mathrm{Nb}(\theta_0, \epsilon) \times \mathcal{X}$ and it takes the form $\Lambda^d(\tilde{\theta}, x) := \frac{\partial \mu^d}{\partial \theta^\intercal}(\tilde{\theta}, \pi(x, \tilde{\theta})) + \frac{\partial \mu^d}{\partial p}(\tilde{\theta}, \pi(x, \tilde{\theta}))(\mathrm{D}_\theta \pi)(x, \tilde{\theta})$.

Assumption 1.7 (Differentiability of Regression Functions)**.** *The maps* $(\theta, p) \mapsto \mu^d(\theta, p)$ *are continuously differentiable on* $\mathcal{S}_{\theta_0,\epsilon}$ *with partial derivatives* $\frac{\partial \mu^d}{\partial \theta} : \Theta \times [0, 1] \to \mathbb{R}^K$ *and* $\frac{\partial \mu^d}{\partial p} : \Theta \times [0, 1] \to \mathbb{R}$ *uniformly bounded on* $\mathcal{S}_{\theta_0,\epsilon}$ *for all* $d \in \{0, 1\}$.

To ensure the smoothness, and to control the magnitude of the disturbance term $\varepsilon_i(\theta) := Y_i - \mu^{D_i}(\theta, \pi(X_i, \theta)), i \in [n]$, we require that the functions

$$\sigma_d^r(\theta, p) := \mathbb{E}\left[(Y - \mu^D(\theta, p))^r \,\middle|\, D = d, \pi(X, \theta) = p\right], \quad r \in \{2, 4\}, \, d \in \{0, 1\},$$

satisfy the following conditions.

Assumption 1.8 (Smooth Parametric Disturbance Term)**.** *(i) The* σ_d^2 *satisfy the Lipschitz-condition* $|\sigma_d^2(\theta, p) - \sigma_d^2(\theta', p')| \leq L_\sigma(\|\theta - \theta'\| + |p - p'|)$ *for all* $p \in [\underline{p}_\theta, \bar{p}_\theta]$ *and* $p' \in [\underline{p}_{\theta'}, \bar{p}_{\theta'}]$ *for all* $\theta, \theta' \in \mathrm{Nb}(\theta_0, \epsilon)$ *for some constant* $0 < L_\sigma < \infty$ *and the lower bound* $\inf_{p \in [\underline{p}_{\theta_0}, \bar{p}_{\theta_0}]} \sigma_d^2(\theta_0, p) > 0$ *for all* $d \in \{0, 1\}$.

(ii) The σ_d^4 *satisfy the condition* $\sup_{\theta \in \mathrm{Nb}(\theta_0,\epsilon)} \sup_{p \in [\underline{p}_\theta, \bar{p}_\theta]} \sigma_d^4(\theta, p) < \infty$ *for all* $d \in \{0, 1\}$.

Finally, we need that the estimator $\hat{\theta}$ of the propensity score parameter converges to θ_0 in an appropriate sense. For instance, if $\hat{\theta}$ is the maximum

likelihood estimator, it converges appropriately under regularity conditions. We further assume that θ_0 is estimated from a sample that is independent of $((Y_i, D_i, X_i))_{i\in[n]}$. In practice, sample splitting may be applied to ensure the independence: one can halve a $2n$-large sample and use the first half to estimate θ_0, and plug the resulting estimator $\hat{\theta}$ back into the second half to compute $\hat{\tau}_{\hat{\pi}}, \hat{\tau}_{\mathrm{t},\hat{\pi}}$.

Assumption 1.9 (Estimator of the Propensity Score Parameter)**.** *(i) $\hat{\theta}$ is asymptotically normal with $\sqrt{n}(\hat{\theta} - \theta_0) \rightsquigarrow \mathcal{N}(0, V_{\theta_0})$ as $n \to \infty$ for a finite invertible matrix V_{θ_0}.*

(ii) $\hat{\theta}$ is independent of the data set from which the matching estimator is computed: $\hat{\theta} \perp\!\!\!\perp ((Y_i, X_i, D_i))_{i\in[n]}$.

To accommodate the propensity score estimation, we introduce

$$\widehat{\underline{\Delta}}_n := \max_{i\in[n]} \min_{j\in[n]:D_j\neq D_i} |\pi(X_i, \hat{\theta}) - \pi(X_j, \hat{\theta})|,$$

the estimated analogue of $\underline{\Delta}_n$, and the corresponding caliper choices

$$\delta_n := s\frac{\log n}{n} \quad \text{or} \quad \delta_n := \widehat{\underline{\Delta}}_n \vee \frac{\log N_0}{N_0 + 1} \vee \frac{\log N_1}{N_1 + 1} \tag{1.13}$$

for any constant $0 < s \leq 2/\log(2)$ and $n \geq 2$. Proposition 1.5 shows that the number of matches based on the estimated propensity scores and the caliper choice (1.13) is also of the order $\log n$ as in Proposition 1.2. This yields Theorems 1.3 and 1.4, establishing the asymptotic normality of caliper matching on the estimated propensity score.

Proposition 1.5 (Number of Matches for Estimated Propensity Score)**.** *Suppose that the caliper δ_n satisfies (1.13). If Assumptions 1.6 and 1.9 hold, then there exist constants $0 < \bar{c}_l, \bar{c}_u < \infty$ such that*

$$\bar{c}_l(1 + o_P(1)) \log n \leq \min_{i\in[n]} M_i(\hat{\theta}) \leq \max_{i\in[n]} M_i(\hat{\theta}) \leq \bar{c}_u(1 + o_P(1)) \log n$$

as $n \to \infty$. Thus, $\mathbb{P}\left(\min_{i\in[n]} M_i(\hat{\theta}) \geq 1\right) \to 1$ as $n \to \infty$.

Theorem 1.3 (Asymptotic Normality for Estimated Propensity Score (ATE))**.** *Suppose that the caliper δ_n satisfies* (1.13)*. If Assumptions 1.1, 1.3 and 1.6 to 1.9 all hold, then*

$$\sqrt{n}(\hat{\tau}_{\hat{\pi}} - \tau) \rightsquigarrow \mathcal{N}(0, V_{\hat{\pi}}) \quad \text{as } n \to \infty,$$

where $V_{\hat{\pi}} \coloneqq V_\tau + V_{\sigma,\pi} + (q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0)$ for $V_\tau, V_{\sigma,\pi}$ of Theorem 1.1, V_{θ_0} of Assumption 1.9, and $q_d \in \mathbb{R}^K$ arising as the probability limit $\frac{1}{n}\sum_{i\in[n]} \Lambda^d(\hat{\theta}, X_i) \xrightarrow{P} q_d^\intercal$ as $n \to \infty$ for $d \in \{0, 1\}$.

Theorem 1.4 (Asymptotic Normality for Estimated Propensity Score (ATT))**.** *Suppose that the caliper δ_n satisfies* (1.13)*. If Assumptions 1.1, 1.3 and 1.6 to 1.9 all hold, then*

$$\sqrt{n}(\hat{\tau}_{\mathrm{t},\hat{\pi}} - \tau_{\mathrm{t}}) \rightsquigarrow \mathcal{N}(0, V_{\mathrm{t},\hat{\pi}}) \quad \text{as } n \to \infty,$$

where $V_{\mathrm{t},\hat{\pi}} \coloneqq V_{\tau_{\mathrm{t}}} + V_{\mathrm{t},\sigma,\pi} + (1/p_1^2)(q_{\mathrm{t},1} - q_{\mathrm{t},0})^\intercal V_{\theta_0}(q_{\mathrm{t},1} - q_{\mathrm{t},0})$ for $V_{\tau_{\mathrm{t}}}$, $V_{\mathrm{t},\sigma,\pi}$ and p_1 of Theorem 1.2, V_{θ_0} of Assumption 1.9, and $q_{\mathrm{t},d} \in \mathbb{R}^K$ arising as the probability limit

$$\frac{1}{n}\sum_{i\in[n]} D_i \Lambda^d(\hat{\theta}, X_i) \xrightarrow{P} q_{\mathrm{t},d}^\intercal$$

as $n \to \infty$ for $d \in \{0, 1\}$.

Compared to Theorems 1.1 and 1.2, the variances are increased by $(q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0)$ for the ATE estimator $\hat{\tau}_{\hat{\pi}}$, and by $(1/p_1^2)(q_{\mathrm{t},1} - q_{\mathrm{t},0})^\intercal V_{\theta_0}(q_{\mathrm{t},1} - q_{\mathrm{t},0})$ for the ATT estimator $\hat{\tau}_{\mathrm{t},\hat{\pi}}$, representing the uncertainty from the propensity score estimation. The more precisely we can estimate the propensity score, the smaller V_{θ_0} is, resulting in smaller differences. Alternatively, if $q_1 \approx q_0$ or $q_{\mathrm{t},1} \approx q_{\mathrm{t},0}$, then the respective increments are also small. This is the case if the derivatives Λ^1 and Λ^0 are close to each other, although it is difficult to see if and when that happens, even for simple linear regressions in Example 1.1. As a consequence, it remains unclear whether caliper or nearest neighbor matching (Abadie and Imbens, 2016) is more efficient when the parametric propensity score is estimated.

Remark 1 (Variance Comparison)**.** *The asymptotic variances of $\hat{\tau}_\pi, \hat{\tau}_{\mathrm{t},\pi}$ and $\hat{\tau}_{\hat{\pi}}, \hat{\tau}_{\mathrm{t},\hat{\pi}}$ are comparable as in the preceding paragraph if and only if we use only half of a $2n$-large sample to compute $\hat{\tau}_\pi, \hat{\tau}_{\mathrm{t},\pi}$ and the caliper* (1.3)*, because of*

the sample-splitting in the computation of $\hat{\tau}_{\hat{\pi}}, \hat{\tau}_{\mathrm{t},\hat{\pi}}$ *and* (1.13)*. If we use the whole* $2n$*-large sample to compute* $\hat{\tau}_{\pi}, \hat{\tau}_{\mathrm{t},\pi}$*,* (1.3)*, and only* n *observations to evaluate* $\hat{\tau}_{\hat{\pi}}, \hat{\tau}_{\mathrm{t},\hat{\pi}}$*,* (1.13) *— with the remaining* n *observations reserved to estimate* θ_0 *—, then the standard error of* $\hat{\tau}_{\pi}$ *is* $\sqrt{\frac{V_\tau + V_{\sigma,\pi}}{2n}}$*, while that of* $\hat{\tau}_{\hat{\pi}}$ *is* $\sqrt{\frac{V_\tau + V_{\sigma,\pi} + (q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0)}{n}}$*, which is an increment by a factor of* $\sqrt{2}$ *even without the contribution of* $(q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0)$*. The same applies to* $\hat{\tau}_{\mathrm{t},\pi}$ *and* $\hat{\tau}_{\mathrm{t},\hat{\pi}}$*.*

Abadie and Imbens (2016) account for the estimation of the propensity score by considering a shifted law of $(Y, D, X) \sim \mathbb{P}_{\theta_0}$. They assume that conditional expectations under the shifted law converge weakly to conditional expectations under the nonshifted law. We pursue a different approach. Our Assumptions 1.6 to 1.8 do not involve shifted laws. Rather, they impose smoothness of conditional expectations in θ and may be regarded as local versions of Assumptions 1.2, 1.4 and 1.5 in the neighbourhood of θ_0. Moreover, we verify the assumptions of Theorems 1.3 and 1.4 for the models in Example 1.1.

Proposition 1.6 (Admissible Models (Estimated Propensity Score))**.** *Consider the family of models described in Example 1.1, with the propensity score model* $\{\pi(x, \theta) = g(\theta^\intercal x) : \theta \in \Theta\}$ *estimated with maximum likelihood on an independent* n*-large i.i.d. sample from the distribution of* (D, X)*. Then Assumptions 1.1, 1.3 and 1.6 to 1.9 are all satisfied.*

1.3.4. Variance Estimation

In this section, we provide consistent estimators for the components of $V_{\hat{\pi}}$ and $V_{\mathrm{t},\hat{\pi}}$ so that we can construct asymptotically valid confidence intervals for ATE and ATT. To prove consistency, we impose some further assumptions, which are all in accordance with the models in Example 1.1.

Namely, we need that certain estimators are almost surely bounded, which is implied if the outcome is almost surely bounded. Furthermore, $\theta \mapsto \pi(\cdot, \theta)$ may take many forms in general, which renders Λ^d intractable. Requiring that the propensity score follows a single-index model, such as the logit or the probit, and that the covariates have a well-behaved density, alleviates these difficulties, provided the outcome regression is smooth enough. Imposing $K \geq 2$

continuously distributed covariates and certain smoothness conditions implies that Λ^d is expressible in a way suitable for showing consistency.

Assumption 1.10 (Outcome and Covariate Distribution)**.** *(i) The outcome is almost surely bounded: there exists a constant $0 < \bar{y} < \infty$ such that $\mathbb{P}(|Y| > \bar{y}) = 0$.*

(ii) The covariate vector X has at least $K \geq 2$ coordinates, and X admits a density Ψ on the compact $\mathcal{X}$; the Ψ is as specified in Example 1.1.

Assumption 1.11 (Single-Index Propensity Score and Smooth Outcome Regression)**.**

(i) The propensity score model of Assumption 1.3 is $\pi(x, \theta) = g(\theta^\intercal x)$ for g as specified in Example 1.1.

(ii) The $m(x) \coloneqq \mathbb{E}[Y \mid X = x]$ is bounded, and there exist two covariates — X_1 and X_2 without loss of generality — such that $\frac{\partial m}{\partial x_1}$ and $\frac{\partial m}{\partial x_2}$ are well-defined and continuous for all $x \in \mathcal{X}$.

The variance estimators are

$$\hat{V}_{\hat{\pi}} \coloneqq \hat{V}_\tau + \hat{V}_{\sigma,\pi} + (\hat{q}_1 - \hat{q}_0)^\intercal \hat{V}_{\theta_0} (\hat{q}_1 - \hat{q}_0), \tag{1.14}$$

$$\hat{V}_{\mathrm{t},\hat{\pi}} \coloneqq \hat{V}_{\tau_\mathrm{t}} + \hat{V}_{\mathrm{t},\sigma,\pi} + (1/\hat{p}_1^2)(\hat{q}_{\mathrm{t},1} - \hat{q}_{\mathrm{t},0})^\intercal \hat{V}_{\theta_0} (\hat{q}_{\mathrm{t},1} - \hat{q}_{\mathrm{t},0}), \tag{1.15}$$

where the component estimators are as follows. We assume that $\hat{V}_{\theta_0} \xrightarrow{P} V_{\theta_0}$ is a consistent estimator of V_{θ_0}. In practice, $\hat{\theta}$ is usually the maximum likelihood estimator, as supported by Proposition 1.6, in which case, under Assumption 1.11,

$$\hat{V}_{\theta_0} \coloneqq \left(\frac{1}{n} \sum_{i \in [n]} \frac{(g'(\hat{\theta}^\intercal X_i))^2}{g(\hat{\theta}^\intercal X_i)(1 - g(\hat{\theta}^\intercal X_i))} X_i X_i^\intercal \right)^{-1}$$

is well-known to be consistent for V_{θ_0}. The p_1 is consistently estimated with $\hat{p}_1 \coloneqq \frac{1}{n} \sum_{i \in [n]} D_i$ by the law of large numbers. The nonparametric estimators of the remaining components in (1.14) and (1.15) are developed in Section 1.A.

Proposition 1.7 (Consistent Variance Estimators)**.** *Suppose Assumptions 1.1, 1.3 and 1.6 to 1.11 all hold, the caliper δ_n satisfies (1.13), and $\hat{V}_{\theta_0} \xrightarrow{P} V_{\theta_0}$ as*

$n \to \infty$. Then $\hat{V}_{\hat{\pi}} \xrightarrow{P} V_{\hat{\pi}}$ and $\hat{V}_{\mathrm{t},\hat{\pi}} \xrightarrow{P} V_{\mathrm{t},\hat{\pi}}$ as $n \to \infty$. In particular, the estimators on the right side of (1.14) *and* (1.15) *are all consistent for their respective estimands.*

In view of Theorems 1.3 and 1.4, an immediate implication is that we can construct asymptotic confidence intervals for ATE and ATT. Let $z_{1-\alpha/2}$ be the $(1-\alpha/2)$th quantile of the standard normal distribution for $\alpha \in (0,1)$, and let $[a \pm b]$ denote the interval $[a-b, a+b]$ for $a, b \in \mathbb{R}$, $b \geq 0$. Then the intervals $[\hat{\tau}_{\hat{\pi}} \pm z_{1-\alpha/2}(\hat{V}_{\hat{\pi}}/n)^{1/2}]$ and $[\hat{\tau}_{\mathrm{t},\hat{\pi}} \pm z_{1-\alpha/2}(\hat{V}_{\mathrm{t},\hat{\pi}}/n)^{1/2}]$ are asymptotically valid confidence intervals for ATE and ATT, respectively.

Corollary 1.1 (Asymptotic Confidence Intervals)**.** *Suppose Assumptions 1.1, 1.3 and 1.6 to 1.11 all hold, the caliper δ_n satisfies* (1.13)*, and $\hat{V}_{\theta_0} \xrightarrow{P} V_{\theta_0}$ as $n \to \infty$. Then, as $n \to \infty$, $\mathbb{P}\left(\left[\hat{\tau}_{\hat{\pi}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{V}_{\hat{\pi}}}{n}}\right] \ni \tau\right) \to 1-\alpha$ and $\mathbb{P}\left(\left[\hat{\tau}_{\mathrm{t},\hat{\pi}} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{V}_{\mathrm{t},\hat{\pi}}}{n}}\right] \ni \tau_{\mathrm{t}}\right) \to 1-\alpha$.*

1.4. Conclusion

We studied the caliper matching estimator when matching is performed on the (estimated) propensity scores. We proposed a caliper, and proved that the resulting estimator of the Average Treatment Effect (ATE), and of the Average Treatment Effect on the Treated (ATT), is asymptotically unbiased and normal.

When the propensity score is known, our estimator of ATE reaches the semiparametric lower bound in the restricted model where only the propensity scores and not the covariates are observed in the sample or where the outcome regression on the covariates only depend on the propensity score. In this restricted model, the estimator of ATT only reaches the larger lower bound corresponding to unknown propensity score. Even in the unrestricted model, both our estimators are more efficient than nearest neighbor matching estimators on the known propensity scores, and are, therefore, preferred over the latter method in the large sample limit, provided our assumptions hold. When the parametric propensity score is estimated, the variances of both our esti-

mators increase, hence it remains unclear whether caliper or nearest neighbor matching will be more efficient.

We facilitated the empirical application of the estimator by verifying our assumptions for a family of often employed models, and by constructing asymptotic confidence intervals for the average treatment effects. An interesting avenue for future research is to study in-sample estimation of the propensity score, and to allow for nonparametric propensity score estimators. The main challenge arising is to see how uncertainty from the propensity score estimation propagates to the matching estimator, which is more difficult to quantify for nonparametric models.

1.A. Variance Estimation

In this section, we define the variance estimators of Section 1.3.4 for the components in (1.14) and (1.15). Let $K(u) = (2\pi)^{-1/2}e^{-u^2/2}, u \in \mathbb{R}$, be the Gaussian kernel and $K'(u)$ its derivative. Let $0 < \gamma_n \lesssim a_n$ be two arbitrary sequences $\gamma_n := \kappa_0 n^{-\beta}$, $a_n := \kappa_1 n^{-\alpha}$ for fixed finite constants $0 < \alpha < \beta < 1/4$ and $\kappa_0, \kappa_1 > 0$. We employ a truncation strategy to avoid bias at the boundaries. Define the intervals $A_n := [\underline{p}_{\hat{\theta}} + a_n, \bar{p}_{\hat{\theta}} - a_n]$ and $\hat{A}_n := [\min_{i\in[n]} g(\hat{\theta}^\intercal X_i) + a_n, \max_{i\in[n]} g(\hat{\theta}^\intercal X_i) - a_n]$, which are well-defined with probability tending to one as $\hat{\theta} \xrightarrow{P} \theta_0$ under Assumption 1.9; see the proof of Proposition 1.7.[6] Let $\hat{N} := \sum_{i\in[n]} \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n}$. The estimators of the first components are

$$\hat{V}_\tau := \left(\frac{1}{\hat{N}}\sum_{i\in[n]}[\hat{\mu}^1(\hat{\theta}, g(\hat{\theta}^\intercal X_i)) - \hat{\mu}^0(\hat{\theta}, g(\hat{\theta}^\intercal X_i))]^2 \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n}\right) - \hat{\tau}^2_{\hat{\pi}}, \quad (1.16)$$

$$\hat{V}_{\tau_t} := \frac{1}{\hat{p}_1^2}\left(\frac{1}{\hat{N}}\sum_{i\in[n]} D_i[\hat{\mu}^1(\hat{\theta}, g(\hat{\theta}^\intercal X_i)) - \hat{\mu}^0(\hat{\theta}, g(\hat{\theta}^\intercal X_i))]^2 \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n}\right) - \frac{\hat{\tau}^2_{t,\hat{\pi}}}{\hat{p}_1},$$

where

$$\hat{\mu}^d(\theta, p) := \frac{\hat{q}_{\mu,d}(\theta, p)}{\hat{h}_d(\theta, p)}, \quad \hat{q}_{\mu,d}(\theta, p) := \frac{1}{N_d\gamma_n}\sum_{j\in[n]} \mathbb{1}_{D_j=d} Y_j K\left(\frac{g(\theta^\intercal X_j) - p}{\gamma_n}\right),$$

$$\hat{h}_d(\theta, p) := \hat{f}_{\theta,d}(p) := \frac{1}{N_d\gamma_n}\sum_{j\in[n]} \mathbb{1}_{D_j=d} K\left(\frac{g(\theta^\intercal X_j) - p}{\gamma_n}\right), \quad d \in \{0, 1\};$$

[6] In practice, especially for moderate sample sizes, γ_n and a_n should be chosen carefully to ensure nonnegative variance estimates. The a_n should be chosen small enough to enlarge $A_n, \hat{A}_n$; for instance, one could choose κ_1 arbitrary close to zero and $\alpha := 1/(4 + \varepsilon_\alpha)$ for an $\varepsilon_\alpha > 0$ arbitrarily close to zero. As a rule, γ_n should be set small too to minimise the bias of the variance estimates by standard nonparametric theory, thereby avoiding negative values; to accommodate $\alpha < \beta$, one can set $\beta = 1/(4 + \varepsilon_\beta)$ with $0 < \varepsilon_\beta < \varepsilon_\alpha$, for example, $\varepsilon_\beta := \varepsilon_\alpha/2$. The κ_0 should be chosen to accommodate the different scales of $(g(\hat{\theta}^\intercal X_i))_{i\in[n]}$ and $(\hat{\theta}^\intercal X_i)_{i\in[n]}$ present in the estimation of the μ^d and their derivate, respectively. A small κ_0 is a safe but conservative choice. Note that asymptotically the effect of truncation disappears ($\mathbb{E}\hat{N}/n \to 1$) as shown in Proposition 1.7.

and those of the second components are

$$\hat{V}_{\sigma,\pi} \coloneqq \frac{1}{\hat{N}} \sum_{i\in[n]} \left(\frac{\hat{\sigma}_0^2(\hat{\theta}, g(\hat{\theta}^\intercal X_i))}{1 - g(\hat{\theta}^\intercal X_i)} + \frac{\hat{\sigma}_1^2(\hat{\theta}, g(\hat{\theta}^\intercal X_i))}{g(\hat{\theta}^\intercal X_i)} \right) \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n},$$

$$\hat{V}_{\mathrm{t},\sigma,\pi} \coloneqq \frac{1}{\hat{p}_1^2 \hat{N}} \sum_{i\in[n]} \left(\frac{g(\hat{\theta}^\intercal X_i)^2 \hat{\sigma}_0^2(\hat{\theta}, g(\hat{\theta}^\intercal X_i))}{1 - g(\hat{\theta}^\intercal X_i)} + g(\hat{\theta}^\intercal X_i)\hat{\sigma}_1^2(\hat{\theta}, g(\hat{\theta}^\intercal X_i)) \right) \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n},$$

where

$$\hat{\sigma}_d^2(\theta, p) \coloneqq \hat{\mu}_2^d(\theta, p) - (\hat{\mu}^d(\theta, p))^2, \quad \hat{\mu}_2^d(\theta, p) \coloneqq \frac{\hat{q}_{\mu_2,d}(\theta, p)}{\hat{h}_d(\theta, p)},$$
$$\hat{q}_{\mu_2,d}(\theta, p) \coloneqq \frac{1}{N_d \gamma_n} \sum_{j\in[n]} \mathbb{1}_{D_j=d} Y_j^2 K\left(\frac{g(\theta^\intercal X_j) - p}{\gamma_n} \right), \quad d \in \{0, 1\},$$

is an estimator of $\sigma_d^2(\theta, p) = \mu_2^d(\theta, p) - (\mu^d(\theta, p))^2$ with

$$\mu_2^d(\theta, p) \coloneqq \mathbb{E}\left[Y^2 \,\middle|\, D = d, \pi(X, \theta) = p \right], \quad d \in \{0, 1\}.$$

Last, the probability limits of the derivatives are estimated by

$$\hat{q}_d^\intercal \coloneqq \frac{1}{\hat{N}} \sum_{i\in[n]} \hat{\Lambda}^d(\hat{\theta}, X_i) \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n}, \quad \hat{q}_{\mathrm{t},d}^\intercal \coloneqq \frac{1}{\hat{N}} \sum_{i\in[n]} D_i \hat{\Lambda}^d(\hat{\theta}, X_i) \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n},$$
$$\hat{\Lambda}^d(\theta, x) \coloneqq \widehat{\left(\frac{\partial \mu^d}{\partial \theta^\intercal} \right)}(\theta, g(\theta^\intercal x)) + \widehat{\left(\frac{\partial \mu^d}{\partial p} \right)}(\theta, g(\theta^\intercal x)) g'(\theta^\intercal x) x^\intercal,$$

where

$$\widehat{\left(\frac{\partial \mu^d}{\partial \theta_k} \right)}(\theta, p) \coloneqq \frac{\widehat{\left(\frac{\partial q_{\mu,d}}{\partial \theta_k} \right)}(\theta, p) \hat{h}_d(\theta, p) - \hat{q}_{\mu,d}(\theta, p) \widehat{\left(\frac{\partial h_d}{\partial \theta_k} \right)}(\theta, p)}{(\hat{h}_d(\theta, p))^2},$$
$$\widehat{\left(\frac{\partial q_{\mu,d}}{\partial \theta_k} \right)}(\theta, p) \coloneqq \frac{(g^{-1})'(p)}{N_d \gamma_n^2} \sum_{j\in[n]} \mathbb{1}_{D_j=d} Y_j X_{j,k} K'\left(\frac{\theta^\intercal X_j - g^{-1}(p)}{\gamma_n} \right),$$
$$\widehat{\left(\frac{\partial h_d}{\partial \theta_k} \right)}(\theta, p) \coloneqq \frac{(g^{-1})'(p)}{N_d \gamma_n^2} \sum_{j\in[n]} \mathbb{1}_{D_j=d} X_{j,k} K'\left(\frac{\theta^\intercal X_j - g^{-1}(p)}{\gamma_n} \right),$$

with θ_k ($X_{j,k}$) being the kth coordinate of θ (X_j) for $k \in [K]$, and

$$\widehat{\left(\frac{\partial \mu^d}{\partial p}\right)}(\theta, p) := \frac{\left(\frac{\partial}{\partial p}\hat{q}_{\mu,d}(\theta,p)\right)\hat{h}_d(\theta,p) - \hat{q}_{\mu,d}(\theta,p)\frac{\partial}{\partial p}\hat{h}_d(\theta,p)}{(\hat{h}_d(\theta,p))^2},$$
$$\frac{\partial}{\partial p}\hat{q}_{\mu,d}(\theta,p) = -\frac{1}{N_d\gamma_n^2}\sum_{j\in[n]} \mathbb{1}_{D_j=d} Y_j K'\left(\frac{g(\theta^\intercal X_j) - p}{\gamma_n}\right),$$
$$\frac{\partial}{\partial p}\hat{h}_d(\theta,p) = -\frac{1}{N_d\gamma_n^2}\sum_{j\in[n]} \mathbb{1}_{D_j=d} K'\left(\frac{g(\theta^\intercal X_j) - p}{\gamma_n}\right)$$

for $d \in \{0,1\}$. An intercept in the propensity score model can be accommodated by defining $X_{j,K+1} := 1$ for $j \in [n]$ and considering derivatives with respect to θ_{K+1} too.

1.B. Proofs of Main Results

In this section, we prove the main results, Propositions 1.1, 1.2 and 1.4 to 1.6 and Theorems 1.1 to 1.4 together with supporting Lemmas 1.1 to 1.4. The proofs of Propositions 1.3 and 1.7 and Lemmas 1.5 to 1.7 are in Section 1.C. For simplicity, we give the proofs for the caliper choice $\delta_n = s\frac{\log n}{n}$, $s := 1$, and provide remarks for the choices $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$ and $\delta_n = \widehat{\overline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$ when necessary.

We adopt the following notation. Let $D^{(n)} := (D_i)_{i\in[n]}$ and $PS^{(n)} := (\pi(X_i))_{i\in[n]}$. For $a, b \in \mathbb{R}, a < b$, let

$$F_d[a,b] := \mathbb{P}\left(\pi(X) \in [a,b] \mid D = d\right),$$
$$\mathbb{F}_{N_d}[a,b] := \frac{1}{N_d}\sum_{i:D_i=d} \mathbb{1}_{\pi(X_i)\in[a,b]}, \tag{1.17}$$

be the conditional (empirical) measures of intervals $[a,b]$ for $d \in \{0,1\}$. Similarly, under Assumption 1.3, define the conditional (empirical) measures

$$F_{d,\theta}[a,b] := \mathbb{P}\left(\pi(X,\theta) \in [a,b] \mid D = d\right),$$
$$\mathbb{F}_{N_d,\theta}[a,b] := \frac{1}{N_d}\sum_{i:D_i=d} \mathbb{1}_{\pi(X_i,\theta)\in[a,b]} \quad \text{for } \theta \in \Theta. \tag{1.18}$$

Let $\mathbb{G}_{N_d} := \sqrt{N_d}(\mathbb{F}_{N_d} - F_d)$ be the empirical process of

$$((\pi(X_i))_{i:D_i=d} \mid D^{(n)}) \overset{\text{i.i.d.}}{\sim} F_d,$$

and $[a \pm b]$ denote the interval $[a-b, a+b]$. In the proofs, the value of constants may change from equation to equation without explicit notice.

Proof of Proposition 1.1. Spacings and Their Order. Let $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(N_1)}$ be the order statistics of $(U_1, \ldots, U_{N_1} \mid D^{(n)}) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)$. Let $\tilde{U}_1 := U_{(1)}$, $\tilde{U}_i := U_{(i)} - U_{(i-1)}$ for $i = 2, \ldots, N_1$ and $\tilde{U}_{N_1+1} := 1 - U_{(N_1)}$ be the spacings generated by $(U_i)_{i\in[N_1]}$. Let $\tilde{U}_{(1)} \leq \tilde{U}_{(2)} \leq \ldots \leq \tilde{U}_{(N_1+1)}$ be the ordered spacings. Shorack and Wellner (2009, Chapter 21) prove that $\mathbb{E}\left[\tilde{U}_{(N_1+1)} \,\middle|\, D^{(n)}\right] = \frac{1}{N_1+1}\sum_{i=1}^{N_1+1} \frac{1}{N_1+2-i}$, which we apply as follows.

Bounding $\mathbb{E}\underline{\Delta}_n$ with Spacings. Let $\underline{\Delta}_i := \min_{j:D_j\neq D_i} |\pi(X_i) - \pi(X_j)|$ for $i \in [n]$, so that $\underline{\Delta}_n \leq \sum_{d\in\{0,1\}} \max_{i:D_i=d} \underline{\Delta}_i$. Consider $\max_{i:D_i=0} \underline{\Delta}_i$. Let $\pi_{(1)} \leq \pi_{(2)} \leq \ldots \leq \pi_{(N_1)}$ be the order statistics of $(\pi(X_i))_{i:D_i=1}$, and let $\tilde{\pi}_1 := \pi_{(1)} - \underline{p}$, $\tilde{\pi}_i := \pi_{(i)} - \pi_{(i-1)}$ for $i \in \{2, \ldots, N_1\}$ and $\tilde{\pi}_{N_1+1} := \bar{p} - \pi_{(N_1)}$ be the corresponding spacings for $\underline{p}, \bar{p}$ of Assumption 1.2. Let $\tilde{\pi}_{(1)} \leq \tilde{\pi}_{(2)} \leq \ldots \leq \tilde{\pi}_{(N_1+1)}$ be the order statistics of these spacings. Every propensity score $\pi(X_j)$ of the control units falls either to the left of $\pi_{(1)}$ or to the right of $\pi_{(N_1)}$ or between two propensity scores $\pi_{(i)}$ and $\pi_{(i-1)}$ for some $i \in \{2, 3, \ldots, N_1\}$. In all three cases, the closest treated propensity score to $\pi(X_j)$ is within $\tilde{\pi}_{(N_1+1)}$-distance. Hence, $\max_{i:D_i=0} \underline{\Delta}_i \leq \tilde{\pi}_{(N_1+1)}$.

Given $D^{(n)}$, the $\pi(X_i)$ restricted to $i : D_i = 1$ are i.i.d., with $(\pi(X_i) \mid D_i = 1) \sim F_1$. By Assumption 1.2, $\inf_{p\in[\underline{p},\bar{p}]} f_1(p) > 0$, thus F_1 is strictly increasing on $[\underline{p}, \bar{p}]$, and therefore has a strictly increasing inverse F_1^{-1} on $[F_1(\underline{p}), F_1(\bar{p})]$. Because F_1^{-1} is increasing, $((\pi_{(i)})_{i\in[N_1]} \mid D^{(n)}) \sim ((F_1^{-1}(U_{(i)}))_{i\in[N_1]} \mid D^{(n)})$ by the quantile transform. We distinguish three cases.

- Case 1: $\tilde{\pi}_{(N_1+1)} = \pi_{(i)} - \pi_{(i-1)}$ for some $i \in \{2, \ldots, N_1\}$. Then we bound $\tilde{\pi}_{(N_1+1)}$ by noting that $(\pi_{(i)} - \pi_{(i-1)} \mid D^{(n)}) \sim (F_1^{-1}(U_{(i)}) - F_1^{-1}(U_{(i-1)}) \mid D^{(n)})$, which is bounded by $\left\|(F_1^{-1})'\right\|_\infty (U_{(i)} - U_{(i-1)}) = \left\|(F_1^{-1})'\right\|_\infty \tilde{U}_i$, because F_1^{-1} is Lipschitz with constant $\left\|(F_1^{-1})'\right\|_\infty$ with $(F_1^{-1})'(u) = \frac{1}{f_1(F_1^{-1}(u))}$ finite as $\inf_{p\in[\underline{p},\bar{p}]} f_1(p) > 0$ by Assumption 1.2.

- Case 2: $\tilde{\pi}_{(N_1+1)} = \pi_{(1)} - \underline{p}$. Then write $\underline{p} = F_1^{-1}(F_1(\underline{p})) = F_1^{-1}(0)$, so $(\tilde{\pi}_{(N_1+1)} \mid D^{(n)})$ is distributed as a random variable that is bounded by $\left\|(F_1^{-1})'\right\|_\infty (U_{(1)} - 0) = \left\|(F_1^{-1})'\right\|_\infty \tilde{U}_1$.

- Case 3: $\tilde{\pi}_{(N_1+1)} = \bar{p} - \pi_{(N_1)}$. Then write $\bar{p} = F_1^{-1}(F_1(\bar{p})) = F_1^{-1}(1)$, so $(\tilde{\pi}_{(N_1+1)} \mid D^{(n)})$ is distributed as a random variable that is bounded by $\left\|(F_1^{-1})'\right\|_\infty (1 - U_{(N_1)}) = \left\|(F_1^{-1})'\right\|_\infty \tilde{U}_{N_1+1}$.

Conclude that $(\tilde{\pi}_{(N_1+1)} \mid D^{(n)})$ is distributed as a random variable that is bounded by $\left\|(F_1^{-1})'\right\|_\infty \tilde{U}_{(N_1+1)}$. Thus,

$$\begin{aligned}
\mathbb{E} \max_{i:D_i=0} \underline{\Delta}_i &\leq \mathbb{E}\mathbb{E}\left[\tilde{\pi}_{(N_1+1)} \mid D^{(n)}\right] \lesssim \mathbb{E}\left[\frac{1}{N_1+1}\sum_{i=1}^{N_1+1}\frac{1}{N_1+2-i}\right] \\
&= \sum_{n_1=0}^{n}\left[\frac{1}{n_1+1}\sum_{i=1}^{n_1+1}\frac{1}{n_1+2-i}\right]\binom{n}{n_1}p_1^{n_1}(1-p_1)^{n-n_1} \\
&= (1-p_1)^n + \sum_{n_1=1}^{n}\left[\frac{1}{n_1+1}\sum_{i=1}^{n_1+1}\frac{1}{n_1+2-i}\right]\binom{n}{n_1}p_1^{n_1}(1-p_1)^{n-n_1}
\end{aligned} \tag{1.19}$$

by Shorack and Wellner (2009) where $p_1 = \mathbb{P}(D=1)$. The first term in (1.19) decays exponentially. In the second term of (1.19), the integrand in the square brackets is asymptotic to $\frac{\log n_1}{n_1+1} \leq \frac{\log n}{n_1+1}$. That is, there exist some constants $c, \bar{n}_1 > 0$ such that $\frac{1}{n_1+1}\sum_{i=1}^{n_1+1}\frac{1}{n_1+2-i} \leq c\frac{\log n}{n_1+1}$ if $n_1 > \bar{n}_1$. Since the integrand in the square brackets is bounded by one, it follows that the second term in (1.19) is bounded by

$$\sum_{n_1=1}^{\bar{n}_1}\binom{n}{n_1}p_1^{n_1}(1-p_1)^{n-n_1} + c(\log n)\mathbb{E}\left[(1+N_1)^{-1}\right],$$

where the first term is $O\left(n^{\bar{n}_1}(1-p_1)^n\right) = O\left(\log n/n\right)$ and the second term is $O\left(\log n/n\right)$ by Cribari-Neto et al. (2000). Similar arguments hold for $\mathbb{E}\max_{i:D_i=1}\underline{\Delta}_i$ by symmetry. ■

Proof of Proposition 1.2. We have for the R_{di} in (1.32) of Lemma 1.1,

$$\min_{i\in[n]} M_i = \min_{i\in[n]} N_{1-D_i}F_{1-D_i}[\pi(X_i) \pm \delta_n](1 + R_{1-D_i,i}). \tag{1.20}$$

If $\min_{i\in[n]}(1 + R_{1-D_i,i}) \geq 0$, which happens with probability tending to one, then (1.20) is less than or equal to

$$\left(1 \wedge \min_{d\in\{0,1\}}\inf_{p\in[\underline{p},\bar{p}]} f_d(p)\right)(N_0 \wedge N_1)\delta_n \min_{i\in[n]}(1 + R_{1-D_i,i}).$$

By Assumption 1.2, $\inf_{p\in[\underline{p},\bar{p}]} f_d(p) > 0$. By the strong law of large numbers and the continuous mapping theorem, $(1/\log n)(N_0 \wedge N_1)\delta_n = (N_0 \wedge N_1)/n \xrightarrow{a.s.} (1-p_1)\wedge p_1 > 0$. By Lemma 1.1, $\max_{i\in[n]} |R_{1-D_i,i}| = o_P(1)$. Then $\min_{i\in[n]} M_i \simeq (1 + o_P(1))\log n$, from which the lower bound in Proposition 1.2, and thus $\mathbb{P}\left(\min_{i\in[n]} M_i \geq 1\right) \to 1$, follows. The same reasoning applies to the upper bound in Proposition 1.2.

When the caliper is $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, $\min_{i\in[n]} M_i \geq 1$. Lemma 1.1(iv)–(vi), the continuous mapping theorem, combined with the law of large numbers and Proposition 1.1 prove the assertion. ■

Proof of Proposition 1.5. Follows from arguments proving Proposition 1.2 and Lemma 1.1 (iv)–(vi). When the caliper is $\delta_n = \widehat{\overline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, it is bounded by $\widehat{\overline{\Delta}}_n + \frac{\log N_0}{N_0+1} + \frac{\log N_1}{N_1+1}$, where $\mathbb{E}\frac{\log N_0}{N_0+1} = O\left(\frac{\log n}{n}\right)$ by Cribari-Neto et al. (2000). A spacings argument on $(\pi(X_i,\hat{\theta}))_{i\in[n]}$, similarly to the proof of Proposition 1.1, combined with Assumptions 1.6 and 1.9 yields $\widehat{\overline{\Delta}}_n = O_P\left(\frac{\log n}{n}\right)$. Then arguments in Proposition 1.2 and Lemma 1.1(iv)–(vi) prove Proposition 1.5. ■

Proof of Theorem 1.1. By Assumption 1.2, $\tau(\pi(X))$ of (1.5) is well-defined, satisfying $\mathbb{E}\tau(\pi(X)) = \tau$ by Assumption 1.1. By Assumption 1.5, the μ^d are Lipschitz continuous on the compact set $[\underline{p},\bar{p}]$, hence are bounded, and then so is $V_\tau < \infty$. Then the central limit theorem implies $\sqrt{n}(\overline{\tau(\pi(X))} - \tau) \rightsquigarrow \mathcal{N}(0, V_\tau)$. Combine this with Lemma 1.2, to get

$$\begin{bmatrix} V_\tau^{-1/2}\sqrt{n}(\overline{\tau(\pi(X))} - \tau) \\ V_E^{-1/2}\sqrt{n}E \end{bmatrix} \rightsquigarrow \mathcal{N}(0, \mathrm{I}_2),$$

along subsequences, where I_2 is the 2-by-2 identity matrix. By Lemma 1.4, $V_E \xrightarrow{P} V_{\sigma,\pi}$, which is finite by Assumptions 1.2 and 1.4. Then the continuous mapping theorem and Slutsky's lemma imply $(V_\tau + V_{\sigma,\pi})^{-1/2}\sqrt{n}(\overline{\tau(\pi(X))} - \tau + E) \rightsquigarrow \mathcal{N}(0,1)$. The event $\{\min_{i\in[n]} M_i > 0\}$ happens with probability tending to one by Proposition 1.2. On this event, $\sqrt{n}|B| \lesssim \sqrt{n}\delta_n = \frac{\log n}{\sqrt{n}} = o(1)$ by Assumption 1.5. Thus $\sqrt{n}B = o_P(1)$.

When the caliper is $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, Lemma 1.2 continues to apply. Then the continuous mapping theorem, the law of large numbers and Proposition 1.1 imply $\sqrt{n}|B| = o_P(1)$. ■

Proof of Theorem 1.2. A decomposition similar to (1.4)–(1.8) holds, whereby

$$\begin{aligned}
\sqrt{n}(\hat{\tau}_{\mathrm{t},\pi} - \tau_{\mathrm{t}}) &= \sqrt{n}(\overline{\tau_{\mathrm{t}}(\pi(X))} - \tau_{\mathrm{t}}) + \sqrt{n}E_{\mathrm{t}} + \sqrt{n}B_{\mathrm{t}}, \\
\overline{\tau_{\mathrm{t}}(\pi(X))} &\coloneqq \frac{1}{N_1}\sum_{i\in[n]} D_i \tau(\pi(X_i)), \\
E_{\mathrm{t}} &\coloneqq \frac{1}{N_1}\sum_{i\in[n]} E_{\mathrm{t},i}, \quad E_{\mathrm{t},i} \coloneqq (\mathbb{1}_{M_i>0} D_i - (1-D_i)w_i)\varepsilon_i, \\
B_{\mathrm{t}} &\coloneqq \frac{1}{N_1}\sum_{i\in[n]} B_{\mathrm{t},i}, \\
B_{\mathrm{t},i} &\coloneqq D_i(\mathbb{1}_{M_i>0} - 1)(\mu^{D_i}(\pi(X_i)) + \mu^{1-D_i}(\pi(X_i))) \\
&\quad + D_i \frac{\mathbb{1}_{M_i>0}}{M_i} \sum_{j\in\mathcal{J}(i)} (\mu^0(\pi(X_i)) - \mu^0(\pi(X_j))).
\end{aligned}$$

As $\mathbb{E}D(\tau(\pi(X)) - \tau_{\mathrm{t}}) = \mathbb{E}\left[\tau(\pi(X)) - \tau_{\mathrm{t}} \mid D = 1\right] p_1 = 0$, $\sqrt{n}(\overline{\tau_{\mathrm{t}}(\pi(X))} - \tau_{\mathrm{t}})$ is mean zero with finite variance by Assumptions 1.1 and 1.5. By the central limit theorem, continuous mapping theorem and Slutsky's lemma,

$$\sqrt{n}(\overline{\tau_{\mathrm{t}}(\pi(X))} - \tau_{\mathrm{t}}) = (N_1/n)^{-1} n^{-1/2} \sum_{i\in[n]} D_i(\tau(\pi(X_i)) - \tau_{\mathrm{t}}) \rightsquigarrow \mathcal{N}(0, V_{\tau_{\mathrm{t}}})$$

since $(N_1/n) \xrightarrow{a.s.} p_1$. The $\sqrt{n}E_{\mathrm{t}}$ has mean zero and a Lindeberg-Feller central limit theorem establishes that $\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(V_{E_{\mathrm{t}}}^{-1/2} \sqrt{n} E_{\mathrm{t}} \leq x \;\middle|\; D^{(n)}, PS^{(n)} \right) - \Phi(x) \right| \xrightarrow{P} 0$, where $V_{E_{\mathrm{t}}} \coloneqq \frac{1}{n}\sum_{i\in[n]} (\mathbb{1}_{M_i>0} D_i - (1-D_i)w_i)^2 \sigma^2_{D_i}(\pi(X_i))$, similarly to arguments in Lemma 1.2. By arguments similar to those of Lemma 1.4, $V_{E_{\mathrm{t}}} \xrightarrow{P} V_{\mathrm{t},\sigma,\pi}$. By Proposition 1.2 and arguments in Theorem 1.1, $\sqrt{n}B_{\mathrm{t}} = o_P(1)$. ∎

Proof of Theorem 1.3. Let $w_i(\hat{\theta}) := \sum_{j \in \mathcal{J}_{\hat{\theta}}(i)} \frac{1}{M_j(\hat{\theta})}, i \in [n]$, and, as in (1.4)–(1.8), decompose

$$\sqrt{n}(\hat{\tau}_{\hat{\pi}} - \tau) = \frac{1}{\sqrt{n}} \sum_{i \in [n]} \left\{ \mu^1(\hat{\theta}, \pi(X_i, \hat{\theta})) - \mu^0(\hat{\theta}, \pi(X_i, \hat{\theta})) - \tau \right\} \tag{1.21}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i \in [n]} (2D_i - 1)(\mathbb{1}_{M_i(\hat{\theta})>0} + w_i(\hat{\theta}))\varepsilon_i(\hat{\theta}) \tag{1.22}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i \in [n]} (2D_i - 1)(\mathbb{1}_{M_i(\hat{\theta})>0} - 1)(\mu^{1-D_i}(\hat{\theta}, \pi(X_i, \hat{\theta})) - \mu^{D_i}(\hat{\theta}, \pi(X_i, \hat{\theta}))) \tag{1.23}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i \in [n]} (2D_i - 1) \frac{\mathbb{1}_{M_i(\hat{\theta})>0}}{M_i(\hat{\theta})} \sum_{j \in \mathcal{J}_{\hat{\theta}}(i)} \left[\mu^{1-D_i}(\hat{\theta}, \pi(X_i, \hat{\theta})) - \mu^{1-D_i}(\hat{\theta}, \pi(X_j, \hat{\theta})) \right]. \tag{1.24}$$

We show first that (1.21) and (1.22) are, asymptotically, jointly normal and independent, and then that (1.23) and (1.24) are asymptotically negligible.

Terms (1.21) *and* (1.22). We apply the following result with (1.21) and (1.22) corresponding to V_n and W_n respectively. Let V_n, W_n, $n = 1, 2, \ldots$ be two sequences of random variables defined on some probability space. To show that $(V_n, W_n) \rightsquigarrow (V, W) \sim \mathcal{N}(0, \Sigma)$, for a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_V^2, \sigma_W^2)$, it suffices that $\mathbb{E}h_1(V_n)h_2(W_n) \to (\mathbb{E}h_1(V))(\mathbb{E}h_2(W))$ for all bounded continuous functions h_1, $h_2 : \mathbb{R} \to \mathbb{R}$. Let $\mathcal{F}_{n0}$ be a sub-σ-algebra such that V_n is $\mathcal{F}_{n0}$-measurable for all $n \geq 1$. As $\mathbb{E}h_1(V_n)h_2(W_n) = \mathbb{E}\left[h_1(V_n)\mathbb{E}\left[h_2(W_n) \mid \mathcal{F}_{n0}\right]\right]$, it suffices, by the Portmanteau lemma, that $V_n \rightsquigarrow \mathcal{N}(0, \sigma_V^2)$ and $\mathbb{P}\left(W_n \leq w \mid \mathcal{F}_{n0}\right) \xrightarrow{P} \Phi(w/\sigma_W)$ for all $w \in \mathbb{R}$.

Convergence of (1.21). Expand (1.21) as

$$\frac{1}{\sqrt{n}} \sum_{i \in [n]} (\mu^1_{\theta_0}(\pi(X_i, \theta_0)) - \mu^0_{\theta_0}(\pi(X_i, \theta_0)) - \tau) \tag{1.25}$$

$$+ \frac{1}{\sqrt{n}} \sum_{i \in [n]} \Big\{ \mu^1(\hat{\theta}, \pi(X_i, \hat{\theta})) - \mu^0(\hat{\theta}, \pi(X_i, \hat{\theta})) - \left[\mu^1_{\theta_0}(\pi(X_i, \theta_0)) - \mu^0_{\theta_0}(\pi(X_i, \theta_0))\right] \Big\}. \tag{1.26}$$

By Assumptions 1.1 and 1.5, (1.25) converges weakly to $\mathcal{N}(0, V_\tau)$ by the standard central limit theorem. By Assumptions 1.3 and 1.7, $\Lambda^d(\tilde{\theta}, x)$ is well-defined for all $(\tilde{\theta}, x) \in \text{Nb}(\theta_0, \epsilon) \times \mathcal{X}$. Then by the mean-value theorem,

$$\mu^d(\hat{\theta}, \pi(x, \hat{\theta})) = \mu^d_{\theta_0}(\pi(x, \theta_0)) + \Lambda^d(\tilde{\theta}^d, x)(\hat{\theta} - \theta_0),$$

for some $\tilde{\theta}^d$ on the line segment between $\hat{\theta}$ and θ_0. Rewrite (1.26) as

$$\left(\frac{1}{n} \sum_{i \in [n]} \left[\Lambda^1(\tilde{\theta}^1, X_i) - \Lambda^0(\tilde{\theta}^0, X_i) \right] \right) \sqrt{n}(\hat{\theta} - \theta_0). \tag{1.27}$$

By Assumptions 1.3 and 1.7, $\tilde{\theta} \mapsto \Lambda^d(\tilde{\theta}, x)$ is continuous and uniformly bounded for all $(\tilde{\theta}, x) \in \text{Nb}(\theta_0, \epsilon) \times \mathcal{X}$, therefore $\frac{1}{n} \sum_{i \in [n]} \Lambda^d(\tilde{\theta}^d, X_i) \xrightarrow{P} q_d^\intercal$ for some finite $q_d \in \mathbb{R}^K$. Then (1.27) and Slutsky's lemma imply that (1.26) converges weakly to $\mathcal{N}(0, (q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0))$ by Assumption 1.9. But, by Assumption 1.9 again, (1.25) is independent of $\sqrt{n}(\hat{\theta} - \theta_0)$, thus (1.21), being the sum of (1.25) and (1.26), converges weakly to $\mathcal{N}(0, V_\tau + (q_1 - q_0)^\intercal V_{\theta_0}(q_1 - q_0))$.

Conditional Convergence of (1.22). Let

$$\mathcal{F}_{n0} \coloneqq \sigma\{D_1, \ldots, D_n, \pi(X_1, \hat{\theta}), \ldots, \pi(X_n, \hat{\theta}), \hat{\theta}\},$$

so that (1.21) is $\mathcal{F}_{n0}$-measurable. We show that given $\mathcal{F}_{n0}$, (1.22) converges weakly to a normal variate in probability. We construct a martingale array and apply Lemma 1.7. Let $\xi_{ni} \coloneqq (2D_i - 1)(1 + w_i(\hat{\theta}))\varepsilon_i(\hat{\theta})/\sqrt{n}$ and

$$\mathcal{F}_{ni} \coloneqq \sigma\{D_1, \ldots, D_n, \pi(X_1, \hat{\theta}), \ldots, \pi(X_n, \hat{\theta}), \varepsilon_1(\hat{\theta}), \ldots, \varepsilon_i(\hat{\theta}), \hat{\theta}\}$$

for $i \in [n]$. Assume temporarily that $\min_{i \in [n]} M_i(\hat{\theta}) > 0$, so that (1.22) is equal to $\sum_{i=1}^n \xi_{ni}$. One can verify that $\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn}$ are martingale differences relative to the filtration $\mathcal{F}_{n1} \subset \mathcal{F}_{n2} \subset \ldots \subset \mathcal{F}_{nn}$, using that $\mu^{D_i}(\hat{\theta}, \pi(X_i, \hat{\theta}))$ is $\mathcal{F}_{n,i-1}$-measurable and Assumption 1.9(ii), which implies that the observations are i.i.d. given $\hat{\theta}$.

First, we verify the variance condition (1.56) of Lemma 1.7. Consider the sum

$$\sum_{i=1}^n \mathbb{E}\left[\xi_{ni}^2 \,\middle|\, \mathcal{F}_{n,i-1} \right].$$

We have

$$\mathbb{E}\left[\xi_{ni}^2 \,\middle|\, \mathcal{F}_{n,i-1}\right] = (1+w_i(\hat{\theta}))^2\sigma_{D_i}^2(\hat{\theta},\pi(X_i,\hat{\theta}))/n,$$

where we used Assumption 1.9(ii) again, and that $w_i(\hat{\theta})$ is $\mathcal{F}_{n,i-1}$-measurable. But then $\mathbb{E}\left[\xi_{ni}^2 \,\middle|\, \mathcal{F}_{n,i-1}\right]$ is $\mathcal{F}_{n0}$-measurable for all $i \in [n]$ for all $n \geq 1$, thus for condition (1.56) it suffices that $\sum_{i=1}^n \mathbb{E}\left[\xi_{ni}^2 \,\middle|\, \mathcal{F}_{n,i-1}\right]$ converges in $\mathbb{P}$-probability to a finite constant. Assumption 1.3(iii) implies that

$$\max_{i\in[n]} |\pi(X_i,\hat{\theta}) - \pi(X_i,\theta_0)| = O_P\left(\|\hat{\theta}-\theta_0\|\right).$$

Then, under Assumption 1.8, we can write $\sigma_d^2(\hat{\theta},\pi(X_i,\hat{\theta})) = \sigma_d^2(\theta_0,\pi(X_i,\theta_0)) + S_i$, where $\max_{i\in[n]}|S_i| = O_P\left(\|\hat{\theta}-\theta_0\|\right)$. Write

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[\xi_{ni}^2 \,\middle|\, \mathcal{F}_{n,i-1}\right] &= \frac{1}{n}\sum_{i\in[n]}(1+w_i(\hat{\theta}))^2\sigma_{D_i}^2(\theta_0,\pi(X_i,\theta_0)) \\ &\quad + \frac{1}{n}\sum_{i\in[n]}(1+w_i(\hat{\theta}))^2 S_i.\end{aligned} \tag{1.28}$$

Under Assumptions 1.3, 1.6 and 1.9, the arguments in the proof of Lemma 1.4 continue to apply in view of Lemma 1.1(iv)–(vi). Then for any sequence of random variables $(Q_i)_{i\in[n]}$ and any constant $r \in \mathbb{R}$, we have, by Lemma 1.1(iv)–(vi),

$$\begin{aligned}\sum_{i\in[n]} w_i(\hat{\theta})^r Q_i = (1+o_P(1))\Bigg\{&\left(\frac{N_1}{N_0}\right)^r \sum_{i:D_i=0}\left(\frac{f_{1,\hat{\theta}}(\pi(X_i,\hat{\theta}))}{f_{0,\hat{\theta}}(\pi(X_i,\hat{\theta}))}\right)^r Q_i \\ &+ \left(\frac{N_0}{N_1}\right)^r \sum_{i:D_i=1}\left(\frac{f_{0,\hat{\theta}}(\pi(X_i,\hat{\theta}))}{f_{1,\hat{\theta}}(\pi(X_i,\hat{\theta}))}\right)^r Q_i\Bigg\}.\end{aligned} \tag{1.29}$$

The second term in (1.28) is bounded by

$$O_P\left(\|\hat{\theta}-\theta_0\|\right)\left(1+\frac{1}{n}\sum_{i\in[n]} w_i(\hat{\theta})^2\right).$$

Assumption 1.6, bounding the ratios $f_{d,\theta}(p)/f_{1-d,\theta}(p)$ uniformly in $p \in [\underline{p}_\theta, \bar{p}_\theta]$ and $\theta \in \mathrm{Nb}(\theta_0,\epsilon)$, combined with (1.29) and $(N_d/N_{1-d})^r \xrightarrow{a.s.} \left(\frac{p_d}{1-p_d}\right)^r$, where $p_0 := 1-p_1$, bounds $\frac{1}{n}\sum_{i\in[n]} w_i(\hat{\theta})^2$ by a $(1+o_P(1))$-term up to a constant

factor. Thus, the second term in (1.28) is $O_P\left(\|\hat{\theta}-\theta_0\|\right) = o_P(1)$ under Assumption 1.9. Put $Q_i \coloneqq \sigma^2_{D_i}(\theta_0, \pi_i)$ and apply a mean-value expansion in θ around θ_0 to the right side of (1.29). This is feasible under Assumptions 1.3 and 1.6, which also bound the derivative uniformly in $(x, \tilde{\theta}) \in \mathcal{X} \times \mathrm{Nb}(\theta_0, \epsilon)$. Then Assumptions 1.8 and 1.9 imply that the first term in (1.28) is $V_{\sigma,\pi} + o_P(1)$ under Assumption 1.3, as in the proof of Lemma 1.4, where $V_{\sigma,\pi}$ is finite by Assumptions 1.6 and 1.8. Conclude that (1.56) holds.

Second, we verify the Lindeberg-condition (1.57) of Lemma 1.7. We need to show

$$\sum_{i=1}^{n} \mathbb{E}\left[\xi_{ni}^2 \mathbb{1}_{|\xi_{ni}|\geq\eta} \,\middle|\, \mathcal{F}_{n0}\right] \xrightarrow{P} 0 \quad \text{for each } \eta > 0.$$

As $\mathbb{1}_{|\xi_{ni}|\geq\eta}$ is bounded by ξ_{ni}^2/η^2,

$$\sum_{i=1}^{n} \mathbb{E}\left[\xi_{ni}^2 \mathbb{1}_{|\xi_{ni}|\geq\eta} \,\middle|\, \mathcal{F}_{n0}\right]$$
$$\leq \sum_{t=1}^{n} \frac{\mathbb{E}\left[\xi_{ni}^4 \,\middle|\, \mathcal{F}_{n0}\right]}{\eta^2} = \frac{1}{n\eta^2} \sum_{i\in[n]} \frac{(1+\tilde{w}_i(\hat{\theta}))^4 \mathbb{E}\left[\varepsilon_i(\hat{\theta})^4 \,\middle|\, \mathcal{F}_{n0}\right]}{n},$$

with $\mathbb{E}\left[\varepsilon_i(\hat{\theta})^4 \,\middle|\, \mathcal{F}_{n0}\right] = \sigma^4_{D_i}(\hat{\theta}, \pi(X_i, \hat{\theta}))$ by Assumption 1.9(ii), which is bounded uniformly in $i \in [n]$ by Assumption 1.8. In view of (1.29), $\frac{1}{n^2}\sum_{i\in[n]}(1+\tilde{w}_i(\hat{\theta}))^4 = o_P(1)$, so (1.57) is met.

Conclude that under the temporary assumption $\min_{i\in[n]} M_i(\hat{\theta}) > 0$, Lemma 1.7 applies, so (1.22) converges weakly to $\mathcal{N}(0, V_{\sigma,\pi})$ in probability. To remove this assumption, define the $\mathcal{F}_{n0}$-measurable set $A_n \coloneqq \{\min_{i\in[n]} M_i(\hat{\theta}) > 0\}$. On A_n, (1.22) is equal to $\sum_{i\in[n]} \xi_{ni}$. As $\mathbb{P}(A_n) \to 1$ by Proposition 1.5, the desired convergence follows.

Vanishing (1.23). By Assumption 1.7, the $p \mapsto \mu^d(\theta, p)$ are continuous on a compact set $[\underline{p}_\theta, \bar{p}_\theta]$ for all $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$. By Assumption 1.9, $\mathbb{P}\left(\hat{\theta} \in \mathrm{Nb}(\theta_0, \epsilon)\right) \to 1$. By Proposition 1.5, $\mathbb{P}(A_n) \to 1$, thus (1.23) is $o_P(1)$.

Vanishing (1.24). By Assumption 1.7, $|\mu^d(\theta,p') - \mu^d(\theta,p)| \leq L|p'-p|$ for all $p', p \in [\underline{p}_\theta, \bar{p}_\theta]$ for all $\theta \in \text{Nb}(\theta_0, \epsilon)$. Then

$$\max_{i\in[n]} \max_{j\in\mathcal{J}_{\hat{\theta}}(i)} |\mu^{1-D_i}(\hat{\theta}, \pi(X_i,\hat{\theta})) - \mu^{1-D_i}(\hat{\theta}, \pi(X_j,\hat{\theta}))|$$
$$\lesssim \max_{i\in[n]} \max_{j\in\mathcal{J}_{\hat{\theta}}(i)} |\pi(X_i,\hat{\theta}) - \pi(X_j,\hat{\theta})|.$$

By the construction of $\mathcal{J}_{\hat{\theta}}(i)$, the right side is bounded by δ_n, where $\sqrt{n}\delta_n \to 0$.

When the caliper is $\delta_n = \widehat{\underline{\Delta}}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, the above arguments continue to hold. Specifically, in showing the conditional convergence of (1.22), $w_i(\hat{\theta})$ is still $\mathcal{F}_{n,i-1}$-measurable; (1.23) is exactly zero, and, in (1.24), $\sqrt{n}\delta_n \leq \sqrt{n}\left(\widehat{\underline{\Delta}}_n + \frac{\log n}{N_0+1} + \frac{\log n}{N_1+1}\right) = o_P(1)$ in view of the proof of Proposition 1.5. ■

Proof of Theorem 1.4. Follows those of Theorem 1.2 and Theorem 1.3. ■

Proof of Proposition 1.6. Assumptions 1.1, 1.3 and 1.9(ii) hold by construction as $\mathcal{X}$ is bounded. The assumptions of Example 1.1 imply Assumption 1.9(i) by standard asymptotic theory (Van der Vaart, 1998). We assume $K = 2$ covariates in the following so that $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$; the general case $K \geq 2$ follows analogously. First, we establish some general results. Let θ_k ($\theta_{0,k}$) denote the kth entry of θ (θ_0). For $t \in \mathcal{T} := \{\theta^\intercal x : \theta \in \Theta, x \in \mathcal{X}\}$, $\theta^\intercal X$ has density $f_{\theta^\intercal X}(t) = \int_{\mathcal{X}_1} \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1$, which is strictly positive by assumptions on Ψ, the density of X.[7] Let $h : \mathcal{X} \to \mathbb{R}^J, J \geq 1$, be an arbitrary integrable function. We have

$$\mathbb{E}\left[h(X_1,X_2) \mid \theta^\intercal X = t\right] = \frac{1}{f_{\theta^\intercal X}(t)} \int_{\mathcal{X}_1} h\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1. \tag{1.30}$$

Combine this with the tower property of expectations to get the conditional density

$$f_{\theta^\intercal X|D}(t \mid d) = \begin{cases} \frac{1}{1-p_1} \int_{\mathcal{X}_1} (1 - g_{\theta_0}(x_1,\theta,t)) \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1 & \text{if } d = 0, \\ \frac{1}{p_1} \int_{\mathcal{X}_1} g_{\theta_0}(x_1,\theta,t) \Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1 & \text{if } d = 1, \end{cases}$$

[7] If X included an intercept, then we would have $f_{\theta^\intercal X}(t) = \int_{\mathcal{X}_1} \Psi\left(x_1, \frac{t-\theta_3-\theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1$, where θ_3 is the coefficient on the intercept. Below, the right side of (1.30), $f_{\theta^\intercal X|D}(t \mid D)$ and (1.31) would need to be adjusted in a similar manner to accommodate an intercept.

where $g_{\theta_0}(x_1, \theta, t) \coloneqq g\left(\theta_{0,1}x_1 + \theta_{0,2}\frac{t-\theta_1 x_1}{\theta_2}\right) \in (0,1)$ as $0 < g(t') < 1$ for all t' in the bounded $\mathcal{T}$. Hence, $f_{\theta^\intercal X|D}$ is strictly positive. It is also continuously differentiable in t by assumptions on g and Ψ. Moreover, $\theta \mapsto f_{\theta^\intercal X|D}(t \mid d)$ is continuously differentiable at $\tilde{\theta} \in \mathrm{Nb}(\theta_0, \epsilon)$ with bounded derivative for a $\theta_{0,2} \neq 0$ as $\sup_{t\in\mathbb{R}} g'(t) < \infty$ and the derivatives of Ψ are bounded. One can also show that $\mathbb{E}\left[h(X_1, X_2) \mid D = d, \theta^\intercal X = t\right]$ is

$$\begin{aligned}
&\frac{1-p_1}{f_{\theta^\intercal X|D}(t \mid 0)} \int_{\mathcal{X}_1} h\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right)(1 - g_{\theta_0}(x_1, \theta, t))\,\Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right)\mathrm{d}x_1, \\
&\frac{p_1}{f_{\theta^\intercal X|D}(t \mid 1)} \int_{\mathcal{X}_1} h\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right) g_{\theta_0}(x_1, \theta, t)\Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right)\mathrm{d}x_1, \quad (1.31)
\end{aligned}$$

respectively for $d = 0$ and $d = 1$, which is continuously differentiable in t and θ by assumptions on g, Ψ, and the properties of $f_{\theta^\intercal X|D}$ derived above, provided h is continuously differentiable. We are now ready to verify the remaining assumptions.

Assumption 1.6. The distributions are

$$F_{d,\theta}(p) = \begin{cases} \frac{1}{1-p_1}\int_{\mathcal{X}} \mathbb{1}_{g(\theta^\intercal x)\leq p}(1 - g(\theta_0^\intercal x))\Psi(x)\mathrm{d}x & \text{if } d = 0, \\ \frac{1}{p_1}\int_{\mathcal{X}} \mathbb{1}_{g(\theta^\intercal x)\leq p} g(\theta_0^\intercal x)\Psi(x)\mathrm{d}x & \text{if } d = 1. \end{cases}$$

Since g is increasing and $\mathcal{T}_\theta \coloneqq \{\theta^\intercal x : x \in \mathcal{X}\}$ is compact, for all $\theta \in \Theta$ there exist $0 < \underline{p}_\theta < \bar{p}_\theta < 1$ such that $F_{d,\theta}(\underline{p}_\theta) = 0$ and $F_{d,\theta}(\bar{p}_\theta) = 1$. Specifically, $\underline{p}_\theta = g(\inf \mathcal{T}_\theta)$ and $\bar{p}_\theta = g(\sup \mathcal{T}_\theta)$. On $[\underline{p}_\theta, \bar{p}_\theta]$, the $F_{d,\theta}$ admit densities

$$f_{d,\theta}(p) = f_{\theta^\intercal X|D}(g^{-1}(p) \mid d)(g^{-1})'(p) = \frac{f_{\theta^\intercal X|D}(g^{-1}(p) \mid d)}{g'(g^{-1}(p))},$$

where g^{-1} is well-defined, as well as its derivative by the inverse function theorem. Then the assumptions on g and the properties of $f_{\theta^\intercal X|D}$ imply that Assumption 1.6 holds.

Assumption 1.7. Under Assumption 1.1, the tower property of expectation gives

$$\mu^d(\theta, p) = \mathbb{E}\left[Y \mid D = d, \theta^\intercal X = g^{-1}(p)\right] = \mathbb{E}\left[m_d(X) \mid D = d, \theta^\intercal X = g^{-1}(p)\right].$$

But then the properties of (1.31), g and m_d imply that Assumption 1.7 holds.

Assumption 1.8. The lower bound in (i) is satisfied by assumptions on ν_d in Example 1.1. The Lipschitz condition in (i) and (ii) both follow by the mean-value theorem as the σ_d^2 and σ_d^4 are polynomials in terms of the form $\mathbb{E}\left[h(X_1, X_2) \mid D = d, \theta^\intercal X = g^{-1}(p)\right]$ for continuously differentiable functions h by assumptions on m_d and ν_d, so (1.31) applies. ■

Proof of Proposition 1.4 . Follows that of Proposition 1.6. ■

Lemma 1.1 (Convergence of Ratios)**.** *Suppose that the caliper δ_n satisfies* (1.3)*. For the measures in* (1.17)*, define*

$$R_{di} \coloneqq \frac{\mathbb{F}_{N_d}[\pi(X_i) \pm \delta_n]}{F_d[\pi(X_i) \pm \delta_n]} - 1, \tag{1.32}$$

$$\tilde{R}_{di} \coloneqq \frac{F_d[\pi(X_i) \pm \delta_n]}{2\delta_n f_d(\pi(X_i))} - 1, \tag{1.33}$$

$$\check{R}_{dji} \coloneqq \frac{f_d(\pi(X_j))}{f_d(\pi(X_i))} - 1 \tag{1.34}$$

for $j \in \mathcal{J}(i)$, $i \in \{i \in [n] : D_i = 1 - d\}$ and $d \in \{0, 1\}$. If Assumption 1.2 holds, then

(i) $\max_{i:D_i=1-d} |R_{di}| = o_P(1)$*; and*

(ii) $\max_{i:D_i=1-d} |\tilde{R}_{di}| = o_P(1)$*; and*

(iii) $\max_{i:D_i=1-d} \max_{j\in\mathcal{J}(i)} |\check{R}_{dji}| = o_P(1)$

as $n \to \infty$ for all $d \in \{0, 1\}$. Suppose that Assumption 1.3 holds and the caliper δ_n satisfies (1.13)*. For the measures in* (1.18)*, define*

$$R_{di}^\theta \coloneqq \frac{\mathbb{F}_{N_d,\theta}[\pi(X_i, \theta) \pm \delta_n]}{F_{d,\theta}[\pi(X_i, \theta) \pm \delta_n]} - 1, \tag{1.35}$$

$$\tilde{R}_{di}^\theta \coloneqq \frac{F_{d,\theta}[\pi(X_i, \theta) \pm \delta_n]}{2\delta_n f_{d,\theta}(\pi(X_i, \theta))} - 1, \tag{1.36}$$

$$\check{R}_{dji}^\theta \coloneqq \frac{f_{d,\theta}(\pi(X_j, \theta))}{f_{d,\theta}(\pi(X_i, \theta))} - 1 \tag{1.37}$$

for $j \in \mathcal{J}_\theta(i)$, $i \in \{i \in [n] : D_i = 1 - d\}$, $d \in \{0, 1\}$ and $\theta \in \mathrm{Nb}(\theta_0, \epsilon)$. If Assumption 1.6 holds, then for any $\hat{\theta}$ satisfying Assumption 1.9,

(iv) $\max_{i:D_i=1-d} |R_{di}^{\hat{\theta}}| = o_P(1)$*; and*

(v) $\max_{i:D_i=1-d} |\tilde{R}_{di}^{\hat{\theta}}| = o_P(1)$*; and*

(vi) $\max_{i:D_i=1-d} \max_{j\in\mathcal{J}_{\hat{\theta}}(i)} |\check{R}^{\hat{\theta}}_{dji}| = o_P(1)$

as $n \to \infty$ *for all* $d \in \{0,1\}$.

Proof. Assertion (i). Consider

$$|R_{1i}| \leq \sup_{p\in[\underline{p},\bar{p}]} \left| \frac{\mathbb{F}_{N_1}[p \pm \delta_n]}{F_1[p \pm \delta_n]} - 1 \right| = \sup_{p\in[\underline{p},\bar{p}]} \frac{|\mathbb{G}_{N_1}[p \pm \delta_n]|}{\sqrt{N_1} F_1[p \pm \delta_n]} =: W_1. \tag{1.38}$$

Fix a constant $\zeta > 0$. We bound $\mathbb{P}\left(W_1 > \zeta \mid D^{(n)}\right)$ using ratio and tail bounds of empirical processes. To this end, note that for any finite $\delta_n > 0$, $\mathcal{C}_{\delta_n} := \{[p \pm \delta_n] : p \in [\underline{p}, \bar{p}]\}$ is a VC-class, with VC-dimension equal to two (e.g. Van der Vaart and Wellner (1996, Example 2.6.1)). Let $\gamma_n := \delta_n(1 \wedge \inf_{p\in[\underline{p},\bar{p}]} f_1(p))$, so that $\inf_{p\in[\underline{p},\bar{p}]} F_1[p \pm \delta_n] \geq \gamma_n$. The event $\{W_1 > \zeta\}$ is equal to

$$\left\{ \sup \left\{ \frac{|\mathbb{G}_{N_1}[p \pm \delta_n]|}{F_1[p \pm \delta_n]} : p \in [\underline{p}, \bar{p}], F_1[p \pm \delta_n] \geq \gamma_n \right\} > \sqrt{N_1}\zeta \right\} \subset (\mathcal{E}_1 \cup \mathcal{E}_2),$$

$$\mathcal{E}_1 := \left\{ \sup \left\{ \frac{|\mathbb{G}_{N_1}[p \pm \delta_n]|}{F_1[p \pm \delta_n]} : p \in [\underline{p}, \bar{p}], F_1[p \pm \delta_n] \geq \gamma_n, F_1[p \pm \delta_n] \leq \frac{1}{2} \right\} > \sqrt{N_1}\zeta \right\}, \tag{1.39}$$

$$\mathcal{E}_2 := \left\{ \sup \left\{ \frac{|\mathbb{G}_{N_1}[p \pm \delta_n]|}{F_1[p \pm \delta_n]} : p \in [\underline{p}, \bar{p}], F_1[p \pm \delta_n] \geq \gamma_n, F_1[p \pm \delta_n] > \frac{1}{2} \right\} > \sqrt{N_1}\zeta \right\}. \tag{1.40}$$

First, we bound the probability of the event in (1.39). Take $A := \{p \in [\underline{p}, \bar{p}] : F_1[p \pm \delta_n] \geq \gamma_n\}$, $B := \{p \in [\underline{p}, \bar{p}] : F_1[p \pm \delta_n] \leq \frac{1}{2}\}$. If $\gamma_n > \frac{1}{2}$, $A \cap B$ is empty and by the convention $\sup \emptyset = -\infty$, the set (1.39) has measure zero. So assume without loss of generality that $\gamma_n \leq F_1[p \pm \delta_n] \leq \frac{1}{2}$. Thus, $\frac{\gamma_n}{2} \leq \sigma_1^2[p \pm \delta_n] := (F_1[p \pm \delta_n])(1 - F_1[p \pm \delta_n]) \leq \frac{1}{4}$, where note that $\sigma_1^2[p \pm \delta_n] < F_1[p \pm \delta_n]$ for $F_1[p \pm \delta_n] \geq \gamma_n > 0$. As $N_1\gamma_n \xrightarrow{a.s.} \infty$ and $N_1^{-1}\log(e \vee \log(e \vee N_1)) = o(\gamma_n)$ a.s., conditions (2.2)–(2.3) in Alexander (1987, Theorem 2.1) hold a.s.. Therefore, the bounds in the proof of Theorem 5.1 therein apply. Hence, for

$$\bar{\mathcal{E}}_1 := \left\{ |\mathbb{G}_{N_1}[p \pm \delta_n]| > (\sigma_1^2[p \pm \delta_n])\sqrt{N_1}\zeta \text{ for some } p \in [\underline{p}, \bar{p}] : \frac{\gamma_n}{2} \leq \sigma_1^2[p \pm \delta_n] \leq \frac{1}{4} \right\}$$

we have, as in (7.66)–(7.67) of Alexander (1987),

$$\mathbb{P}\left(\mathcal{E}_1 \mid D^{(n)}\right) \leq \mathbb{P}\left(\bar{\mathcal{E}}_1 \mid D^{(n)}\right) \leq 36\int_{\gamma_n/2}^{1/4} t^{-1}e^{-\zeta^2 N_1 t/512}\mathrm{d}t + 68e^{-\zeta N_1\gamma_n/256}$$
$$\leq \frac{36}{\zeta^2 N_1\gamma_n}e^{-\zeta^2 N_1\gamma_n/1024} + 68e^{-\zeta N_1\gamma_n/256}.$$

Second, the probability of the event in (1.40) is bounded by

$$\begin{aligned}
&\mathbb{P}\left(\sup\left\{\frac{|\mathbb{G}_{N_1}[p\pm\delta_n]|}{F_1[p\pm\delta_n]} : p\in[\underline{p},\bar{p}], F_1[p\pm\delta_n] > \frac{1}{2}\right\} > \sqrt{N_1}\zeta \,\middle|\, D^{(n)}\right)\\
&\leq \mathbb{P}\left(\sup\left\{|\mathbb{G}_{N_1}[p\pm\delta_n]| : p\in[\underline{p},\bar{p}], F_1[p\pm\delta_n] > \frac{1}{2}\right\} > \frac{\sqrt{N_1}}{2}\zeta \,\middle|\, D^{(n)}\right)\\
&\leq \mathbb{P}\left(\sup_{p\in[\underline{p},\bar{p}]}|\mathbb{G}_{N_1}[p\pm\delta_n]| > \frac{\sqrt{N_1}}{2}\zeta \,\middle|\, D^{(n)}\right)
\end{aligned} \tag{1.41}$$

as the supremum over a larger set cannot decrease. Van der Vaart and Wellner (1996, Theorem 2.14.9) bound (1.41) by $cN_1\zeta^2 e^{-\frac{N_1\zeta^2}{2}}$. Therefore,

$$\mathbb{P}\left(W_1 > \zeta \mid D^{(n)}\right) \leq \frac{c_1}{\zeta^2 N_1\gamma_n}e^{-\zeta^2 N_1\gamma_n/1024} + c_2 e^{-\zeta N_1\gamma_n/256} + c_3 N_1\zeta^2 e^{-N_1\zeta^2/2}, \tag{1.42}$$

on the set where $N_0, N_1 \geq 1$, which happens with probability tending to one. The left side is bounded by one, the right side converges to zero in probability; then the left side also converges to zero in expectation. Similar arguments hold for

$$W_0 \coloneqq \sup_{p\in[\underline{p},\bar{p}]}\frac{|\mathbb{G}_{N_0}|[p\pm\delta_n]|}{\sqrt{N_0}F_0[p\pm\delta_n]},$$

bounding $\max_{i:D_i=1}|R_{0i}|$. Conclude that $\max_{i:D_i=1-d}|R_{di}| = o_P(1)$. When the caliper is $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, the same arguments yield the assertion by letting $\gamma_n \coloneqq \gamma_{N_d} \coloneqq (1\wedge\inf_{p\in[\underline{p},\bar{p}]} f_d(p))\frac{\log N_d}{N_d+1}$ when bounding W_d, $d\in\{0,1\}$.

Assertion (ii). By Assumption 1.2, the f_d are continuous on the compact set $[\underline{p},\bar{p}]$, hence are uniformly continuous. By the mean-value theorem, uniform continuity of f_d implies uniform differentiability of F_d. By uniform differentiability of F_0,

$$\sup_{j:D_j=1}|F_0[\pi(X_j)\pm\delta_n] - 2\delta_n f_0(\pi(X_j))| = o_P(\delta_n)$$

for $\delta_n = o_P(1)$. To see this, fix a constant $\zeta > 0$. By uniform differentiability of F_0, for all ζ there exists a constant $\bar{\delta} > 0$ such that

$$\sup_{p \in [\underline{p}, \bar{p}]} \frac{|F_0(p + \delta_n) - F_0(p) - \delta_n f_0(p)|}{\delta_n} \leq \zeta$$

whenever $\delta_n \leq \bar{\delta}$. The event $\{\sup_{p \in [\underline{p}, \bar{p}]} \frac{|F_0(p+\delta_n) - F_0(p) - \delta_n f_0(p)|}{\delta_n} > \zeta\}$ is equal to

$$\left\{ \sup_{p \in [\underline{p}, \bar{p}]} \frac{|F_0(p + \delta_n) - F_0(p) - \delta_n f_0(p)|}{\delta_n} > \zeta, \delta_n \leq \bar{\delta} \right\}$$
$$\cup \left\{ \sup_{p \in [\underline{p}, \bar{p}]} \frac{|F_0(p + \delta_n) - F_0(p) - \delta_n f_0(p)|}{\delta_n} > \zeta, \delta_n > \bar{\delta} \right\}.$$

The first event has measure zero by uniform differentiability. The probability of the second event is dominated by $\mathbb{P}\left(\delta_n > \bar{\delta}\right)$, which is $o(1)$. Then the statement follows by noting that

$$F_0[p \pm \delta_n] = F_0(p + \delta_n) - F_0(p) + F_0(p) - F_0(p - \delta_n) \text{ and}$$
$$\max_{i \in [n]} \max_{j \in \mathcal{J}(i)} |f_0(\pi(X_i)) - f_0(\pi(X_j))| = o_P(1)$$

(see the proof of Assertion (iii)), so $2f_0(p) = f_0(p) + f_0(p - \delta_n) + f_0(p) - f_0(p - \delta_n) = f_0(p) + f_0(p - \delta_n) + o_P(1)$. As $\inf_{p \in [\underline{p}, \bar{p}]} f_0(p) > 0$ by Assumption 1.2, Assertion (ii) follows. When the caliper is $\delta_n = \overline{\underline{\Delta}}_n \vee \frac{\log N_0}{N_0 + 1} \vee \frac{\log N_1}{N_1 + 1}$, Proposition 1.1, the continuous mapping theorem and the law of large numbers imply $\delta_n = o_P(1)$, whence the assertion follows by the above arguments.

Assertion (iii). As f_0 is uniformly continuous by Assumption 1.2,

$$\max_{i \in [n]} \max_{j \in \mathcal{J}(i)} |f_0(\pi(X_j)) - f_0(\pi(X_i))| = o_P(1).$$

To see this, fix a constant $\zeta > 0$. By uniform continuity of f_0, for all ζ there exists an $\eta > 0$ such that $|f_0(p) - f_0(p')| \leq \zeta$ whenever $|p - p'| \leq \eta$ for all $p, p' \in [\underline{p}, \bar{p}]$. The event $\{|f_0(\pi(X_i)) - f_0(\pi(X_j))| > \zeta\}$ is equal to $\{|f_0(\pi(X_i)) - f_0(\pi(X_j))| > \zeta, |\pi(X_i) - \pi(X_j)| \leq \eta\} \cup \{|f_0(\pi(X_i)) - f_0(\pi(X_j))| > \zeta, |\pi(X_i) - \pi(X_j)| > \eta\}$. The first event has measure zero by uniform continuity. As $j \in \mathcal{J}(i)$, the probability of the second event is dominated by

$$\mathbb{P}\left(|\pi(X_i) - \pi(X_j)| > \eta\right) \leq \mathbb{P}\left(\delta_n > \eta\right),$$

which is $o(1)$. Hence $\max_{i:D_i=0}\max_{j\in\mathcal{J}(i)}|\check{R}_{0ji}| = o_P(1)$, as $\inf_{p\in[\underline{p},\bar{p}]} f_0(p) > 0$ by Assumption 1.2. When the caliper is $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, arguments proving Assertion (ii) apply.

Assertions (iv), (v), (vi) follow along the same arguments by conditioning on $\hat{\theta}$, exploiting Assumptions 1.6 and 1.9 and that the constants of the bounds of Alexander (1987) and Van der Vaart and Wellner (1996) do not depend on the underlying distribution. This holds for any caliper choice in (1.13) as $\widehat{\overline{\Delta}}_n = o_P(1)$ by the proof of Proposition 1.5. ■

Lemma 1.2 (Error Term)**.** *Suppose Assumptions 1.2 and 1.4 hold, and the caliper δ_n satisfies (1.3). Then*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(V_E^{-1/2}\sqrt{n}E \le x \;\middle|\; D^{(n)}, PS^{(n)}\right) - \Phi(x)\right| \xrightarrow{P} 0 \quad \text{as } n\to\infty,$$

for $V_E := \frac{1}{n}\sum_{i\in[n]}(\mathbb{1}_{M_i>0} + w_i)^2\sigma^2_{D_i}(\pi(X_i))$ and standard normal distribution function Φ.

Proof. We apply a Lindeberg–Feller central limit theorem as the E_i, given $(D^{(n)}, PS^{(n)})$, are independently, but not identically, distributed with mean zero across $i\in[n]$ (the M_i, w_i are constants given $(D^{(n)}, PS^{(n)})$). By Assumption 1.2, the μ^d, and hence ε, are well-defined. By definition of V_E,

$$\mathbb{V}\left[\sum_{i\in[n]} E_i/\sqrt{nV_E} \mid D^{(n)}, PS^{(n)}\right] = 1.$$

Thus, we only need to verify the Lindeberg–Feller condition:

$$\sum_{i\in[n]}\mathbb{E}\left[(E_i/\sqrt{nV_E})^2\mathbb{1}_{|E_i/\sqrt{nV_E}|\ge\eta}\;\middle|\; D^{(n)}, PS^{(n)}\right] \xrightarrow{P} 0 \quad \text{for all constants } \eta > 0. \tag{1.43}$$

Since $\mathbb{1}_{|E_i/\sqrt{nV_E}|\ge\eta}$ is bounded by $E_i^2/(\eta^2 nV_E)$ on $\{E_i^2/(\eta^2nV_E)\ge 1\}$, the expectation in (1.43) satisfies

$$\begin{aligned}\frac{1}{nV_E}\mathbb{E}\left[E_i^2\mathbb{1}_{|E_i|\ge\eta\sqrt{nV_E}}\;\middle|\; D^{(n)}, PS^{(n)}\right] &\le \frac{1}{nV_E}\mathbb{E}\left[E_i^4/(\eta^2nV_E)\mid D^{(n)}, PS^{(n)}\right] \\ &= \frac{\mathbb{E}\left[E_i^4 \mid D^{(n)}, PS^{(n)}\right]}{(\eta nV_E)^2},\end{aligned} \tag{1.44}$$

where $\mathbb{E}\left[E_i^4 \,\middle|\, D^{(n)}, PS^{(n)}\right] = (\mathbb{1}_{M_i>0} + w_i)^4 \mathbb{E}\left[\varepsilon_i^4 \,\middle|\, D_i, \pi(X_i)\right]$, with

$$\mathbb{E}\left[\varepsilon_i^4 \,\middle|\, D_i, \pi(X_i)\right] \leq \sup_{d\in\{0,1\}, p\in[0,1]} \mathbb{E}\left[\varepsilon_i^4 \mid D_i = d, \pi(X_i) = p\right] < \infty$$

by Assumption 1.4. By Lemma 1.4, $V_E \xrightarrow{P} V_{\sigma,\pi} \in (0,\infty)$, so (1.44) is bounded by $(1 + o_P(1))n^{-1}\left(\frac{1}{n}\sum_{i\in[n]}(\mathbb{1}_{M_i>0} + w_i)^4\right)$ up to a constant factor. Fix a constant $C > 0$. By Markov's inequality,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i\in[n]}(\mathbb{1}_{M_i>0} + w_i)^4 > C\right) \leq C^{-1}\max_{i\in[n]}\mathbb{E}(\mathbb{1}_{M_i>0} + w_i)^4.$$

Below, we show that $n^{-\varrho}\max_{i\in[n]}\mathbb{E}(\mathbb{1}_{M_i>0} + w_i)^4 = O(1)$ for any $\varrho > 0$, so that

$$\frac{1}{n}\sum_{i\in[n]}(\mathbb{1}_{M_i>0} + w_i)^4 = O_P\left(n^{\varrho}\right).$$

Then (1.44) is bounded by $(1 + o_P(1))O_P\left(n^{\varrho-1}\right)$, which is $o_P(1)$ for $\varrho < 1$, so (1.43) is met.

As $w_i \geq 0$ and $(1+x)^4 \leq (2x)^4$ for $x \geq 1$,

$$(\mathbb{1}_{M_i>0} + w_i)^4 \leq 2^4 + 2^4 w_i^4 \mathbb{1}_{w_i>1}.$$

We have $w_i \leq M_i \max_{j\in\mathcal{J}(i)} M_j^{-1}$ and hence

$$\mathbb{E}w_i^4\mathbb{1}_{w_i>1} \leq \mathbb{E}\left[\mathbb{1}_{w_i>1}M_i^4 \max_{j\in\mathcal{J}(i)} M_j^{-4}\right].$$

This yields the result by Lemma 1.3.

When the caliper is $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, it is $\sigma\left\{D^{(n)}, PS^{(n)}\right\}$-measurable, hence the M_i, w_i are constants given $(D^{(n)}, PS^{(n)})$. Lemmas 1.3 and 1.4 complete the proof. ■

Lemma 1.3 (Lindeberg–Feller Bound)**.** *Suppose that the caliper δ_n satisfies (1.3) and that Assumption 1.2 holds. Then for any finite fixed constant integer $r \geq 2$ and any finite fixed constant $\varrho > 0$,*

$$\max_{i\in[n]}\mathbb{E}\mathbb{1}_{w_i>1}\max_{j\in\mathcal{J}(i)}\left(\frac{M_i}{M_j}\right)^r = o\left(n^{\varrho}\right). \tag{1.45}$$

Proof. If $w_i > 1$ for some $i \in [n]$, then $M_i \geq 1$ and hence also $M_j \geq 1$ for all $j \in \mathcal{J}(i)$, as $j \in \mathcal{J}(i)$ if and only if $i \in \mathcal{J}(j)$. This in turn implies $N_0, N_1 \geq 1$. Thus, $\mathbb{1}_{w_i>1} \leq \mathbb{1}_{M_j \geq 1}\mathbb{1}_{N_0 \geq 1}\mathbb{1}_{N_1 \geq 1}$ for all $j \in \mathcal{J}(i)$ for all $i \in [n]$. By definition, $M_i = N_{1-d}\mathbb{F}_{N_{1-d}}[\pi(X_i) \pm \delta_n]$ if $D_i = d$. Then, in the notation of (1.17) and (1.32), the left side of (1.45) is bounded by

$$\begin{aligned}
&\sum_{d\in\{0,1\}} \max_{i:D_i=d} \mathbb{E}\left[\left(\frac{N_{1-d}\mathbb{1}_{N_d\geq 1}}{N_d}\right)^r \right. \\
&\qquad \left. \times \max_{j\in\mathcal{J}(i)} \left(\frac{F_{1-d}[\pi(X_i)\pm\delta_n]}{F_d[\pi(X_j)\pm\delta_n]} \frac{1+R_{1-d,i}}{1+R_{d,j}} \mathbb{1}_{N_{1-d}\geq 1}\mathbb{1}_{M_j\geq 1}\right)^r\right] \\
&\lesssim \sum_{d\in\{0,1\}} \max_{i:D_i=d} \mathbb{E}\left[\left(\frac{N_{1-d}\mathbb{1}_{N_d\geq 1}}{N_d}\right)^r \max_{j\in\mathcal{J}(i)} \left(\frac{1+R_{1-d,i}}{1+R_{d,j}} \mathbb{1}_{N_{1-d}\geq 1}\mathbb{1}_{M_j\geq 1}\right)^r\right],
\end{aligned} \tag{1.46}$$

where the third line follows from Assumption 1.2: because f_0, f_1 have the same support, are bounded away from zero and infinity, we have, for $c \neq 0$, $0 < \frac{c(1\wedge\inf_{p\in[\underline{p},\bar{p}]} f_{1-d}(p))}{c\sup_{p\in[\underline{p},\bar{p}]} f_d(p)} \leq F_{1-d}[a\pm c]/F_d[b\pm c] \leq \frac{8c\sup_{p\in[\underline{p},\bar{p}]} f_{1-d}(p)}{c(1\wedge\inf_{p\in[\underline{p},\bar{p}]} f_d(p))} < \infty$. We address the case $d = 0$ in (1.46); $d = 1$ follows by symmetry. For the expectation in (1.46), two applications of the Cauchy–Schwarz inequality give

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{N_1\mathbb{1}_{N_0\geq 1}}{N_0}\right)^r \max_{j\in\mathcal{J}(i)} \left(\frac{1+R_{1i}}{1+R_{0j}}\mathbb{1}_{N_1\geq 1}\mathbb{1}_{M_j\geq 1}\right)^r\right] \\
&\leq \sqrt{\mathbb{E}(N_1\mathbb{1}_{N_0\geq 1}/N_0)^{2r}\sqrt{\mathbb{E}(1+R_{1i})^{4r}\mathbb{1}_{N_1\geq 1}\mathbb{E}\left(\max_{j\in\mathcal{J}(i)}(\mathbb{1}_{M_j\geq 1}/(1+R_{0j}))^r\right)^4}}.
\end{aligned} \tag{1.47}$$

Here, $\mathbb{E}(N_1\mathbb{1}_{N_0\geq 1}/N_0)^{2r} \leq n^{2r}\mathbb{E}(\mathbb{1}_{N_0\geq 1}/N_0)^{2r}$ with

$$\begin{aligned}
\mathbb{E}(\mathbb{1}_{N_0\geq 1}/N_0)^{2r} &= \sum_{n_0=1}^{n} \binom{n}{n_0}(1-p_1)^{n_0}p_1^{n-n_0}\left(\frac{1}{n_0}\right)^{2r} \\
&= \sum_{n_0=0}^{n-1} \frac{n}{n_0+1}\binom{n-1}{n_0}(1-p_1)^{n_0+1}p_1^{n-(n_0+1)}\left(\frac{1}{n_0+1}\right)^{2r} \\
&= n\frac{1-p_1}{p_1}\sum_{n_0=0}^{n-1}\binom{n-1}{n_0}(1-p_1)^{n_0}p_1^{n-n_0}\left(\frac{1}{n_0+1}\right)^{2r+1}.
\end{aligned}$$

The last line is $O\left(n^{-2r}\right)$ by Cribari-Neto et al. (2000), thus $\mathbb{E}(N_1\mathbb{1}_{N_0\geq 1}/N_0)^{2r} = O(1)$.

Next, we bound the second factor of (1.47). We have $\mathbb{E}(1+R_{1i})^{4r}\mathbb{1}_{N_1\geq 1} \leq 2^{4r} + c\mathbb{E}|R_{1i}|^{4r}\mathbb{1}_{N_1\geq 1}$ for some constant $c>0$, where $\max_{i:D_i=0}|R_{1i}| \leq W_1$ for W_1 in (1.38). Because $W_1 \geq 0$,

$$\mathbb{E}\left[W_1^{4r} \mid D^{(n)}\right] = \int_0^\infty \mathbb{P}\left(W_1^{4r} > w \mid D^{(n)}\right) \mathrm{d}w = \int_0^\infty \mathbb{P}\left(W_1 > w^{\frac{1}{4r}} \mid D^{(n)}\right) \mathrm{d}w.$$

By (1.42) of Lemma 1.1,

$$\mathbb{E}\left[W_1^{4r} \mid D^{(n)}\right] \leq \frac{c_0}{N_1\gamma_n}\int_0^\infty \frac{1}{w^{\frac{2}{4r}}} e^{-w^{\frac{2}{4r}}N_1\gamma_n/1024}\mathrm{d}w + c_1\int_0^\infty e^{-w^{\frac{1}{4r}}N_1\gamma_n/256}\mathrm{d}w \tag{1.48}$$

$$+ cN_1\int_0^\infty w^{\frac{2}{4r}} e^{-\frac{N_1 w^{\frac{2}{4r}}}{2}}\mathrm{d}w. \tag{1.49}$$

The integral in the first term of (1.48) is $\frac{c_0}{N_1\gamma_n}\frac{1}{\lambda_{n,N_1}}\int_0^\infty t^{-1}\lambda_{n,N_1}e^{-\lambda_{n,N_1}t}t^{2r-1}\mathrm{d}t$, where $\lambda_{n,N_1} := \frac{N_1\gamma_n}{1024}$ is strictly positive for $N_1 \geq 1$. This integral is the $(2r-2)$th moment of an Exponential(λ_{n,N_1}) variable, which is well-defined for $r \geq 2$ and for finite integer r is bounded by $\left(\frac{1}{\lambda_{n,N_1}}\right)^{2r-2}$ up to a constant factor. Hence, the first term of (1.48) is bounded by $c_0\left(\frac{1}{N_1\gamma_n}\right)^{2r} \simeq c_0\left(\frac{n}{N_1\log n}\right)^{2r}$. Similar arguments show that the second integrals of (1.48) and (1.49) are bounded by $c_1\left(\frac{1}{N_1\gamma_n}\right)^{4r}$ and $c\left(\frac{1}{N_1}\right)^{2r}$, respectively. Conclude that

$$\mathbb{E}(1+R_{1i})^{4r}\mathbb{1}_{N_1\geq 1} \leq O(1) + c_0 n^{2r}\mathbb{E}\left(\frac{\mathbb{1}_{N_1\geq 1}}{N_1}\right)^{2r} + c_1 n^{4r}\mathbb{E}\left(\frac{\mathbb{1}_{N_1\geq 1}}{N_1}\right)^{4r}.$$

By arguments bounding $\mathbb{E}(N_1\mathbb{1}_{N_0\geq 1}/N_0)^{2r}$ of (1.47), the right side is $O(1)$.

Finally, the last factor of (1.47) satisfies

$$\begin{aligned}\max_{j\in\mathcal{J}(i)}\left(\mathbb{1}_{M_j\geq 1}/(1+R_{0j})\right)^r &= \max_{j\in\mathcal{J}(i)}\left(\frac{N_0 F_0[\pi(X_j)\pm\delta_n]}{M_j}\mathbb{1}_{M_j\geq 1}\right)^r \\ &\leq (8n\|f_0\|_\infty\delta_n)^r \\ &\lesssim (\log n)^r.\end{aligned} \tag{1.50}$$

As $(\log n)^r = o(n^\varrho)$ for any $\varrho > 0$, (1.45) follows.

When the caliper is $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$, (1.48) holds with $\gamma_n := (1 \wedge \inf_{p\in[\underline{p},\bar{p}]} f_1(p))\frac{\log N_1}{N_1+1}$ in view of the proof of Lemma 1.1, showing

$$\mathbb{E}(1+R_{1i})^{4r}\mathbb{1}_{N_1\geq 1} = O(1).$$

Consider (1.50), wherein $\delta_n^r \le 2^r \left(\overline{\Delta}_n^r + (2\log n)^r \left(\frac{1}{N_0+1}\right)^r + \left(\frac{1}{N_1+1}\right)^r\right)$. Here, $\mathbb{E}\left(\frac{1}{N_d+1}\right)^r = O\left(n^{-r}\right)$ by Cribari-Neto et al. (2000). Arguments in the proof of Proposition 1.1 and Lemma 1.5 imply $n^r\mathbb{E}\overline{\Delta}_n^r = O\left((\log n)^r\right)$. Conclude that $n^r\mathbb{E}\delta_n^r = O\left((\log n)^r\right)$ and hence (1.45) follows from (1.50). ■

Lemma 1.4 (Semiparametric Efficiency). *If Assumptions 1.2 and 1.4 hold, and the caliper δ_n satisfies (1.3), then $V_E \xrightarrow{P} V_{\sigma,\pi}$ as $n \to \infty$.*

Proof. By Proposition 1.2, $\min_{i\in[n]} M_i > 0$ with probability tending to one, so with probability tending to one,

$$\begin{aligned} V_E &= \frac{1}{n}\sum_{i\in[n]}(1+w_i)^2\sigma^2_{D_i}(\pi(X_i)) = \frac{1}{n}\sum_{i\in[n]}(1+2w_i+w_i^2)\sigma^2_{D_i}(\pi(X_i)) \\ &= \mathbb{E}\sigma^2_D(\pi(X)) + o_P(1) + \frac{2}{n}\sum_{i\in[n]} w_i\sigma^2_{D_i}(\pi(X_i)) + \frac{1}{n}\sum_{i\in[n]} w_i^2\sigma^2_{D_i}(\pi(X_i)) \end{aligned} \tag{1.51}$$

by the law of large numbers. In the notation of (1.17), we have, by definition,

$$\begin{aligned} \sum_{i\in[n]} w_i^r\sigma^2_{D_i}(\pi(X_i)) &= \left(\frac{N_1}{N_0}\right)^r \sum_{i:D_i=0}\left(\frac{1}{N_1}\sum_{j\in\mathcal{J}(i)}\frac{1}{\mathbb{F}_{N_0}[\pi(X_j)\pm\delta_n]}\right)^r \sigma^2_{D_i}(\pi(X_i)) \\ &\quad + \left(\frac{N_0}{N_1}\right)^r \sum_{i:D_i=1}\left(\frac{1}{N_0}\sum_{j\in\mathcal{J}(i)}\frac{1}{\mathbb{F}_{N_1}[\pi(X_j)\pm\delta_n]}\right)^r \sigma^2_{D_i}(\pi(X_i)). \end{aligned} \tag{1.52}$$

Write $\mathbb{F}_{N_0}[\pi(X_j)\pm\delta_n] = (1+R_{0j})F_0[\pi(X_j)\pm\delta_n]$ and $F_0[\pi(X_j)\pm\delta_n] = 2\delta_n f_0(\pi(X_j))(1+\tilde{R}_{0j})$ for $R_{0j}, \tilde{R}_{0j}$ of (1.32), (1.33). By Lemma 1.1,

$$\max_{j:D_j=1}|R_{0j}| = o_P(1) \quad \text{and} \quad \max_{j:D_j=1}|\tilde{R}_{0j}| = o_P(1).$$

Then we can write

$$\begin{aligned} \frac{1}{N_1}\sum_{j\in\mathcal{J}(i)}\frac{1}{\mathbb{F}_{N_0}[\pi(X_j)\pm\delta_n]} &= \frac{1}{N_1}\sum_{j\in\mathcal{J}(i)}\frac{1}{F_0[\pi(X_j)\pm\delta_n]}\frac{1}{1+R_{0j}} \\ &= \frac{1}{N_1}\sum_{j\in\mathcal{J}(i)}\frac{1}{2\delta_n f_0(\pi(X_j))}\frac{1}{1+\tilde{R}_{0j}}\frac{1}{1+R_{0j}} \\ &= \frac{1+o_P(1)}{N_1}\sum_{j\in\mathcal{J}(i)}\frac{1}{2\delta_n f_0(\pi(X_j))} \end{aligned}$$

where the $o_P(1)$ terms are uniform in $i \in [n]$. Write $f_0(\pi(X_j)) = f_0(\pi(X_i))(1+\check{R}_{0ji})$ for $\check{R}_{0ji}$ of (1.34). By Lemma 1.1, $\max_{i:D_i=0} \max_{j\in\mathcal{J}(i)} |\check{R}_{0ji}| = o_P(1)$. Then, by the continuous mapping theorem,

$$\begin{aligned}\frac{1}{N_1}\sum_{j\in\mathcal{J}(i)}\frac{1}{2\delta_n f_0(\pi(X_j))} &= \frac{1}{N_1}\sum_{j\in\mathcal{J}(i)}\frac{1}{2\delta_n f_0(\pi(X_i))}\frac{1}{1+\check{R}_{0ji}} \\ &= (1+o_P(1))\frac{f_1(\pi(X_i))}{f_0(\pi(X_i))}\frac{\mathbb{F}_{N_1}[\pi(X_i)\pm\delta_n]}{2\delta_n f_1(\pi(X_i))},\end{aligned}$$

because $\inf_{p\in[\underline{p},\bar{p}]} f_1(p) > 0$ by Assumption 1.2. Write

$$\frac{\mathbb{F}_{N_1}[\pi(X_i)\pm\delta_n]}{2\delta_n f_1(\pi(X_i))} = \frac{\mathbb{F}_{N_1}[\pi(X_i)\pm\delta_n]}{F_1[\pi(X_i)\pm\delta_n]}\frac{F_1[\pi(X_i)\pm\delta_n]}{2\delta_n f_1(\pi(X_i))},$$

where $\max_{i:D_i=0}\left|\frac{\mathbb{F}_{N_1}[\pi(X_i)\pm\delta_n]}{F_1[\pi(X_i)\pm\delta_n]}-1\right| = o_P(1)$ and $\max_{i:D_i=0}\left|\frac{F_1[\pi(X_i)\pm\delta_n]}{2\delta_n f_1(\pi(X_i))}-1\right| = o_P(1)$ by Lemma 1.1. By symmetry, similar arguments apply to the second term of (1.52). Then (1.52), divided by n, is equal to

$$\begin{aligned}&\frac{1+o_P(1)}{n}\left[\left(\frac{N_1}{N_0}\right)^r\sum_{i:D_i=0}\left(\frac{f_1(\pi(X_i))}{f_0(\pi(X_i))}\right)^r\sigma^2_{D_i}(\pi(X_i))\right. \\ &\qquad\left.+\left(\frac{N_0}{N_1}\right)^r\sum_{i:D_i=1}\left(\frac{f_0(\pi(X_i))}{f_1(\pi(X_i))}\right)^r\sigma^2_{D_i}(\pi(X_i))\right] \\ &\xrightarrow{P}\mathbb{E}\left[(1-D)\left(\frac{p_1}{1-p_1}\frac{f_1(\pi(X))}{f_0(\pi(X))}\right)^r\sigma^2_D(\pi(X))\right] \qquad (1.53) \\ &\qquad+\mathbb{E}\left[D\left(\frac{1-p_1}{p_1}\frac{f_0(\pi(X))}{f_1(\pi(X))}\right)^r\sigma^2_D(\pi(X))\right], \qquad (1.54)\end{aligned}$$

by the weak law of large numbers as $(N_d/N_{1-d})^r \xrightarrow{a.s.} \left(\frac{p_d}{1-p_d}\right)^r$ and the f_d/f_{1-d} are uniformly bounded by Assumption 1.2 and the σ^2_d are bounded by Assumption 1.4. For $r=1$, (1.53) is

$$\begin{aligned}\mathbb{E}\left[\frac{p_1}{1-p_1}\frac{f_1(\pi(X))}{f_0(\pi(X))}\sigma^2_0(\pi(X))\,\middle|\,D=0\right](1-p_1) &= \int_{\underline{p}}^{\bar{p}} p_1\frac{f_1(p)}{f_0(p)}\sigma^2_0(p)f_0(p)\mathrm{d}p \\ &= \mathbb{E}\pi(X)\sigma^2_0(\pi(X)),\end{aligned}$$

where we used that $f_1(p) = \frac{p}{p_1}f_{\pi(X)}(p)$, with $f_{\pi(X)}$ being the density of $\pi(X)$. Noting that $f_0(p) = \frac{1-p}{1-p_1}f_{\pi(X)}(p)$, (1.54) is $\mathbb{E}(1-\pi(X))\sigma^2_1(\pi(X))$. For $r=2$, the same arguments yield $\mathbb{E}\frac{\pi(X)^2}{1-\pi(X)}\sigma^2_0(\pi(X))$ for (1.53), and $\mathbb{E}\frac{(1-\pi(X))^2}{\pi(X)}\sigma^2_1(\pi(X))$

for (1.54). Note that $\mathbb{E}\sigma_D^2(\pi(X)) = \mathbb{E}(1-\pi(X))\sigma_0^2(\pi(X)) + \mathbb{E}\pi(X)\sigma_1^2(\pi(X))$. Collect the terms to get the assertion, which also holds for $\delta_n = \overline{\Delta}_n \vee \frac{\log N_0}{N_0+1} \vee \frac{\log N_1}{N_1+1}$ in view of Lemma 1.1. ■

1.C. Auxiliary Results

Proof of Proposition 1.3. The semiparametric lower bound is the variance of the efficient influence function. When the observed sample is $((Y_i, D_i, X_i))_{i\in[n]}$, the efficient influence function of ATE is

$$\chi_{\mathcal{X}}(Y,D,X) \coloneqq \frac{D(Y-\mu_{\mathcal{X}}^1(X))}{\pi(X)} - \frac{(1-D)(Y-\mu_{\mathcal{X}}^0(X))}{1-\pi(X)} + \mu_{\mathcal{X}}^1(X) - \mu_{\mathcal{X}}^0(X) - \tau,$$

see Hahn (1998). Then $V = V_{\text{eff}}$ follows under (1.10) in Proposition 1.3. One can verify that $V = \mathbb{V}[\chi(Y,D,\pi(X))]$, where

$$\chi(Y,D,\pi(X)) \coloneqq \frac{D(Y-\mu^1(\pi(X)))}{\pi(X)} - \frac{(1-D)(Y-\mu^0(\pi(X)))}{1-\pi(X)} + \mu^1(\pi(X)) - \mu^0(\pi(X)) - \tau.$$

Assumption 1.1 and $\mathbb{E}[D \mid X] = \pi(X)$ imply that

$$\mathbb{E}\chi_{\mathcal{X}}(Y,D,X)(\chi(Y,D,\pi(X)) - \chi_{\mathcal{X}}(Y,D,X)) = 0,$$

thus

$$\begin{aligned} V = \mathbb{V}[\chi(Y,D,\pi(X))] &= \mathbb{V}[\chi(Y,D,\pi(X)) - \chi_{\mathcal{X}}(Y,D,X)] + \mathbb{V}[\chi_{\mathcal{X}}(Y,D,X)] \\ &\geq \mathbb{V}[\chi_{\mathcal{X}}(Y,D,X)] = V_{\text{eff}}. \end{aligned}$$

For ATT, $V_{\text{t,eff}} \leq V_{\text{t}}$ follows similarly as the efficient influence function of ATT under unknown propensity score (Hahn, 1998) is

$$\chi_{\text{t},\mathcal{X}}(Y,D,X) \coloneqq \frac{D}{p_1}(Y-\mu_{\mathcal{X}}^1(X)) - \frac{1-D}{p_1}\frac{\pi(X)}{1-\pi(X)}(Y-\mu_{\mathcal{X}}^0(X)) + \frac{D}{p_1}(\mu_{\mathcal{X}}^1(X) - \mu_{\mathcal{X}}^0(X) - \tau_{\text{t}}).$$

Under Assumption 1.1, one verifies that $V_{\mathrm{t}} = \mathbb{V}\left[\chi_{\mathrm{t}}(Y, D, \pi(X))\right]$, where

$$\chi_{\mathrm{t}}(Y, D, \pi(X)) \coloneqq \frac{D}{p_1}(Y - \mu^1(\pi(X))) - \frac{1-D}{p_1}\frac{\pi(X)}{1-\pi(X)}(Y - \mu^0(\pi(X))) + \frac{D}{p_1}(\mu^1(\pi(X)) - \mu^0(\pi(X)) - \tau_{\mathrm{t}}),$$

and that

$$\mathbb{E}\chi_{\mathrm{t},\mathcal{X}}(Y, D, X)(\chi_{\mathrm{t}}(Y, D, \pi(X)) - \chi_{\mathrm{t},\mathcal{X}}(Y, D, X)) = 0,$$

thus proving the assertion. ■

Lemma 1.5 (Moment Bounds of Ordered Uniform Spacings)**.** *Let the order statistics of* $(U_1, U_2, \ldots, U_n) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0,1)$ *be* $U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(n)}$*. Let* $\tilde{U}_1 \coloneqq U_{(1)}$*,* $\tilde{U}_i \coloneqq U_{(i)} - U_{(i-1)}$ *for* $i = 2, \ldots, n$ *and* $\tilde{U}_{n+1} \coloneqq 1 - U_{(n)}$ *be the spacings generated by* $(U_i)_{i\in[n]}$*. Let* $\tilde{U}_{(1)} \leq \tilde{U}_{(2)} \leq \ldots \leq \tilde{U}_{(n+1)}$ *be the ordered spacings. Then for any finite fixed integer* $1 \leq a < \frac{n+1}{2}$ *and* $n \geq 2$*,*

$$\mathbb{E}\tilde{U}^a_{(r)} = \begin{cases} O\left(\left(\frac{1}{n}\right)^{2a}\right) & \text{for } r = 1 \\ O\left(\left(\frac{\log r}{n}\right)^{a}\right) & \text{for all } r = 2, 3, \ldots, n+1. \end{cases}$$

In particular, $\mathbb{E}\tilde{U}^a_{(r)} = o(1)$ *for all* $r \in [n+1]$ *and finite fixed integer* $1 \leq a < \frac{n+1}{2}$*.*

Proof. By Shorack and Wellner (2009, Chapter 21), $\tilde{U}_{(r)} \sim \frac{Z_{r:n+1}}{\sum_{i\in[n+1]} Z_i}$, for $(Z_1, Z_2, \ldots, Z_{n+1}) \overset{\text{i.i.d.}}{\sim} \text{Exponential}(1)$ with order statistics $Z_{1:n+1} \leq Z_{2:n+1} \leq \ldots \leq Z_{n+1:n+1}$. The Cauchy–Schwarz inequality gives

$$\mathbb{E}\tilde{U}^a_{(r)} \leq \sqrt{\mathbb{E}Z^{2a}_{r:n+1}\mathbb{E}\left(\sum_{i\in[n+1]} Z_i\right)^{-2a}}.$$

Here, $\mathbb{E}Z^{2a}_{1:n+1} = O\left(n^{-2a}\right)$ and $\mathbb{E}Z^{2a}_{r:n+1} = O\left((\log r)^{2a}\right)$ for $r \geq 2$ and for finite fixed integer $a \geq 1$ by Lemma 1.6. The sum of $n+1$ i.i.d. Exponential(1) variates follows a Gamma$(n+1, 1)$ distribution. Thus, $(\sum_{i\in[n+1]} Z_i)^{-1}$ follows an Inverse-Gamma$(n+1, 1)$ distribution, whose $2a$-th moment is equal to $\frac{(n-2a)!}{n!} \lesssim n^{-2a}$ for $2a < n+1$, where the last inequality follows from $\sqrt{2\pi}n^{n+1/2}e^{-n} \leq n! \leq en^{n+1/2}e^{-n}$ (Robbins, 1955), because $n - 2a + 1/2 \geq 0$ for a positive integer a. ■

Lemma 1.6 (Moments of Exponential(1) Order Statistics). *Let* $(Z_1, Z_2, \ldots, Z_n) \overset{i.i.d.}{\sim}$ *Exponential(1) with order statistics* $Z_{1:n} \le Z_{2:n} \le \ldots \le Z_{n:n}$. *Then for any finite fixed integer* $k \ge 1$, *we have for all* $n \ge 2$,

$$\mathbb{E}Z_{r:n}^k = k! \sum_{t_1=1}^{r} \frac{1}{n+1-t_1} \sum_{t_2=1}^{t_1} \frac{1}{n+1-t_2} \cdots \sum_{t_{k-1}=1}^{t_{k-2}} \frac{1}{n+1-t_{k-1}} \sum_{t_k=1}^{t_{k-1}} \frac{1}{n+1-t_k} \tag{1.55}$$

for all $r = 1, 2, \ldots, n$. *The right side in* (1.55) *is* $O\left((\log r)^k\right)$ *for* $r \ge 2$, *and* $\mathbb{E}Z_{1:n}^k = O\left(n^{-k}\right)$.

Proof. The right side of (1.55) and $\mathbb{E}Z_{1:n}^k = O\left(n^{-k}\right)$ are obtained by solving the recursion in Balakrishnan and Gupta (1998, Theorems 1 and 2). The innermost sum in (1.55) satisfies $\sum_{t_k=1}^{t_{k-1}} \frac{1}{n+1-t_k} \le \sum_{j=1}^{r} \frac{1}{n+1-j}$ as $t_k \le r$. Because every fraction in (1.55) is positive, we can upper bound the right side of (1.55) by

$$k! \left(\sum_{j=1}^{r} \frac{1}{n+1-j} \right) \left(\sum_{t_1=1}^{r} \frac{1}{n+1-t_1} \sum_{t_2=1}^{t_1} \frac{1}{n+1-t_2} \cdots \sum_{t_{k-1}=1}^{t_{k-2}} \frac{1}{n+1-t_{k-1}} \right).$$

Apply the same bound for the remaining $k-1$ sums noting that $1 \le t_1 \le t_1 \le \ldots \le t_{k-1} \le r$, to obtain the bound $k! \left(\sum_{j=1}^{r} \frac{1}{n+1-j} \right)^k$ on (1.55). The proof is complete as $\sum_{j=1}^{r} \frac{1}{n+1-j}$ is $O(\log r)$ for $r \in [n]$ as $n \to \infty$. ∎

Lemma 1.7 (Conditional Martingale Central Limit Theorem). *Let* $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ *be a sequence of probability spaces. Let* $\xi_{n1}, \xi_{n2}, \ldots, \xi_{nn} : \Omega_n \to \mathbb{R}$ *be martingale differences with respect to sub-*σ*-algebras* $\mathcal{F}_{n1} \subset \mathcal{F}_{n2} \subset \ldots \subset \mathcal{F}_{nn} \subset \mathcal{F}_n$. *Let* $\mathcal{F}_{n0} \subset \mathcal{F}_{n1}$ *be a sub-*σ*-algebra. For* $k = 1, 2, \ldots, n$, *let* $\sigma_{nk}^2 := \mathbb{E}_n\left[\xi_{nk}^2 \,\middle|\, \mathcal{F}_{n,k-1}\right]$. *If there exists a finite constant* $\sigma > 0$ *such that*

$$\mathbb{P}_n\left(\left| \sum_{k=1}^{n} \sigma_{nk}^2 - \sigma^2 \right| > \epsilon \,\middle|\, \mathcal{F}_{n0} \right) \xrightarrow{\mathbb{P}_n} 0 \quad \textit{for all constants } \epsilon > 0 \textit{ and} \tag{1.56}$$

$$\sum_{k=1}^{n} \mathbb{E}_n\left[\xi_{nk}^2 \mathbb{1}_{|\xi_{nk}| \ge \eta} \,\middle|\, \mathcal{F}_{n0}\right] \xrightarrow{\mathbb{P}_n} 0 \quad \textit{for all constants } \eta > 0, \tag{1.57}$$

then $\mathbb{P}_n\left(\sigma^{-1} \sum_{k=1}^{n} \xi_{nk} \le x \,\middle|\, \mathcal{F}_{n0}\right) \xrightarrow{\mathbb{P}_n} \Phi(x)$ *as* $n \to \infty$ *for all* $x \in \mathbb{R}$, *where* Φ *is the standard normal distribution function.*

Proof. Follows from Billingsley (1995, Theorem 35.12) by conditioning on $\mathcal{F}_{n0}$ throughout. ∎

Proof of Proposition 1.7. Existence of $A_n, \hat{A}_n$. We show that $A_n, \hat{A}_n$ are well-defined with probability tending to one. For A_n, this happens if and only if

$$\mathbb{P}\left(\underline{p}_{\hat{\theta}} + a_n < \bar{p}_{\hat{\theta}} - a_n\right) = \mathbb{P}\left(2a_n < \bar{p}_{\hat{\theta}} - \underline{p}_{\hat{\theta}}\right) \to 1.$$

If $\underline{p}_{\hat{\theta}} = \underline{p} + o_P(1)$ and $\bar{p}_{\hat{\theta}} = \bar{p} + o_P(1)$, the probability is $\mathbb{P}\left(2a_n < \bar{p} - \underline{p} + o_P(1)\right)$ $= \mathbb{P}\left(2 < a_n^{-1}(\bar{p} - \underline{p}) + o_P\left(a_n^{-1}\right)\right)$, which goes to one as $\bar{p} > \underline{p}$ and $a_n \downarrow 0$. To show that these conditions hold, define $T_n(x) \coloneqq \hat{\theta}^\intercal x$ and $T(x) \coloneqq \theta_0^\intercal x$, so by definition $\bar{p} = \sup_{x\in\mathcal{X}} g(T(x)) = g(\sup_{x\in\mathcal{X}} T(x))$ and $\bar{p}_{\hat{\theta}} = \sup_{x\in\mathcal{X}} g(T_n(x)) = g(\sup_{x\in\mathcal{X}} T_n(x))$ because g is increasing by Assumption 1.11. Since g is continuous by Assumption 1.11, it suffices by the continuous mapping theorem to show $\sup_{x\in\mathcal{X}} T_n(x) \xrightarrow{P} \sup_{x\in\mathcal{X}} T(x)$. Because $\mathcal{X}$ is bounded and $\hat{\theta} \xrightarrow{P} \theta_0$ by Assumption 1.9, $\sup_{x\in\mathcal{X}} |T_n(x) - T(x)| \lesssim \|\hat{\theta} - \theta_0\| = o_P(1)$, so that $\bar{p}_{\hat{\theta}} = \bar{p} + o_P(1)$. Similar arguments yield $\underline{p}_{\hat{\theta}} = \underline{p} + o_P(1)$.

For $\hat{A}_n$, the desired result follows from that for A_n above, and that $\min_{i\in[n]} g(\hat{\theta}^\intercal X_i) = \underline{p}_{\hat{\theta}} + o_P(1)$ and $\max_{i\in[n]} g(\hat{\theta}^\intercal X_i) = \bar{p}_{\hat{\theta}} + o_P(1)$. Because $F_{\hat{\theta}}^{-1}$, the inverse of $F_{\hat{\theta}}(p) = p_1 F_{1,\hat{\theta}}(p) + (1 - p_1)F_{0,\hat{\theta}}(p)$ which is the distribution of $(\pi(X, \hat{\theta}) \mid \hat{\theta})$ under Assumption 1.9, is strictly increasing by Assumption 1.6, $(\min_{i\in[n]} \pi(X_i, \hat{\theta}) \mid \hat{\theta})$ is distributed as $F_{\hat{\theta}}^{-1}(U_{(1)})$, where $U_{(1)}$ is the sample minimum of $((U_1, U_2, \ldots, U_n) \mid \hat{\theta}) \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$. Then $(\min_{i\in[n]} \pi(X_i, \hat{\theta}) - \underline{p}_{\hat{\theta}} \mid \hat{\theta})$ is distributed as $F_{\hat{\theta}}^{-1}(U_{(1)}) - F_{\hat{\theta}}^{-1}(F_{\hat{\theta}}(\underline{p}_{\hat{\theta}})) = F_{\hat{\theta}}^{-1}(U_{(1)}) - F_{\hat{\theta}}^{-1}(0)$ given $\hat{\theta}$. By Assumption 1.6, $F_{\hat{\theta}}^{-1}$ is Lipschitz with constant $\left\|(F_{\hat{\theta}}^{-1})'\right\|_\infty$ with $(F_{\hat{\theta}}^{-1})'(u) = \frac{1}{f_{\hat{\theta}}(F_{\hat{\theta}}^{-1}(u))}$ finite for $\inf_{p\in\underline{p}_{\hat{\theta}},\bar{p}_{\hat{\theta}}} f_{\hat{\theta}}(p) > 0$ by Assumption 1.6 for $\hat{\theta} \in \text{Nb}(\theta_0, \epsilon)$. Thus, $F_{\hat{\theta}}^{-1}(U_{(1)}) - F_{\hat{\theta}}^{-1}(0) \lesssim U_{(1)}$. Here, $\mathbb{E}U_{(1)}$ goes to zero by the proof of Proposition 1.1. Hence, by Assumption 1.11, $\min_{i\in[n]} g(\hat{\theta}^\intercal X_i) = \underline{p}_{\hat{\theta}} + o_P(1)$ and $\max_{i\in[n]} g(\hat{\theta}^\intercal X_i) = \bar{p}_{\hat{\theta}} + o_P(1)$ similarly. In the following, we prove the consistency of the variance component estimators for $V_{\hat{\pi}}$; similar arguments give the result for $V_{\text{t},\hat{\pi}}$. We show below that $\hat{N}/n \xrightarrow{P} 1$ (since (1.62) is $o_P(1)$). Therefore, in the following, we prove the consistency of the estimators $\hat{V}_\tau, \hat{V}_{\tau_\text{t}}, \hat{V}_{\sigma,\pi}, \hat{V}_{\text{t},\sigma,\pi}, \hat{q}_d$ and $\hat{q}_{\text{t},d}$ normalised by n rather than $\hat{N}$.

Consistency of $\hat{V}_\tau$. By Theorem 1.3 and the continuous mapping theorem, $(\hat{\tau}_{\hat{\pi}})^2 \xrightarrow{P} \tau^2$. For short, put $\pi_i \coloneqq g(\theta_0^\intercal X_i)$, $\hat{\pi}_i \coloneqq g(\hat{\theta}^\intercal X_i)$ and $\hat{\mathbb{1}}_i \coloneqq \mathbb{1}_{g(\hat{\theta}^\intercal X_i)\in\hat{A}_n}$. The first term in (1.16) is

$$\frac{1}{n}\sum_{i\in[n]}[\mu^1(\hat{\theta},\hat{\pi}_i)-\mu^0(\hat{\theta},\hat{\pi}_i)]^2\hat{\mathbb{1}}_i \tag{1.58}$$

$$+\frac{1}{n}\sum_{i\in[n]}[\hat{\mu}^1(\hat{\theta},\hat{\pi}_i)-\hat{\mu}^0(\hat{\theta},\hat{\pi}_i)-(\mu^1(\hat{\theta},\hat{\pi}_i)-\mu^0(\hat{\theta},\hat{\pi}_i))]^2\hat{\mathbb{1}}_i \tag{1.59}$$

$$+\frac{2}{n}\sum_{i\in[n]}[\hat{\mu}^1(\hat{\theta},\hat{\pi}_i)-\hat{\mu}^0(\hat{\theta},\hat{\pi}_i)-(\mu^1(\hat{\theta},\hat{\pi}_i)-\mu^0(\hat{\theta},\hat{\pi}_i))][\mu^1(\hat{\theta},\hat{\pi}_i)-\mu^0(\hat{\theta},\hat{\pi}_i)]\hat{\mathbb{1}}_i. \tag{1.60}$$

Here, (1.58) converges to $\mathbb{E}(\mu^1(\theta_0,\pi_i)-\mu^0(\theta_0,\pi_i))^2$. To see this, first note that under Assumption 1.9,

$$\left|\frac{1}{n}\sum_{i\in[n]}[\mu^1(\hat{\theta},\hat{\pi}_i)-\mu^0(\hat{\theta},\hat{\pi}_i)]^2\hat{\mathbb{1}}_i-\frac{1}{n}\sum_{i\in[n]}[\mu^1(\theta_0,\pi_i)-\mu^0(\theta_0,\pi_i)]^2\hat{\mathbb{1}}_i\right|\xrightarrow{P}0,$$

which follows from a mean-value expansion of $[\mu^1(\hat{\theta},\hat{\pi}_i)-\mu^0(\hat{\theta},\hat{\pi}_i)]^2$ in $\hat{\theta}$, similarly to the treatment of (1.26) in the proof of Theorem 1.3. Second, as the $\mu^d(\theta,\cdot)$, $\theta\in\mathrm{Nb}(\theta_0,\epsilon)$, are bounded by Assumption 1.5,

$$\left|\frac{1}{n}\sum_{i\in[n]}[\mu^1(\theta_0,\pi_i)-\mu^0(\theta_0,\pi_i)]^2\hat{\mathbb{1}}_i-\frac{1}{n}\sum_{i\in[n]}[\mu^1(\theta_0,\pi_i)-\mu^0(\theta_0,\pi_i)]^2\right| \tag{1.61}$$

is of the order

$$\begin{aligned}\frac{1}{n}\sum_{i\in[n]}(1-\hat{\mathbb{1}}_i)=&\frac{1}{n}\sum_{i\in[n]}\left(\mathbb{1}_{g(\hat{\theta}^\intercal X_i)\notin\hat{A}_n}-\mathbb{E}\left[\mathbb{1}_{g(\hat{\theta}^\intercal X_i)\notin\hat{A}_n}\,\middle|\,\hat{\theta}\right]\right)\\&+\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left[\mathbb{1}_{g(\hat{\theta}^\intercal X_i)\notin\hat{A}_n}\,\middle|\,\hat{\theta}\right]\end{aligned} \tag{1.62}$$

with probability tending to one. Here the first term has mean zero and variance bounded by $1/n$, so it converges to zero in the first mean, and then so in

probability. By Assumption 1.9, the second term of (1.62) is

$$\mathbb{P}\left(g(\hat{\theta}^\intercal X_i) \notin \hat{A}_n \,\middle|\, \hat{\theta}\right) = \mathbb{P}\left(g(\hat{\theta}^\intercal X_i) < \min_{i\in[n]} g(\hat{\theta}^\intercal X_i) + a_n \,\middle|\, \hat{\theta}\right) \tag{1.63}$$

$$+ \mathbb{P}\left(\max_{i\in[n]} g(\hat{\theta}^\intercal X_i) - a_n < g(\hat{\theta}^\intercal X_i) \,\middle|\, \hat{\theta}\right) \tag{1.64}$$

for some $i \in [n]$. Let $\underline{G}_{\hat{\theta}} \coloneqq \min_{i\in[n]} g(\hat{\theta}^\intercal X_i)$. To bound (1.63), we have, by Shanmugam and Arnold (1988),

$$\mathbb{P}\left(g(\hat{\theta}^\intercal X_i) \leq \underline{G}_{\hat{\theta}} + a_n \,\middle|\, \hat{\theta}, \underline{G}_{\hat{\theta}}\right) = \frac{1}{n} + \frac{n-1}{n}\frac{F_{\hat{\theta}}(\underline{G}_{\hat{\theta}} + a_n) - F_{\hat{\theta}}(\underline{G}_{\hat{\theta}})}{1 - F_{\hat{\theta}}(\underline{G}_{\hat{\theta}})} \tag{1.65}$$

under Assumptions 1.9 and 1.11, where $F_{\hat{\theta}} \coloneqq p_1 F_{\hat{\theta},1} + (1-p_1)F_{\hat{\theta},0}$ is the distribution function of $g(\hat{\theta}^\intercal X)$ given $\hat{\theta}$. By Assumption 1.6, $F_{\hat{\theta}}$ is continuous, and by arguments on $\hat{A}_n$ above $\underline{G}_{\hat{\theta}} = \underline{p}_{\hat{\theta}} + o_P(1)$. Then $a_n \downarrow 0$ implies that (1.65) is $o_P(1)$, which in turn implies that (1.63), being bounded by one, is also $o_P(1)$. Term (1.64) is $o_P(1)$ by similar arguments, noting that

$$\left\{\max_{i\in[n]} g(\hat{\theta}^\intercal X_i) - a_n < g(\hat{\theta}^\intercal X_i)\right\} = \left\{\min_{i\in[n]} -g(\hat{\theta}^\intercal X_i) + a_n > -g(\hat{\theta}^\intercal X_i)\right\}.$$

Thus, (1.62) is $o_P(1)$. Conclude that (1.58) converges in probability to $V_\tau + \tau^2$. Write (1.59) as

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in[n]} [\hat{\mu}^1(\hat{\theta}, \hat{\pi}_i) - \mu^1(\hat{\theta}, \hat{\pi}_i)]^2 \hat{\mathbb{1}}_i \\
+\frac{2}{n}\sum_{i\in[n]} &(\hat{\mu}^1(\hat{\theta}, \hat{\pi}_i) - \mu^1(\hat{\theta}, \hat{\pi}_i))(\hat{\mu}^0(\hat{\theta}, \hat{\pi}_i) - \mu^0(\hat{\theta}, \hat{\pi}_i))\hat{\mathbb{1}}_i \\
&+\frac{1}{n}\sum_{i\in[n]} [\hat{\mu}^0(\hat{\theta}, \hat{\pi}_i) - \mu^0(\hat{\theta}, \hat{\pi}_i)]^2 \hat{\mathbb{1}}_i.
\end{aligned}$$

As Assumption 1.10(i) bounds both $\hat{\mu}^d$ and μ^d, $\sup_{p\in A_n} |\hat{\mu}^d(\hat{\theta}, p) - \mu^d(\hat{\theta}, p)| \xrightarrow{P} 0$ implies that all three terms in the last display are $o_P(1)$ (note that $\hat{A}_n \subset A_n$). This convergence can be established along the same lines as that of other estimators, which are detailed below. Similar arguments show that (1.60) is $o_P(1)$. Conclude that $\hat{V}_\tau \xrightarrow{P} V_\tau$.

Consistency of $\hat{V}_{\sigma,\pi}$. First, $0 < g(\hat{\theta}^\intercal x) < 1$ for all x in compact $\mathcal{X}$ and for all $\hat{\theta} \in \text{Nb}(\theta_0, \epsilon)$. Then under Assumption 1.11, a mean-value expansion and

$\hat{\theta} \xrightarrow{P} \theta_0$, implied by Assumption 1.9, gives that $\sup_{x\in\mathcal{X}} |1/g(\hat{\theta}^\intercal x) - 1/g(\theta_0^\intercal x)| = o_P(1)$ and similarly for $1/(1-g(\hat{\theta}^\intercal x))$. Second, the Lipschitz condition in Assumption 1.8(i) implies

$$\left| \frac{1}{n} \sum_{i\in[n]} \sigma_d^2(\hat{\theta}, \hat{\pi}_i) \hat{\mathbb{1}}_i - \frac{1}{n} \sum_{i\in[n]} \sigma_d^2(\theta_0, \pi_i) \hat{\mathbb{1}}_i \right| \xrightarrow{P} 0,$$

and, as in (1.61) above, we also have

$$\left| \frac{1}{n} \sum_{i\in[n]} \sigma_d^2(\theta_0, \pi_i) \hat{\mathbb{1}}_i - \frac{1}{n} \sum_{i\in[n]} \sigma_d^2(\theta_0, \pi_i) \right| \xrightarrow{P} 0.$$

Thus, if $\sup_{p\in A_n} |\hat{\sigma}_d^2(\hat{\theta}, p) - \sigma_d^2(\hat{\theta}, p)| \xrightarrow{P} 0$, then the law of large numbers gives that $\hat{V}_{\sigma,\pi} \xrightarrow{P} V_{\sigma,\pi}$. This holds if both $\sup_{p\in A_n} |\hat{\mu}^d(\hat{\theta}, p) - \mu^d(\hat{\theta}, p)| \xrightarrow{P} 0$ and $\sup_{p\in A_n} |\hat{\mu}_2^d(\hat{\theta}, p) - \mu_2^d(\hat{\theta}, p)| \xrightarrow{P} 0$ since the former implies $\sup_{p\in A_n} |(\hat{\mu}^d(\hat{\theta}, p))^2 - (\mu^d(\hat{\theta}, p))^2| \xrightarrow{P} 0$ under Assumption 1.10(i). See below for a proof of such convergence (e.g. establishing (1.105)).

Consistency of $\hat{q}_d$. By definition of q_d, the conditions

$$\left\| \frac{1}{n} \sum_{i\in[n]} \Lambda^d(\hat{\theta}, X_i) - \frac{1}{n} \sum_{i\in[n]} \Lambda^d(\hat{\theta}, X_i) \hat{\mathbb{1}}_i \right\| \xrightarrow{P} 0 \tag{1.66}$$

$$\sup_{p\in A_n} \left| \widehat{\left(\frac{\partial \mu^d}{\partial p} \right)}(\hat{\theta}, p) - \frac{\partial \mu^d}{\partial p}(\hat{\theta}, p) \right| \xrightarrow{P} 0 \tag{1.67}$$

$$\sup_{p\in A_n} \left| \widehat{\left(\frac{\partial \mu^d}{\partial \theta_k} \right)}(\hat{\theta}, p) - \frac{\partial \mu^d}{\partial \theta_k}(\hat{\theta}, p) \right| \xrightarrow{P} 0 \quad \text{for all } k = 1, 2, \ldots, K, \tag{1.68}$$

together with Assumptions 1.10 and 1.11, implying the boundedness of $\mathcal{X}$ and g', ensure $\hat{q}_d \xrightarrow{P} q_d$. Assumption 1.7, 1.11(i), and Assumption 1.10 imply that Λ^d is bounded uniformly. Therefore, (1.66) is satisfied, because the arguments treating $\hat{\mathbb{1}}_i$ in the case of $\hat{V}_\tau$ above (e.g. (1.61)) apply.

In the following, we show that (1.67) and (1.68) hold, wherein we also give a detailed proof of the uniform consistency of the nonparametric estimators assumed above. For simplicity of exposition, we only give the proof for the (derivatives of) $\mu(\theta, p) := \mathbb{E}[Y \mid g(\theta^\intercal X) = p] = \mathbb{E}\left[Y \mid \theta^\intercal X = g^{-1}(p)\right]$ and $\mu_2(\theta, p) := \mathbb{E}\left[Y^2 \,\middle|\, g(\theta^\intercal X) = p\right]$. The proof when we also condition on $D = d$ follows along the same lines. To this end, let $h(\theta, p) := p_1 h_1(\theta, p) + (1-p_1) h_0(\theta, p)$ be the density of $\pi(X, \theta)$, $q_\mu(\theta, p) := \mu(\theta, p) h(\theta, p)$, $q_{\mu_2}(\theta, p) :=$

$\mu_2(\theta,p)h(\theta,p)$ and their corresponding estimators be obtained by setting $\mathbb{1}_{D_j=d} := 1$ for all $j \in [n]$ in the formulae for $\hat{h}_d$, $\hat{q}_{\mu,d}$, $\hat{q}_{\mu_2,d}$, respectively. For short, we also let $h'(\theta,p) := (\partial/\partial p)h(\theta,p)$, $\hat{h}'(\theta,p) := (\partial/\partial p)\hat{h}(\theta,p)$ and likewise for $q'_\mu, \hat{q}'_\mu$. Furthermore, we let $\widehat{\left(\frac{\partial h}{\partial \theta_k}\right)}$ and $\widehat{\left(\frac{\partial q_\mu}{\partial \theta_k}\right)}$ be obtained by setting $\mathbb{1}_{D_j=d} := 1$ in the formula for $\widehat{\left(\frac{\partial h_d}{\partial \theta_k}\right)}$ and $\widehat{\left(\frac{\partial q_{\mu,d}}{\partial \theta_k}\right)}$, respectively. For short, we put $\dot{h}_k(\theta,p) := (\partial/\partial\theta_k)h(\theta,p)$, $\hat{\dot{h}}_k(\theta,p) := \widehat{\left(\frac{\partial h}{\partial \theta_k}\right)}(\theta,p)$ and likewise for $\dot{q}_{\mu,k}(\theta,p), \hat{\dot{q}}_{\mu,k}(\theta,p)$. For a function $r : \Theta \times [0,1] \to \mathbb{R}$, we let $\|r\|_{A_n} := \sup_{p\in A_n} |r(\hat{\theta},p)|$.

Condition (1.67). First we show that the numerator of

$$\widehat{\left(\frac{\partial \mu}{\partial p}\right)}(\hat{\theta},p) := \frac{\hat{q}'_\mu(\hat{\theta},p)\hat{h}(\hat{\theta},p) - \hat{q}_\mu(\hat{\theta},p)\hat{h}'(\hat{\theta},p)}{(\hat{h}(\hat{\theta},p))^2} \tag{1.69}$$

converges to that of

$$\frac{\partial \mu}{\partial p}(\hat{\theta},p) = \frac{q'_\mu(\hat{\theta},p)h(\hat{\theta},p) - q_\mu(\hat{\theta},p)h'(\hat{\theta},p)}{h(\hat{\theta},p)^2}. \tag{1.70}$$

Consider

$$\begin{aligned}
\left\|\hat{q}'_\mu\hat{h} - \hat{q}_\mu\hat{h}' - (q'_\mu h - q_\mu h')\right\|_{A_n} &\le \left\|\hat{q}'_\mu\hat{h} - q'_\mu h\right\|_{A_n} + \left\|\hat{q}_\mu\hat{h}' - q_\mu h'\right\|_{A_n}, \\
\left\|\hat{q}'_\mu\hat{h} - q'_\mu h\right\|_{A_n} &= \left\|(\hat{q}'_\mu - q'_\mu + q'_\mu)(\hat{h} - h + h) - q'_\mu h\right\|_{A_n} \\
&\le \left\|\hat{q}'_\mu - q'_\mu\right\|_{A_n}\left\|\hat{h} - h\right\|_{A_n} + \left\|\hat{q}'_\mu - q'_\mu\right\|_{A_n}\|h\|_{A_n} \\
&\quad + \left\|q'_\mu\right\|_{A_n}\left\|\hat{h} - h\right\|_{A_n}, \\
\left\|\hat{q}_\mu\hat{h}' - q_\mu h'\right\|_{A_n} &\le \left\|\hat{h}' - h'\right\|_{A_n}\|\hat{q}_\mu - q_\mu\|_{A_n} \\
&\quad + \left\|\hat{h}' - h'\right\|_{A_n}\|q_\mu\|_{A_n} + \left\|h'\right\|_{A_n}\|\hat{q}_\mu - q_\mu\|_{A_n}.
\end{aligned}$$

By Assumption 1.6(iv), $\|h\|_{A_n}$ is bounded with probability tending to one as $\hat{\theta} \xrightarrow{P} \theta_0$ and combining it with Assumptions 1.7 and 1.11(ii), the same holds for $\|h'\|_{A_n}$, $\|q_\mu\|_{A_n} = \|\mu h\|_{A_n}$ and $\left\|q'_\mu\right\|_{A_n} = \|\mu' h + \mu h'\|_{A_n}$. Therefore, if all

$$\left\|\hat{q}'_\mu - q'_\mu\right\|_{A_n} \xrightarrow{P} 0, \tag{1.71}$$

$$\left\|\hat{h}' - h'\right\|_{A_n} \xrightarrow{P} 0, \tag{1.72}$$

$$\|\hat{q}_\mu - q_\mu\|_{A_n} \xrightarrow{P} 0, \tag{1.73}$$

$$\left\|\hat{h} - h\right\|_{A_n} \xrightarrow{P} 0, \tag{1.74}$$

then the numerator of (1.69) converges to that of (1.70) uniformly in $p \in A_n$. The denominator of (1.69) satisfies

$$\begin{aligned}\left\|(\hat{h})^2 - h^2\right\|_{A_n} &= \left\|(\hat{h}-h)(\hat{h}+h+h-h)\right\|_{A_n} \\ &\leq \left\|\hat{h}-h\right\|_{A_n}\left\{\left\|\hat{h}-h\right\|_{A_n} + 2\left\|h\right\|_{A_n}\right\}.\end{aligned}$$

By Assumption 1.6(iv), $\|h\|_{A_n} < \infty$ with probability tending to one as $\hat{\theta} \xrightarrow{P} \theta_0$, hence (1.74) implies that the denominator of (1.69) converges to that of (1.70) uniformly in $p \in A_n$. Thus, both the numerator and the denominator of (1.69) converges to those of (1.70). Because $\inf_{p\in[\underline{p}_{\hat{\theta}},\bar{p}_{\hat{\theta}}]} h(\hat{\theta},p) > 0$ for all $\hat{\theta} \in \mathrm{Nb}(\theta_0,\epsilon)$, it follows that (1.69) converges to (1.70) uniformly in $p \in A_n$. In the following, we show that (1.71)–(1.74) hold.

Condition (1.67)*, part* (1.71). The proof consists in showing

$$\mathbb{E}\mathbb{E}\left[\sup_{p\in A_n}\left|\hat{q}'_\mu(\hat{\theta},p) - \mathbb{E}\left[\hat{q}'_\mu(\hat{\theta},p)\,\middle|\,\hat{\theta}\right]\right|\,\middle|\,\hat{\theta}\right] \to 0 \quad \text{and} \tag{1.75}$$

$$\sup_{p\in A_n}\left|\mathbb{E}\left[\hat{q}'_\mu(\hat{\theta},p)\,\middle|\,\hat{\theta}\right] - q'_\mu(\hat{\theta},p)\right| \xrightarrow{P} 0. \tag{1.76}$$

We show (1.75) following Bierens (1994). For the imaginary unit i, let

$$\psi(t) \coloneqq \int e^{itx}K(x)\mathrm{d}x, \quad t \in \mathbb{R},$$

be the characteristic function of K, which is $\psi(t) = e^{-t^2/2}$ for the Gaussian kernel K, so that $K(x) = (2\pi)^{-1}\int e^{-itx}\psi(t)\mathrm{d}t$ by the inversion formula for characteristic functions. Then $K'(x) = (2\pi)^{-1}\int(-it)e^{-itx}\psi(t)\mathrm{d}t$, hence

$$\begin{aligned}\hat{q}'_\mu(\hat{\theta},p) &= -\frac{1}{n\gamma_n^2}\sum_{j\in[n]} Y_j(2\pi)^{-1}\int(-it)e^{-it(g(\hat{\theta}^\intercal X_j)-p)/\gamma_n}\psi(t)\mathrm{d}t \\ &= (2\pi)^{-1}\int\left(\frac{1}{n}\sum_{j\in[n]} Y_j e^{-itg(\hat{\theta}^\intercal X_j)}\right)e^{itp}it\psi(\gamma_n t)\mathrm{d}t,\end{aligned}$$

where we used a change of variables and Fubini's theorem (the integral is bounded as $|Y_j| < \bar{y}$ almost surely by Assumption 1.10(i), $|it| \leq 1$, $\sup_{t\in\mathbb{R}}|\psi(t)|$

$< \infty$ and the exponential is bounded too). As $|e^{itp}| \leq 1$ and $|it| \leq |t|$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{p\in A_n}\left|\hat{q}'_\mu(\hat{\theta},p) - \mathbb{E}\left[\hat{q}'_\mu(\hat{\theta},p)\,\middle|\,\hat{\theta}\right]\right|\,\middle|\,\hat{\theta}\right] \\
&\leq (2\pi)^{-1}\int\left\{\mathbb{E}\left[\left|\left(\frac{1}{n}\sum_{j\in[n]} Y_j e^{-itg(\hat{\theta}^\intercal X_j)}\right) - \mathbb{E}\left[Y_j e^{-itg(\hat{\theta}^\intercal X_j)}\,\middle|\,\hat{\theta}\right]\right|\,\middle|\,\hat{\theta}\right]\right. \\
&\qquad\left.\times\,|t||\psi(\gamma_n t)|\right\}\mathrm{d}t.
\end{aligned}$$

As $e^{-ia} = \cos(a) - i\sin(a)$ for $a \in \mathbb{R}$, and $\mathbb{E}|W| \leq \sqrt{\mathbb{E}W^2}$ for any square-integrable random variable W, the expectation on the right of the last display is bounded by

$$\begin{aligned}
&\mathbb{V}\left[\frac{1}{n}\sum_{j\in[n]} Y_j\cos(g(\hat{\theta}^\intercal X_j))\,\middle|\,\hat{\theta}\right]^{1/2} + \mathbb{V}\left[\frac{1}{n}\sum_{j\in[n]} Y_j\sin(g(\hat{\theta}^\intercal X_j))\,\middle|\,\hat{\theta}\right]^{1/2} \\
&\qquad\leq 2\sqrt{\frac{\mathbb{E}Y_j^2}{n}},
\end{aligned}$$

where we used that by Assumption 1.9(ii) the elements in the sum are i.i.d. given $\hat{\theta}$, so the covariances are zero, and that $\mathbb{V}\left[Y_j\cos(g(\hat{\theta}^\intercal X_j))\,\middle|\,\hat{\theta}\right] \leq \mathbb{E}\left[Y_j^2\,\middle|\,\hat{\theta}\right] = \mathbb{E}Y_j^2$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\sup_{p\in A_n}\left|\hat{q}'_\mu(\hat{\theta},p) - \mathbb{E}\left[\hat{q}'_\mu(\hat{\theta},p)\,\middle|\,\hat{\theta}\right]\right|\,\middle|\,\hat{\theta}\right] &\leq \sqrt{\frac{\mathbb{E}Y_j^2}{\pi^2 n}}\int|t||\psi(\gamma_n t)|\mathrm{d}t \\
&\leq \sqrt{\frac{\mathbb{E}Y_j^2}{\pi^2 n\gamma_n^4}}\int|t||\psi(t)|\mathrm{d}t.
\end{aligned}$$

As $\mathbb{E}Y_j^2 < \infty$ by Assumption 1.10(i) and $\int|t||\psi(t)|\mathrm{d}t = \int|t|e^{-t^2/2} < \infty$ for the Gaussian K, the right side is of the order $1/(\gamma_n^2\sqrt{n}) = (\kappa_0)^{-2}n^{2\beta-1/2} = o(1)$ for $\beta < 1/4$.

Next, we show (1.76). As the summands are identically distributed given $\hat{\theta}$ by Assumption 1.9(ii), the tower property of expectations gives

$$\begin{aligned}
\mathbb{E}\left[\hat{q}'_\mu(\hat{\theta},p)\,\middle|\,\hat{\theta}\right] &= -\frac{1}{\gamma_n^2}\mathbb{E}\left[\mathbb{E}\left[Y \mid g(\hat{\theta}^\intercal X),\hat{\theta}\right]K'((g(\hat{\theta}X)-p)/\gamma_n)\,\middle|\,\hat{\theta}\right] \\
&= -\frac{1}{\gamma_n^2}\mathbb{E}\left[\mu(\hat{\theta},g(\hat{\theta}^\intercal X))K'((g(\hat{\theta}X)-p)/\gamma_n)\,\middle|\,\hat{\theta}\right] \\
&= -\frac{1}{\gamma_n^2}\int_{\underline{p}_{\hat{\theta}}}^{\bar{p}_{\hat{\theta}}}\mu(\hat{\theta},\tilde{p})K'((\tilde{p}-p)/\gamma_n)h(\hat{\theta},\tilde{p})\mathrm{d}\tilde{p} \\
&= -\frac{1}{\gamma_n^2}\int_{\underline{p}_{\hat{\theta}}}^{\bar{p}_{\hat{\theta}}}q_\mu(\hat{\theta},\tilde{p})K'((\tilde{p}-p)/\gamma_n)\mathrm{d}\tilde{p} \\
&= -\frac{1}{\gamma_n}\int_{(\underline{p}_{\hat{\theta}}-p)/\gamma_n}^{(\bar{p}_{\hat{\theta}}-p)/\gamma_n}q_\mu(\hat{\theta},\gamma_n v+p)K'(v)\mathrm{d}v
\end{aligned}$$

by definition of $h(\hat{\theta},\cdot)$ as the density of $(g(\hat{\theta}^\intercal X)\mid\hat{\theta})$ under Assumptions 1.9(ii) and 1.11(i), and $q_\mu=\mu h$. Integration by parts gives

$$\mathbb{E}\left[\hat{q}'_\mu(\hat{\theta},p)\,\middle|\,\hat{\theta}\right] = -\frac{1}{\gamma_n}\left\{q(\hat{\theta},\bar{p}_{\hat{\theta}})K((\bar{p}_{\hat{\theta}}-p)/\gamma_n)-q(\hat{\theta},\underline{p}_{\hat{\theta}})K((\underline{p}_{\hat{\theta}}-p)/\gamma_n)\right\} \tag{1.77}$$

$$+\int_{(\underline{p}_{\hat{\theta}}-p)/\gamma_n}^{(\bar{p}_{\hat{\theta}}-p)/\gamma_n}q'_\mu(\hat{\theta},\gamma_n v+p)K(v)\mathrm{d}v. \tag{1.78}$$

As $q_\mu(\hat{\theta},\cdot)$ is bounded for $\hat{\theta}\in\mathrm{Nb}(\theta_0,\epsilon)$, with probability tending to one (1.77) is of the order

$$\gamma_n^{-1}\left[\sup_{p\in A_n}K((\bar{p}_{\hat{\theta}}-p)/\gamma_n)+\sup_{p\in A_n}K((\underline{p}_{\hat{\theta}}-p)/\gamma_n)\right]=2\gamma_n^{-1}K(a_n/\gamma_n).$$

As $\gamma_n^{-1}K(a_n/\gamma_n)\to 0$, (1.77) is $o_P(1)$. Now we show that (1.78) converges uniformly to $q'_\mu(\hat{\theta},p)$ adapting Schuster and Yakowitz (1979, Proof of Lemma 1). The kernel K being a density integrating to one implies

$$q'_\mu(\hat{\theta},p)=q'_\mu(\hat{\theta},p)\gamma_n^{-1}\left\{\int_{-\infty}^{p-\bar{p}_{\hat{\theta}}}K(u/\gamma_n)\mathrm{d}u+\int_{p-\bar{p}_{\hat{\theta}}}^{p-\underline{p}_{\hat{\theta}}}K(u/\gamma_n)\mathrm{d}u+\int_{p-\underline{p}_{\hat{\theta}}}^{\infty}K(u/\gamma_n)\mathrm{d}u\right\}.$$

Combine this with a change of variables in (1.78) (with $u \coloneqq -\gamma_n v$ noting that K is symmetric about zero), to get

$$\sup_{p\in A_n}\left|\int_{(\underline{p}_{\hat\theta}-p)/\gamma_n}^{(\bar p_{\hat\theta}-p)/\gamma_n} q'_\mu(\hat\theta,\gamma_n v+p)K(v)\mathrm{d}v - q'_\mu(\hat\theta,p)\right|$$

$$\leq \sup_{p\in A_n}\left|\int_{-\infty}^{p-\bar p_{\hat\theta}} q'_\mu(\hat\theta,p)\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| + \sup_{p\in A_n}\left|\int_{p-\underline{p}_{\hat\theta}}^{\infty} q'_\mu(\hat\theta,p)\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \tag{1.79}$$

$$+ \sup_{p\in A_n}\left|\int_{p-\bar p_{\hat\theta}}^{p-\underline{p}_{\hat\theta}} [q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right|. \tag{1.80}$$

The two terms in (1.79) are $o_P(1)$. The first one is bounded by

$$\sup_{p\in A_n}|q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{-\infty}^{(p-\bar p_{\hat\theta})/\gamma_n} K(v)\mathrm{d}v = \sup_{p\in A_n}|q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{-\infty}^{-a_n/\gamma_n} K(v)\mathrm{d}v,$$

which vanishes as, on one hand, $\sup_{p\in A_n}|q'_\mu(\hat\theta,p)|<\infty$ with probability tending to one as $\hat\theta \xrightarrow{P} \theta_0$ by Assumptions 1.6(iv) and 1.7, and, on the other hand, $a_n/\gamma_n \to \infty$ implies that the integral of the Gaussian K goes to zero. The second term in (1.79) is bounded by

$$\sup_{p\in A_n}|q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{(p-\underline{p}_{\hat\theta})/\gamma_n}^{\infty} K(v)\mathrm{d}v = \sup_{p\in A_n}|q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{a_n/\gamma_n}^{\infty} K(v)\mathrm{d}v,$$

so it is also $o_P(1)$ by the same argument. To show that (1.80) also vanishes, let $\rho_n>0$ be a sequence satisfying $\rho_n/\gamma_n\to\infty$ and $\rho_n<a_n$, i.e. $\gamma_n \ll \rho_n < a_n$ (as $\gamma_n=\kappa_0 n^{-\beta}$ and $a_n=\kappa_1 n^{-\alpha}$, $\beta>\alpha$, we can take $\rho_n=\kappa_2 n^{-(\beta+\alpha)/2}$, $0<\kappa_2<\kappa_1$). Then (1.80) is bounded by

$$\sup_{p\in A_n}\left|\int_{[p-\bar p_{\hat\theta},p-\underline{p}_{\hat\theta}]\cap\{|u|\leq\rho_n\}} [q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \tag{1.81}$$

$$+ \sup_{p\in A_n}\left|\int_{[p-\bar p_{\hat\theta},p-\underline{p}_{\hat\theta}]\cap\{|u|>\rho_n\}} [q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)]u^{-1}u\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right|. \tag{1.82}$$

Here, (1.81) is bounded by

$$\sup_{p\in A_n}\sup_{|u|\le\rho_n}|q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)|\sup_{p\in A_n}\int_{-\infty}^{\infty}\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u \\ \le \sup_{p\in A_n}\sup_{|u|\le\rho_n}|q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)|.$$

By Assumptions 1.6(iv) and 1.7, $q'_\mu(\hat\theta,\cdot)$ is continuous on the compact set $[\underline{p}_{\hat\theta},\bar{p}_{\hat\theta}]\supset A_n$ and is therefore uniformly continuous. Thus, $|u|\le\rho_n$ for a small enough $\rho_n<a_n$ implies that $p-u,p\in[\underline{p}_{\hat\theta},\bar{p}_{\hat\theta}]$ for all $p\in A_n$. Hence, $\sup_{p\in A_n}\sup_{|u|\le\rho_n}|q'_\mu(\hat\theta,p-u)-q'_\mu(\hat\theta,p)|=o_P(1)$. In the integral of (1.82), $u\in[p-\bar{p}_{\hat\theta},p-\underline{p}_{\hat\theta}]$, so $p-u\in[\underline{p}_{\hat\theta},\bar{p}_{\hat\theta}]$, and thus (1.82) is bounded by $2\sup_{p\in[\underline{p}_{\hat\theta},\bar{p}_{\hat\theta}]}|q'_\mu(\hat\theta,p)|$ times

$$\sup_{p\in A_n}\left|\int_{[p-\bar{p}_{\hat\theta},p-\underline{p}_{\hat\theta}]\cap\{|u|>\rho_n\}}\frac{K(u/\gamma_n)}{\gamma_n}\mathrm{d}u\right|\le\sup_{p\in A_n}\int_{\{|u|>\rho_n\}}\frac{K(u/\gamma_n)}{\gamma_n}\mathrm{d}u \\ \le\int_{\{|v|>\rho_n/\gamma_n\}}K(v)\mathrm{d}v,$$

where we used that $K>0$. As $\gamma_n\ll\rho_n$, and $K(v)\downarrow 0$ as $|v|\to\infty$, the right integral tends to zero. As Assumptions 1.6(iv) and 1.7 control $\sup_{p\in[\underline{p}_{\hat\theta},\bar{p}_{\hat\theta}]}|q'_\mu(\hat\theta,p)|$, (1.82) is $o_P(1)$. Thus, (1.76) holds. Conclude that (1.71) holds.

Condition (1.67)*, part* (1.72). Follows directly along the lines of (1.71), setting $Y_j:=1$ for all $j\in[n]$, in the formulae of (1.71).

Condition (1.67)*, part* (1.73). Analogously to $\hat{q}'_\mu$ above, we can write

$$\hat{q}_\mu(\hat\theta,p)=\frac{1}{n\gamma_n}\sum_{j\in[n]}Y_j(2\pi)^{-1}\int e^{-it(g(\hat\theta^\intercal X_j)-p)/\gamma_n}\psi(t)\mathrm{d}t \\ =(2\pi)^{-1}\int\left(\frac{1}{n}\sum_{j\in[n]}Y_je^{-itg(\hat\theta^\intercal X_j)}\right)e^{itp}\psi(\gamma_nt)\mathrm{d}t,$$

and then

$$\mathbb{E}\left[\sup_{p\in A_n}\left|\hat{q}_\mu(\hat\theta,p)-\mathbb{E}\left[\hat{q}_\mu(\hat\theta,p)\,\middle|\,\hat\theta\right]\right|\,\middle|\,\hat\theta\right]\le\sqrt{\frac{\mathbb{E}Y_j^2}{\pi^2n}}\int|\psi(\gamma_nt)|\mathrm{d}t \\ \le\sqrt{\frac{\mathbb{E}Y_j^2}{\pi^2n\gamma_n^2}}\int|\psi(t)|\mathrm{d}t.$$

Assumption 1.10(i) and $\int |\psi(t)|\mathrm{d}t = \int e^{-t^2/2} = 2\pi$ for the Gaussian kernel K mean that the right side is of the order $1/(\gamma_n\sqrt{n}) = (\kappa_0)^{-1}n^{\beta-1/2} = o(1)$ for $\beta < 1/4$. Next, as for $\hat{q}'_\mu$,

$$\begin{aligned}\mathbb{E}\left[\hat{q}_\mu(\hat{\theta},p)\,\middle|\,\hat{\theta}\right] &= \frac{1}{\gamma_n}\int_{\underline{p}_{\hat{\theta}}}^{\bar{p}_{\hat{\theta}}} \mu(\hat{\theta},\tilde{p})K((\tilde{p}-p)/\gamma_n)h(\hat{\theta},\tilde{p})\mathrm{d}\tilde{p} \\ &= \frac{1}{\gamma_n}\int_{\underline{p}_{\hat{\theta}}}^{\bar{p}_{\hat{\theta}}} q_\mu(\hat{\theta},\tilde{p})K((\tilde{p}-p)/\gamma_n)\mathrm{d}\tilde{p} \\ &= \int_{(\underline{p}_{\hat{\theta}}-p)/\gamma_n}^{(\bar{p}_{\hat{\theta}}-p)/\gamma_n} q_\mu(\hat{\theta},\gamma_n v+p)K(v)\mathrm{d}v.\end{aligned}$$

This converges uniformly to $q_\mu(\hat{\theta},p)$ in $p \in A_n$ by the same arguments as (1.78) does to $q'_\mu(\hat{\theta},p)$, given Assumptions 1.6 and 1.7 ensuring the boundedness and continuity of $q_\mu(\hat{\theta},\cdot)$ with probability tending to one as $\hat{\theta} \xrightarrow{P} \theta_0$.

Condition (1.67)*, part* (1.74). Follows from (1.73) by setting $Y_j := 1$ for all $j \in [n]$.

Condition (1.68). To establish the uniform convergence of

$$\widehat{\left(\frac{\partial\mu}{\partial\theta_k}\right)}(\hat{\theta},p) := \frac{\hat{\dot{q}}_{\mu,k}(\hat{\theta},p)\hat{h}(\hat{\theta},p) - \hat{q}_\mu(\hat{\theta},p)\hat{\dot{h}}_k(\hat{\theta},p)}{(\hat{h}(\hat{\theta},p))^2} \tag{1.83}$$

to

$$\frac{\partial\mu}{\partial\theta_k}(\hat{\theta},p) := \frac{\dot{q}_{\mu,k}(\hat{\theta},p)h(\hat{\theta},p) - \hat{q}_\mu(\hat{\theta},p)\dot{h}_k(\hat{\theta},p)}{(\hat{h}(\hat{\theta},p))^2} \tag{1.84}$$

in $p \in A_n$, we can follow the same steps which led to (1.71)–(1.74) of Condition (1.67), because Assumptions 1.6(iv) and 1.7 ensure that with probability tending to one, $\|h\|_{A_n}$, $\left\|\dot{h}_k\right\|_{A_n}$, $\|q_\mu\|_{A_n}$, $\|\dot{q}_{\mu,k}\|_{A_n}$ are all bounded. As (1.73) and (1.74) were proved above, it is sufficient to show both

$$\left\|\hat{\dot{q}}_{\mu,k} - \dot{q}_{\mu,k}\right\|_{A_n} \xrightarrow{P} 0 \quad \text{and} \tag{1.85}$$

$$\left\|\hat{\dot{h}}_k - \dot{h}_k\right\|_{A_n} \xrightarrow{P} 0. \tag{1.86}$$

Condition (1.68)*, part* (1.85). We proceed by showing

$$\mathbb{E}\mathbb{E}\left[\sup_{p\in A_n}\left|\hat{\dot{q}}_{\mu,k}(\hat{\theta},p) - \mathbb{E}\left[\hat{\dot{q}}_{\mu,k}(\hat{\theta},p)\,\middle|\,\hat{\theta}\right]\right|\,\middle|\,\hat{\theta}\right] \to 0 \quad \text{and} \tag{1.87}$$

$$\sup_{p\in A_n}\left|\mathbb{E}\left[\hat{\dot{q}}_{\mu,k}(\hat{\theta},p)\,\middle|\,\hat{\theta}\right] - \dot{q}_{\mu,k}(\hat{\theta},p)\right| \xrightarrow{P} 0. \tag{1.88}$$

As for $\hat{q}'_\mu$ above,

$$\hat{\dot{q}}_{\mu,k}(\hat{\theta},p) = \frac{(g^{-1})'(p)}{n\gamma_n^2}\sum_{j\in[n]} Y_j X_{j,k}(2\pi)^{-1}\int(-it)e^{-it(\hat{\theta}^\intercal X_j - g^{-1}(p))/\gamma_n}\psi(t)\mathrm{d}t \tag{1.89}$$

$$= (g^{-1})'(p)(2\pi)^{-1}\int\left(\frac{1}{n}\sum_{j\in[n]} Y_j X_{j,k} e^{-it\hat{\theta}^\intercal X_j}\right)e^{itg^{-1}(p)}(-it)\psi(\gamma_n t)\mathrm{d}t. \tag{1.90}$$

Since $|-it| \leq |t|$ and $\mathcal{X}$ is bounded by Assumption 1.10, Euler's formula and Assumption 1.9(ii) imply that the bound of (1.75) also apply here up to a constant, which bounds $\mathcal{X}$, times $\sup_{p\in A_n}|(g^{-1})'(p)|$. As $\sup_{p\in A_n}|(g^{-1})'(p)| \leq 1/\left\|g'\right\|_\infty < \infty$ by Assumption 1.11, (1.87) holds.

Condition (1.68)*, part* (1.85)*,* (1.88). We begin by deriving

$$\dot{q}_{\mu,k}(\hat{\theta},p) = ((\partial/\partial\theta_k)(\mu h))(\hat{\theta},p).$$

For simplicity, assume that we only have two covariates, both continuous, and we are interested in the derivative with respect to the first coordinate of θ ($k \coloneqq 1$). It is straightforward to generalise the arguments below for the general case. By the tower property of expectation,

$$\mu(\theta,p) = \mathbb{E}\left[Y \mid \theta^\intercal X = g^{-1}(p)\right] = \mathbb{E}\left[m(X) \mid \theta^\intercal X = g^{-1}(p)\right].$$

In view of (1.30) of Proposition 1.6, we can write

$$\mathbb{E}\left[m(X) \mid \theta^\intercal X = t\right] = \frac{1}{v(\theta,t)}\int_{\mathcal{X}_1} m\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right)\Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right)\mathrm{d}x_1 \tag{1.91}$$

$$v(\theta,t) \coloneqq \int_{\mathcal{X}_1}\Psi\left(x_1, \frac{t-\theta_1 x_1}{\theta_2}\right)\mathrm{d}x_1, \tag{1.92}$$

where $v(\theta,\cdot)$ is the density of $\theta^\intercal X$ satisfying $v(\theta,\cdot) > 0$ by Assumption 1.10. Thus, $h(\theta,\cdot)$, being the density of $g(\theta^\intercal X)$, is equal to

$$h(\theta,p) = (g^{-1})'(p)\int_{\mathcal{X}_1}\Psi\left(x_1, \frac{g^{-1}(p)-\theta_1 x_1}{\theta_2}\right)\mathrm{d}x_1. \tag{1.93}$$

By Assumption 1.9(ii), (1.91) and (1.93) remain valid once we replace θ with $\hat{\theta}$. It then follows for continuously differentiable Ψ and m (Assumptions 1.10 and 1.11), that for $k = 1$,

$$q_\mu(\theta, p) = \mu(\theta, p)h(\theta, p)$$
$$= (g^{-1})'(p) \int_{\mathcal{X}_1} m\left(x_1, \frac{g^{-1}(p) - \theta_1 x_1}{\theta_2}\right) \Psi\left(x_1, \frac{g^{-1}(p) - \theta_1 x_1}{\theta_2}\right) \mathrm{d}x_1$$
$$\dot{q}_{\mu,k}(\hat{\theta}, p) = (g^{-1})'(p) \int_{\mathcal{X}_1} \left\{ \frac{\mathrm{d}}{\mathrm{d}\theta_1} \left[m\left(x_1, \frac{g^{-1}(p) - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right) \right.\right. \tag{1.94}$$
$$\left.\left. \times \Psi\left(x_1, \frac{g^{-1}(p) - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right) \right] \right\} \mathrm{d}x_1$$
$$= \frac{\mathrm{d}}{\mathrm{d}p} \int_{\mathcal{X}_1} -x_1 m\left(x_1, \frac{g^{-1}(p) - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right) \Psi\left(x_1, \frac{g^{-1}(p) - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right) \mathrm{d}x_1, \tag{1.95}$$

where in the last step we used that $(g^{-1})' > 0$. We proceed by showing the desired convergence. Let $[\underline{t}_{\hat{\theta}}, \bar{t}_{\hat{\theta}}] \coloneqq [g^{-1}(\underline{p}_{\hat{\theta}}), g^{-1}(\bar{p}_{\hat{\theta}})]$. We have for $k = 1$ by the tower property

$$\mathbb{E}\left[\hat{\dot{q}}_{\mu,k}(\hat{\theta}, p) \,\middle|\, \hat{\theta} \right] = \frac{g^{-1}(p)}{\gamma_n^2} \mathbb{E}\left[Y_j X_{j,k} K'((\hat{\theta}^\intercal X_j - g^{-1}(p)/\gamma_n) \,\middle|\, \hat{\theta} \right]$$
$$= \frac{(g^{-1})'(p)}{\gamma_n^2} \mathbb{E}\left[\mathbb{E}\left[m(X_j) X_{j,1} \mid \hat{\theta}^\intercal X_j, \hat{\theta} \right] K'((\hat{\theta}^\intercal X_j - g^{-1}(p))/\gamma_n) \,\middle|\, \hat{\theta} \right]$$
$$= \frac{(g^{-1})'(p)}{\gamma_n^2}$$
$$\times \int_{\underline{t}_{\hat{\theta}}}^{\bar{t}_{\hat{\theta}}} K'\left(\frac{t - g^{-1}(p)}{\gamma_n}\right) \int_{\mathcal{X}_1} x_1 m\left(x_1, \frac{t - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right) \Psi\left(x_1, \frac{t - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right) \mathrm{d}x_1 \, \mathrm{d}t$$
$$= (g^{-1})'(p) \int_{\mathcal{X}_1} \left\{ x_1 \right.$$
$$\left. \times \left[\frac{1}{\gamma_n^2} \int_{\underline{t}_{\hat{\theta}}}^{\bar{t}_{\hat{\theta}}} K'\left(\frac{t - g^{-1}(p)}{\gamma_n}\right) m\left(x_1, \frac{t - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right) \Psi\left(x_1, \frac{t - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right) \mathrm{d}t \right] \right\} \mathrm{d}x_1,$$

where we used that (1.91) holds not only for m, but for any generic function of X by Proposition 1.6. Let $\lambda_n(x_1, t) \coloneqq m\left(x_1, \frac{t - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right) \Psi\left(x_1, \frac{t - \hat{\theta}_1 x_1}{\hat{\theta}_2}\right)$. We show that the term in the square brackets converges to

$$-\frac{\partial}{\partial t} \lambda_n(x_1, g^{-1}(p)) = -\left(\frac{\mathrm{d}}{\mathrm{d}p} \lambda_n(x_1, g^{-1}(p)) \right) / [(g^{-1})'(p)]$$

uniformly in $x_1 \in \mathcal{X}_1$ and $p \in A_n$, which completes the proof for (1.88) in light of (1.95). Put $\lambda_n'(x_1,t) \coloneqq \frac{\partial}{\partial t}\lambda_n(x_1,t)$. Integration by parts gives

$$\frac{1}{\gamma_n^2}\int_{\underline{t}_{\hat{\theta}}}^{\bar{t}_{\hat{\theta}}} K'\left(\frac{t-g^{-1}(p)}{\gamma_n}\right)\lambda_n(x_1,t)\mathrm{d}t \tag{1.96}$$

$$= \frac{1}{\gamma_n}\int_{(\underline{t}_{\hat{\theta}}-g^{-1}(p))/\gamma_n}^{(\bar{t}_{\hat{\theta}}-g^{-1}(p))/\gamma_n} K'(v)\lambda_n(x_1,\gamma_n v+g^{-1}(p))\mathrm{d}t$$

$$= \frac{1}{\gamma_n}\left\{\lambda_n(x_1,\bar{t}_{\hat{\theta}})K((\bar{t}_{\hat{\theta}}-g^{-1}(p))/\gamma_n) - \lambda_n(x_1,\underline{t}_{\hat{\theta}})K((\underline{t}_{\hat{\theta}}-g^{-1}(p))/\gamma_n)\right\} \tag{1.97}$$

$$- \int_{(\underline{t}_{\hat{\theta}}-g^{-1}(p))/\gamma_n}^{(\bar{t}_{\hat{\theta}}-g^{-1}(p))/\gamma_n} \lambda_n'(x_1,\gamma_n v+g^{-1}(p))K(v)\mathrm{d}v. \tag{1.98}$$

Now λ_n is bounded with probability tending to one by Assumption 1.10. Recall that $\underline{t}_{\hat{\theta}} = g^{-1}(\underline{p}_{\hat{\theta}}), \bar{t}_{\hat{\theta}} = g^{-1}(\bar{p}_{\hat{\theta}})$. As g^{-1} is increasing, and K reaches its maximum at zero and satisfies $\lim_{|u|\to\infty} K(u) \to 0$, (1.97) is of the order

$$\gamma_n^{-1}\left[\sup_{p\in A_n} K\left(\frac{g^{-1}(\bar{p}_{\hat{\theta}})-g^{-1}(p)}{\gamma_n}\right) + \sup_{p\in A_n} K\left(\frac{g^{-1}(\underline{p}_{\hat{\theta}})-g^{-1}(p)}{\gamma_n}\right)\right]$$

$$= \gamma_n^{-1}\left[K\left(\frac{g^{-1}(\bar{p}_{\hat{\theta}})-g^{-1}(\bar{p}_{\hat{\theta}}-a_n)}{\gamma_n}\right) + K\left(\frac{g^{-1}(\underline{p}_{\hat{\theta}})-g^{-1}(\underline{p}_{\hat{\theta}}+a_n)}{\gamma_n}\right)\right]. \tag{1.99}$$

By a mean-value expansion, the first term in (1.99) is of the order

$$\gamma_n^{-1}K\left(\left(\inf_{p\in[\underline{p}_{\hat{\theta}},\bar{p}_{\hat{\theta}}]}(g^{-1})'(p)\right)a_n/\gamma_n\right).$$

As $\inf_{p\in[\underline{p}_{\hat{\theta}},\bar{p}_{\hat{\theta}}]}(g^{-1})'(p) > 0$, this is $o_P(1)$ for $a_n/\gamma_n \to \infty$ and Gaussian K. The same applies to the second term of (1.99), so (1.97) is $o_P(1)$. Last, we show that (1.98) converges to $-\lambda_n'(x_1,g^{-1}(p))$. Again, we can write

$$\lambda_n'(x_1,g^{-1}(p)) = \frac{\lambda_n'(x_1,g^{-1}(p))}{\gamma_n}$$

$$\times\left\{\int_{-\infty}^{g^{-1}(p)-\bar{t}_{\hat{\theta}}} K(u/\gamma_n)\mathrm{d}u + \int_{g^{-1}(p)-\bar{t}_{\hat{\theta}}}^{p-\underline{t}_{\hat{\theta}}} K(u/\gamma_n)\mathrm{d}u + \int_{g^{-1}(p)-\underline{t}_{\hat{\theta}}}^{\infty} K(u/\gamma_n)\mathrm{d}u\right\}.$$

A change of variables in (1.98) ($u := -\gamma_n v$) and K being symmetric about zero give

$$\sup_{x_1\in\mathcal{X}_1,p\in A_n}\left|\int_{(\underline{t}_{\hat{\theta}}-g^{-1}(p))/\gamma_n}^{(\bar{t}_{\hat{\theta}}-g^{-1}(p))/\gamma_n}\lambda_n'(x_1,\gamma_n v+g^{-1}(p))K(v)\mathrm{d}v-\lambda_n'(x_1,g^{-1}(p))\right|$$

$$\leq \sup_{x_1\in\mathcal{X}_1,p\in A_n}\left|\int_{-\infty}^{g^{-1}(p)-\bar{t}_{\hat{\theta}}}\lambda_n'(x_1,g^{-1}(p))\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \tag{1.100}$$

$$+\sup_{x_1\in\mathcal{X}_1,p\in A_n}\left|\int_{g^{-1}(p)-\underline{t}_{\hat{\theta}}}^{\infty}\lambda_n'(x_1,g^{-1}(p))\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \tag{1.101}$$

$$+\sup_{x_1\in\mathcal{X}_1,p\in A_n}\left|\int_{g^{-1}(p)-\bar{t}_{\hat{\theta}}}^{g^{-1}(p)-\underline{t}_{\hat{\theta}}}[\lambda_n'(x_1,g^{-1}(p)-u)-\lambda_n'(x_1,g^{-1}(p))]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right|. \tag{1.102}$$

Since $\mathcal{X}_1$ is bounded by Assumption 1.10, $\sup_{x_1\in\mathcal{X}_1,p\in A_n}|\lambda_n'(x_1,g^{-1}(p))|$ is bounded with probability tending to one. Hence, as g^{-1} is increasing, $K\geq 0$ and $\bar{t}_{\hat{\theta}}=g^{-1}(\bar{p}_{\hat{\theta}})$, (1.100) is of the order $\int_{-\infty}^{(g^{-1}(\bar{p}_{\hat{\theta}}-a_n)-g^{-1}(p))/\gamma_n}K(v)\mathrm{d}v$. Here,

$$(g^{-1}(\bar{p}_{\hat{\theta}}-a_n)-g^{-1}(p))/\gamma_n\leq(\inf_{p\in[0,1]}(g^{-1})'(p))(-a_n/\gamma_n)\leq(1/\left\|g'\right\|_\infty)(-a_n/\gamma_n)$$

by a mean-value expansion. As the infimum is positive, (1.100) is $o_P(1)$ for $a_n/\gamma_n\to\infty$, and so is (1.101) by similar arguments. The term (1.102) can be treated analogously to (1.80). First note that taking the supremum over $p\in A_n$ is the same as taking the supremum over $t\in[g^{-1}(\underline{p}_{\hat{\theta}}+a_n),g^{-1}(\bar{p}_{\hat{\theta}}-a_n)]=:T_n$ as g^{-1} is increasing. Take a $\gamma_n\ll\rho_n<(1/\left\|g'\right\|_\infty)a_n$ (e.g. $\rho_n:=(1/\left\|g'\right\|_\infty)\kappa_2 n^{-(\alpha+\beta)/2}$ for some $0<\kappa_2<\kappa_1$). Then (1.102) is bounded by

$$\sup_{x_1\in\mathcal{X}_1,t\in T_n}\left|\int_{[t-\bar{t}_{\hat{\theta}},t-\underline{t}_{\hat{\theta}}]\cap\{|u|\leq\rho_n\}}[\lambda_n'(x_1,t-u)-\lambda_n'(x_1,t)]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right| \tag{1.103}$$

$$\sup_{x_1\in\mathcal{X}_1,t\in T_n}\left|\int_{[t-\bar{t}_{\hat{\theta}},t-\underline{t}_{\hat{\theta}}]\cap\{|u|>\rho_n\}}[\lambda_n'(x_1,t-u)-\lambda_n'(x_1,t))]\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u\right|. \tag{1.104}$$

Here, (1.103) is bounded by

$$\sup_{x_1\in\mathcal{X}_1,t\in T_n}\sup_{|u|\le\rho_n}|\lambda_n'(x_1,t-u)-\lambda_n'(x_1,t)|\sup_{x_1\in\mathcal{X}_1,t\in T_n}\int_{-\infty}^{\infty}\gamma_n^{-1}K(u/\gamma_n)\mathrm{d}u$$
$$\le\sup_{x_1\in\mathcal{X}_1,t\in T_n}\sup_{|u|\le\rho_n}|\lambda_n'(x_1,t-u)-\lambda_n'(x_1,t)|.$$

By Assumptions 1.10 and 1.11, $\lambda_n'(x_1,\cdot)$ is continuous uniformly in x_1 and is therefore uniformly continuous on the compact set

$$[g^{-1}(\underline{p}_{\hat{\theta}}),g^{-1}(\bar{p}_{\hat{\theta}})]\supset[g^{-1}(\underline{p}_{\hat{\theta}}+a_n),g^{-1}(\bar{p}_{\hat{\theta}}-a_n)].$$

Thus, $|u|\le\rho_n$ for a small enough ρ_n implies that $t-u,t\in[g^{-1}(\underline{p}_{\hat{\theta}}),g^{-1}(\bar{p}_{\hat{\theta}})]$ for all $t\in T_n$. (Note that ρ_n is small enough if and only if $|u|\le\rho_n$ implies

$$|u|\le\min\left\{g^{-1}(\underline{p}_{\hat{\theta}}+a_n)-g^{-1}(\underline{p}_{\hat{\theta}}),g^{-1}(\bar{p}_{\hat{\theta}})-g^{-1}(\bar{p}_{\hat{\theta}}-a_n)\right\}.$$

By a mean-value expansion, the right side is smaller than or equal to $\inf_{p\in[0,1]}(g^{-1})'(p)a_n\le(1/\left\|g'\right\|_\infty)a_n$. But then $\rho_n<(1/\left\|g'\right\|_\infty)a_n$ is small enough.) As a consequence,

$$\sup_{x\in\mathcal{X}_1,t\in T_n}\sup_{|u|\le\rho_n}|\lambda_n'(x_1,t-u)-\lambda_n'(x_1,t)|=o_P(1),$$

hence (1.103) is $o_P(1)$. As $2\sup_{x_\in\mathcal{X}_1,t\in[\underline{t}_{\hat{\theta}},\bar{t}_{\hat{\theta}}]}|\lambda_n(x_1,t)|$ is bounded by probability tending to one as $\hat{\theta}_2\xrightarrow{P}\theta_{0,2}\neq0$ by Assumptions 1.10 and 1.11, (1.104) can be shown to be $o_P(1)$ similarly to (1.82). Thus, (1.102) is $o_P(1)$ which shows the desired convergence of (1.98). Hence, (1.88) holds for $k=1$. The case for $k=2$ follows by writing

$$\mathbb{E}\left[m(X)\mid\theta^\intercal X=t\right]=(1/v(\theta,t))\int_{\mathcal{X}_2}m\left(\frac{t-\theta_2x_2}{\theta_1},x_2\right)\Psi\left(\frac{t-\theta_2x_2}{\theta_1},x_2\right)\mathrm{d}x_2$$
$$v(\theta,t):=\int_{\mathcal{X}_2}\Psi\left(\frac{t-\theta_2x_2}{\theta_1},x_2\right)\mathrm{d}x_2$$

instead of (1.91) and (1.92). By Assumption 1.10, these behave equally well. Conclude that the desired convergence of $\hat{q}_{\mu,k}$ (Condition (1.68), part (1.85)) holds.

Condition (1.68)*, part* (1.86). Follows along the same lines as (1.85) by setting $Y_j:=1$ for all $j\in[n]$. Conclude that Condition (1.68) holds.

Remaining Uniform Consistency Results. Showing

$$\sup_{p \in A_n} |\hat{\mu}(\hat{\theta}, p) - \mu(\hat{\theta}, p)| \xrightarrow{P} 0, \tag{1.105}$$

$$\sup_{p \in A_n} |\hat{\mu}_2(\hat{\theta}, p) - \mu_2(\hat{\theta}, p)| \xrightarrow{P} 0 \tag{1.106}$$

completes the proof of Proposition 1.7. As $h(\hat{\theta}, \cdot)$ is bounded away from zero with probability tending to one by Assumption 1.6, (1.105) is implied by (1.73) and (1.74). Likewise, (1.106) is implied by $\|\hat{q}_{\mu_2} - q_{\mu_2}\|_{A_n} \xrightarrow{P} 0$, which can be shown as (1.73). ■

Chapter 2

Combining Experimental and Observational Data: The APOLLO Trial

Abstract

In the APOLLO trial (Tol et al., 2022, 2024), we inferred the causal effects of two hip fracture treatments, Posterolateral Approach (PLA) and Direct Lateral Approach (DLA), on health outcomes of patients in the Netherlands. The starting point of the inference was a Randomised Experiment (RE), where patients were randomly assigned to PLA or DLA, independently of their baseline characteristics. In addition, data from a 'Natural Experiment' (NE, or observational data) were also collected, under the plausible assumption that therein the allocation to PLA or DLA can be considered as good as random conditional on the patients' baseline characteristics. We estimated the average treatment effects of DLA versus PLA in the RE and the NE data separately, using a flexible and asymptotically efficient estimation strategy. We found no significant difference between PLA and DLA in any of the RE or the NE datasets. To improve the precision of the inference by increasing the sample size, we tested whether the RE and the NE datasets can be combined. Having found no evidence against combination, we estimated the average treatment effect on the combined dataset as well. Despite the improved precision, there was still no significant difference between PLA and DLA. Our conclusions were weakened by missing data, but they proved to be robust to our approach in handling the missingness and to our estimation strategy.

2.1. Introduction

Hip fracture imposes a nonnegligible health burden worldwide (Johnell and Kanis, 2004), hence it is desirable to compare medical treatments thereof. When feasible, medical treatments are compared in a randomised experiment, where patients are randomly allocated to the treatment arms. It is well-known that increasing the number of recruited patients improves the accuracy of the comparison by decreasing the length of confidence intervals. Thus, if additional data are also available, one may wish to combine them with the data from the randomised experiment. In this chapter, whose basis is the APOLLO trial (Tol et al., 2022, 2024), we compare two hip fracture treatments combining data from a Randomised Experiment (RE) and from a 'Natural Experiment' (NE or, equivalently, observational data), wherein the allocation of patients to treatments can only be considered random conditional on additional information.

Our contributions are published in Tol et al. (2022, 2024); this chapter is a more detailed discussion of the statistical methods employed therein. The relevant medical and administrative background is given in *ibid*.

In the remainder of this section, we describe briefly the APOLLO trial (Section 2.1.1), the estimand of interest quantifying the superiority of the medical treatments under consideration (Section 2.1.2), and how the estimand is identified from the RE and NE (Section 2.1.3). In Section 2.2, we describe the inferential method with special regard to the combination of the RE and NE datasets, and the issues faced with during inference, such as missing data. Section 2.3 contains the results, the estimates of the treatment effects. In Section 2.4, we consider the limitations of our approach and propose future research directions, only to conclude in Section 2.5.

2.1.1. The APOLLO Trial

The aim of the APOLLO trial was to compare femoral neck fracture (hip fracture) treatments. Set up in the Netherlands, the trial recruited patients who suffered femoral neck fracture and were recommended cemented hemiarthroplasty as a surgical treatment thereof according to the Dutch national guide-

lines. The two main approaches for cemented hemiarthroplasty in the Netherlands are the Posterolateral Approach (PLA) and Direct Lateral Approach (DLA). While it is beyond the scope of this chapter to precisely describe these treatments (medical details are given in Tol et al. (2024, Supplement 1, Section 5)), we note that the two approaches differ in the muscles which are divided during the surgery. The absence of conclusive evidence for the superiority of PLA or DLA supplied the motivation for the trial, whose aim was to compare PLA with DLA.

Although the study population was defined as patients over the age of 18 years who suffered hip fracture, hip fracture mostly affects elderly people: indeed, the mean age of participants was 82 years in both the randomised and the natural experiment with standard deviations around 7 years (Tol et al., 2024, Table on p. 5). The trial was conducted in 14 hospitals; in exactly 9 of these, at each site, the surgeons were qualified to perform *either* PLA or DLA. Patients were admitted to one of 14 sites based on their geographic location: the site of admission was the closest one to the patient's location.

When a patient was admitted to one of the 5 sites providing both PLA and DLA, they were completely randomised — irrespective of their baseline characteristics — to exactly one of PLA or DLA treatments in a 1:1 ratio, provided they consented to participating in the trial. These patients constituted the RE sample.

When a patient was admitted to one of the other 9 PLA-only or DLA-only sites, the patient was allocated to PLA or DLA based on the PLA or DLA specification of the site. Hence, if a patient was closer to a PLA-only site, then they were allocated to PLA, and likewise for DLA. The patients in these 9 sites consenting to the trial participation constituted the NE sample.

2.1.2. Estimand

We denote with D the treatment indicator, letting $D = 0$ correspond to PLA and $D = 1$ to DLA. Our aim is to infer the causal effect of the treatments on various health outcomes listed in Table 2.1. The primary outcome of interest is a health index called EQ-5D-5L, which is constructed from scores of different aspects

of health. In addition, two secondary outcomes, a self-reported assessment of health (EQ-VAS), and fear of falling (FES-I) are analysed.[1] All outcomes considered in this chapter were measured at 6 months following the treatment.

To formalise causality, we adopt the potential outcome framework of Neyman (1924) and Rubin (1974). The potential outcomes are $Y^0 \in \mathbb{R}$ and $Y^1 \in \mathbb{R}$, with Y^0 being the outcome of a given patient under PLA, and Y^1 being the outcome of the patient under DLA. The causal parameter of interest is the Average Treatment Effect (ATE) defined as

$$\tau := \mathbb{E}\left[Y^1 - Y^0\right].$$

Take the outcome EQ-5D-5L as an example: a higher τ means that DLA results in a better health on average, compared to PLA.

Table 2.1: Outcomes of Interest (Tol et al. (2024, eTable 1))

Name	Type (range)	Short Description
EQ-5D-5L	continuous (-0.446,1)	EuroQol Group-5-Dimension Questionnaire, composite health index (higher=healthier)
EQ-VAS	continuous (0,100)	EuroQol Visual Analog Scale, self-reported score of feeling healthy (higher=healthier)
FES-I	continuous (16,64)	fear of falling on Falls Efficacy Scale International (FES-I; higher=more fear)

[1] See the EuroQol website for more information. In Tol et al. (2024) we consider further outcomes as well. For simplicity of exposition, to avoid multiple testing errors, and because of the small variance of these outcomes, we restrict our attention in this chapter to the outcomes in Table 2.1.

2.1.3. Identification

The main challenge of causal inference is that for a given patient we only observe the outcome $Y = (1 - D)Y^0 + DY^1$, which is the potential outcome Y^0 if the patient is treated by PLA, and is Y^1 if treated by DLA. In general, a simple comparison

$$\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0] = \mathbb{E}\left[Y^1 \,\middle|\, D = 1\right] - \mathbb{E}\left[Y^0 \,\middle|\, D = 0\right]$$

of the mean outcome between PLA and DLA treated patients will not equal τ as the potential outcomes Y^d need not be independent of the treatment D. This is called confounding bias, arising from potential, systematic differences between PLA and DLA patients, such as their baseline health status. To account for this bias, the following baseline characteristics of patients are measured before any treatment is administered: Age, Gender, BMI, Living Status, Baseline ASA Scores, Mobility, Dementia; see Tol et al. (2024, eTable 1 and Table on p. 5) for more information. Collecting these characteristics in a vector $X \in \mathbb{R}^K$, $K \coloneqq 7$, we assume that

$$Y^d \perp\!\!\!\perp D \mid X, R = 0 \text{ and } \mathbb{P}(D = d \mid X = x) > 0 \text{ for all } d \in \{0, 1\} \text{ and } x \in \mathcal{X}, \tag{2.1}$$

where $\perp\!\!\!\perp$ denotes statistical independence, $\mathcal{X}$ is the support of X, and R is the randomised experiment indicator with $R = 0$ corresponding to the natural, and $R = 1$ to the randomised experiment.

The first condition of (2.1) is known as unconfoundedness, stating that the covariates X account for all the systematic differences between patients in the PLA versus the DLA group. In the randomised experiment, D was randomised according to a 1:1 allocation ratio to PLA and DLA so that $(Y^d, X) \perp\!\!\!\perp D \mid R = 1$, whereby the unconfoundedness $Y^d \perp\!\!\!\perp D \mid X, R = 1$ holds by design. In the natural experiment, a patient was allocated to PLA or DLA based on their geographic location as described in Section 2.1.1. We assumed that conditional on the covariates X, this allocation could be considered as good as random, implying unconfoundedness. This corresponds to assuming that the geographic location of a patient is noninformative about their outcomes Y^d once the patient's health status is accounted for by the covariates X. Although

there might be differences between hospitals in different geographic locations, which could threaten unconfoundedness (2.1), we assumed that these are not systematic and believe that accounting for the health status of patients' renders (2.1) rather plausible.

The second, overlap condition in (2.1) stipulates that in every region of $\mathcal{X}$, there is a positive probability of finding both PLA and DLA patients, ensuring that X can in fact be used to account for systematic differences between them. The covariates X appear to be balanced among the PLA and DLA arms, at least with respect to their mean and variance (Tol et al., 2024, Table on p. 5). Hence, at least, there appears to be no evidence that the overlap condition is violated.[2]

The conditions (2.1), together with the randomisation procedure in the randomised experiment, are sufficient to infer the ATE *either* from the randomised or the natural experiment alone (Lu et al., 2019, Theorems 1 and 2). However, to improve the accuracy of inference by increasing the sample size, we wish to combine the data from the RE and the NE. In order to do so, it is further assumed that

$$Y^d \perp\!\!\!\perp R \mid X \text{ for all } d \in \{0,1\}\,. \tag{2.2}$$

This combinability (or fusion) condition requires that the covariates X account for the systematic differences between patients who participate in the RE versus who participate in the NE. For instance, it would be violated if, even after accounting for X, patients in the RE were healthier as measured by Y^d than patients in the NE. The plausibility of (2.2) is further studied in Section 2.2.

Based on Lu et al. (2019, Theorem 3), we conclude that under the unconfoundedness and overlap conditions (2.1) and the combinability (2.2),

$$\tau = \mathbb{E}\left[Y^1 - Y^0\right] = \mathbb{E}\left[\mathbb{E}\left[Y \mid D=1, X\right] - \mathbb{E}\left[Y \mid D=0, X\right]\right], \tag{2.3}$$

so that ATE is identified from data on (Y, D, X). Notice that the identification (2.3) does not use data on the experiment indicator R, which is aligned with

[2]Note that $\mathbb{P}\left(D=d \mid X\right) = \frac{p(X \mid D=d)\mathbb{P}(D=d)}{p(X)}$, where $p(X \mid D=d)$ and $p(X)$ are the conditional and marginal density of X, respectively. Hence, a positive $p(X \mid D=d)$, that is, the balance of covariates, implies overlap because $\mathbb{P}\left(D=d\right) > 0$.

the essence of combining the data from the RE and the NE, ignoring their source. This is the consequence of the usual identification formula under unconfoundedness and overlap (2.1) — see Rubin (1974) for intuition — holding in both the RE and NE setting. Indeed,

$$\begin{aligned}\mathbb{E}\left[Y \mid D=d,R,X\right]=\mathbb{E}\left[Y^d\,\middle|\, D=d,R,X\right]=\mathbb{E}\left[Y^d\,\middle|\, R,X\right]\\ =\mathbb{E}\left[Y^d\,\middle|\, X\right],\ d\in\{0,1\},\end{aligned}\tag{2.4}$$

where the second equality is by (2.1) and third is by (2.2). Hence,

$$\mathbb{E}\mathbb{E}\left[Y \mid D=d,R,X\right]=\mathbb{E}\mathbb{E}\left[Y^d\,\middle|\, X\right]=\mathbb{E}Y^d,\ d\in\{0,1\},$$

yielding the identification (2.3).

2.2. Methods

In this section, we present our strategy to infer ATE. There are various methods aimed at the combination of RE and NE data, focusing on different aspects — see Colnet et al. (2023) and Lin et al. (2024) for reviews. Chen et al. (2021) derive convergence rates for ATE estimation in a Gaussian model when the unconfoundedness (2.1) is violated. Oberst et al. (2023) utilise the RE data, where unconfoundedness holds by design, to correct for its violation in the NE data. Dang et al. (2023) propose a method when multiple NE datasets are available for choosing the one that minimises the mean squared error of the ATE estimator. Thus, they may also include the NE data when unconfoundedness is violated as long as the bias stemming from the violation is offset by a decrease in variance. Yang et al. (2022) propose a pretest estimator: they test whether unconfoundedness holds in the NE setting with the help of the RE data, and combine the two data sources if and only if it holds. They show that this pretest estimator exhibits a limiting distribution which is a mixture.

What is common in the above methods is that they approach the question of combinability from the perspective of unconfoundedness (2.1) holding or not in the NE setting. In contrast, our point of departure is (2.1), which we believe to hold. As shown in (2.4), the condition (2.2) implies that the RE and NE data may be combined to identify ATE. Hence, we approach the question

of combinability from the perspective of (2.2): whether there are systematic differences in the health status of patients in the RE and NE setting, once accounting for the covariates, or not.

Suppose there are no such differences so that (2.2) holds. Then we wish to base the inference of ATE on an n-large random sample $\mathcal{S}_{\mathrm{NR}} \coloneqq ((Y_i, D_i, X_i, R_i))_{i\in[n]}$ from the distribution of (Y, D, X, R), which is the combination of the data of patients in the randomised experiment, $\mathcal{S}_{\mathrm{R}} \coloneqq ((Y_i, D_i, X_i))_{i\in[n]:R_i=1}$, and of those in the natural experiment, $\mathcal{S}_{\mathrm{N}} \coloneqq ((Y_i, D_i, X_i))_{i\in[n]:R_i=0}$. Consider the Augmented Inverse Propensity Weighted (AIPW) estimator of τ in Lu et al. (2019, Theorem 3),

$$\hat{\tau} \coloneqq \frac{1}{n}\sum_{i\in[n]}\left\{\frac{D_i}{\hat{\pi}_{\mathcal{X}}(X_i)}(Y_i - \hat{\mu}_{\mathcal{X}}(1, X_i)) - \frac{1-D_i}{1-\hat{\pi}_{\mathcal{X}}(X_i)}(Y_i - \hat{\mu}_{\mathcal{X}}(0, X_i)) + \hat{\mu}_{\mathcal{X}}(1, X_i) - \hat{\mu}_{\mathcal{X}}(0, X_i)\right\}, \quad (2.5)$$

where $\hat{\pi}_{\mathcal{X}}$ and $\hat{\mu}_{\mathcal{X}}$ are estimators of

$$\pi_{\mathcal{X}}(x) \coloneqq \mathbb{P}(D = 1 \mid X = x) \text{ and } \mu_{\mathcal{X}}(d, x) \coloneqq \mathbb{E}[Y \mid D = d, X = x], \quad d \in \{0, 1\}, x \in \mathcal{X},$$

the propensity score and the outcome regression, respectively, and $(\hat{\pi}_{\mathcal{X}}, \hat{\mu}_{\mathcal{X}})$ is constructed from $\mathcal{S}_{\mathrm{NR}}$ by cross-fitting.[3] Notice that (2.5) does not use data on the experiment indicator R_i, and so it coincides with the usual double robust estimator under unconfoundedness (Robins et al., 1994; Rotnitzky et al., 1998; Bang and Robins, 2005). Under conditions (2.1) and (2.2), the AIPW estimator (2.5) is asymptotically normal and efficient under regularity conditions and fast-enough convergence of the estimators $\hat{\pi}_{\mathcal{X}}$ and $\hat{\mu}_{\mathcal{X}}$ to their respective estimands, which is standard in ATE estimation (Lu et al., 2019, Theorem 3).

Importantly, the estimator (2.5) uses data from both the randomised and the natural experiment, hence it is asymptotically more efficient than estima-

[3] In cross-fitting, we first estimate $(\pi_{\mathcal{X}}, \mu_{\mathcal{X}})$ using a random half of $\mathcal{S}_{\mathrm{NR}}$, then evaluate $\hat{\tau}$ on the other half of $\mathcal{S}_{\mathrm{NR}}$. Afterwards, we exchange the roles of the two halves to obtain another estimate of τ; finally we take the average of the thus obtained two estimators of τ to get the final estimate. Employing cross-fitting helps to achieve efficiency while avoiding regularity (Donsker) conditions on the function classes of $(\pi_{\mathcal{X}}, \mu_{\mathcal{X}})$. See e.g. Kennedy (2023).

tors based only on the RE or the NE data (Lu et al., 2019, p. 7). However, there are two issues that need to be addressed before the estimator (2.5) can be applied. First, both the RE and the NE datasets have missing data. Second, the plausibility of the combinability condition (2.2) of the RE and NE datasets needs to be tested. We address these challenges in Section 2.2.1 and Section 2.2.2, respectively.

2.2.1. Missing Data

Missing data were prevalent in the primary outcome EQ-5D-5L (RE: 20%, NE: 35% approximately), the secondary outcomes EQ-VAS (RE, NE: 45% approximately) and FES-I (RE, NE: 50% approximately), and the covariate BMI (RE: 25%, NE: 10% approximately). The other covariates had a low prevalence of missing measurements, under 5%, except for Mobility, reaching 10% in the RE. There were no missing observations on the treatment. See Tol et al. (2024, eFigure 1) for more information.

Measurements in the outcomes were mainly thought to be missing due to nonresponse to the survey recording the self-reported scores from which the outcomes were constructed. Nonresponse was hypothesised to be driven by Dementia, Living Status, and Mobility. Regarding the primary outcome EQ-5D-5L, estimating a logistic model of the missingness indicator on the covariates, we found a significant positive association between Mobility and the missingness indicator in RE (p-value 0.002), and a significant positive association between Dementia and the indicator in NE (p-value 0.004). Therefore, we assumed that the primary outcome is Missing At Random given the covariates (MAR). Under MAR, there were two ways to proceed. As the covariates (partly) explain the missingness, we could use them to impute the outcome. Alternatively, we could perform a complete-outcome analysis: using data only from patients without missing outcome (see e.g. Carpenter and Smuk (2021) for arguments supporting such analysis under MAR). For simplicity, we primarily proceeded with this

latter complete-outcome analysis.[4] Correspondingly, we let $\bar{\mathcal{S}}_{\mathrm{R}}$, $\bar{\mathcal{S}}_{\mathrm{N}}$, $\bar{\mathcal{S}}_{\mathrm{NR}}$ denote subsets of $\mathcal{S}_{\mathrm{R}}$, $\mathcal{S}_{\mathrm{N}}$, $\mathcal{S}_{\mathrm{NR}}$, respectively, formed by removing patients with missing Y from $\mathcal{S}_{\mathrm{R}}$, $\mathcal{S}_{\mathrm{N}}$, $\mathcal{S}_{\mathrm{NR}}$. This left us with 161 PLA and 176 DLA patients in the RE ($\bar{\mathcal{S}}_{\mathrm{R}}$), and 68 PLA and 87 DLA patients in the NE ($\bar{\mathcal{S}}_{\mathrm{N}}$); see Tol et al. (2024, Figure 1). Regarding the secondary outcomes, we pursued the same strategy of complete-outcome analysis.

Regarding the covariates, we employed a multiple imputation (Rubin, 1987, 1996) algorithm called Multiple Imputation Chained Equations (MICE; see Van Buuren and Groothuis-Oudshoorn (2011)). To impute X_{ki}, the missing covariate k of patient i, first, MICE requires a specification of a predictive model for X_k given all the other covariates X_{-k}, *marginally* for each X_k, $k \in [K]$, with missing values. We opted for Classification And Regression Trees (CART, Breiman et al. (2017)), with classification for discrete X_k and regression for continuous X_k. Our choice was motivated by CART being a flexible method, which can capture nonlinear associations.

At the time of the data analysis, the outcome was not used in the prediction model for the missing covariates. This decision was motivated by the desire to preserve unconfoundedness, for it appeared intuitive that including outcome-information into the imputed covariates destroys their exogeneity. However, at the time of completing this thesis, we were pointed to[5] the work of Sterne et al. (2009) and White et al. (2011), (informally) advocating for the inclusion of the outcome. Their argument is that the omission of the outcome constitutes a loss of information, thereby introducing bias in further analysis involving the regression of the outcome on the imputed, and on the other, observed covariates. Investigating the question from the perspective of unconfoundedness, it appears that (the desired) unconfoundedness given the imputed covariate,

$$Y^d \perp\!\!\!\perp D \mid X_{-k}, \hat{X}_k, \tag{2.6}$$

where $\hat{X}_{ki}$ is X_{ki} if it is not missing, and is the imputed value otherwise, does *not* hold in general, regardless of whether the outcome Y is included in the

[4] Nevertheless, we recalculated our results with the outcome imputed as a form of sensitivity analysis, see Section 2.4.

[5] By Dr. ir. Richard A. J. Post, member of the present doctoral committee.

prediction model for X_k.[6] Thus, we conclude that the outcome should have been included, for it provides additional predictive power and its omission does not preserve unconfoundedness, contrary to our intuition at the time of the analysis. Our above inspection, at the same time, highlights that (2.6) may fail, even when the outcome is included; the consequences of which appears to be unclear, calling for a more detailed study.

With a predictive model at hand, MICE generates multiple imputed datasets. For instance, take the NE dataset $\bar{\mathcal{S}}_{\mathrm{N}}$ with observed outcome. Then MICE creates M datasets, $\bar{\mathcal{S}}_{\mathrm{N}}^{(1)}, \bar{\mathcal{S}}_{\mathrm{N}}^{(2)}, \ldots, \bar{\mathcal{S}}_{\mathrm{N}}^{(M)}$, each with possibly different imputed values for X_{ki} based on the CART model and (Gibbs) sampling (Van Buuren and Groothuis-Oudshoorn, 2011). Finally, the causal effect estimator (2.5) is evaluated for each of these datasets, yielding M values of the estimator and their corresponding M standard errors. These M estimates of τ and their standard errors are pooled into a single causal effect estimate and a standard error with Rubin's rule (Rubin, 1987, p. 76–77). The advantage of this approach is that it accounts for uncertainty in the imputation, at least in large samples.

2.2.2. Test of Combinability

To test the combinability condition (2.2) of the RE and the NE datasets, we followed Lu et al. (2019). In particular, we estimated the linear model

$$Y = \alpha_0 + \sum_{j=1}^{K} \alpha_k X_k + \alpha_{K+1} R + \varepsilon, \tag{2.7}$$

where ε is an error term, separately in the PLA and DLA treatment groups. Note that conditional on $D = d$, the outcome Y is the potential outcome Y^d, whence an $\alpha_{K+1} \neq 0$ implies that Y^d is associated with R after controlling for X. Therefore, we tested $H_0 : \alpha_{K+1} = 0$ against $H_1 : \alpha_{K+1} \neq 0$, with H_1 being an evidence against (2.2), implying that the two datasets should not be combined. To accommodate missing data, we began with the complete-outcome dataset

[6] A sufficient condition for (2.6) to hold when the outcome is not included is for X_k to carry no information about Y^d given X_{-k}, and likewise for the missingness indicator of X_k. That is, $Y^d \perp\!\!\!\perp (D, M_k, X_k) \mid X_{-k}$, where M_{ki} indicates whether X_{ki} is missing for patient $i \in [n]$.

$\bar{\mathcal{S}}_{\text{NR}}$, generated one imputed dataset $\bar{\mathcal{S}}_{\text{NR}}^{(0)}$ — imputing the missing values in the combined dataset according to Section 2.2.1 —, and then tested H_0 against H_1. At a significance level of 5%, H_0 was not rejected in both the PLA and the DLA groups, indicating no evidence against data fusion.[7]

2.3. Results

After tackling the missing data issues according to Section 2.2.1, and having found no evidence against data fusion (Section 2.2.2), we evaluated the AIPW estimator (2.5) on mutiple imputed datasets for the combined RE and NE data $\bar{\mathcal{S}}_{\text{NR}}$. In the AIPW estimator, we estimated both the propensity score $\pi_{\mathcal{X}}$ and the outcome regression $\mu_{\mathcal{X}}$ with an ensemble of methods featuring elastic net regression (Zou and Hastie, 2005), generalised linear models (Nelder and Wedderburn, 1972), generalised additive models (Hastie and Tibshirani, 1986), generalised boosted models (Friedman, 2001, 2002) and multivariate adaptive regression splines (Friedman, 1991). The advantage of ensemble estimation is that instead of resorting to an individual estimator such as elastic net, the ensemble estimator combines the individual estimators to improve their predictive accuracy (Van der Laan et al., 2007). By including both parametric models such as generalised linear models and nonparametric estimators such as splines, we further increase the accuracy of the estimators $\hat{\pi}_{\mathcal{X}}$ and $\hat{\mu}_{\mathcal{X}}$, for if the true models are indeed parametric, the convergence of $\hat{\pi}_{\mathcal{X}}$ and $\hat{\mu}_{\mathcal{X}}$ becomes faster by the inclusion.

Notice that while the AIPW estimator (2.5) does not use data on the experiment indicator R explicitly, it uses observations on R implicitly as data on R were used to test the combinability condition (2.2).

The estimates of ATE pooled from the multiple imputed datasets and their pooled standard errors (Section 2.2.1) are depicted in Figure 2.1. For comparison, separate estimates from NE-only and RE-only data are also shown. The

[7] The estimates of α_{K+1} (and their standard errors) for the primary outcome EQ-5D-5L are -0.0625 (0.0420) and 0.0103 (0.0353) in the PLA and DLA groups, respectively (Tol et al., 2024, eAppendix 2).

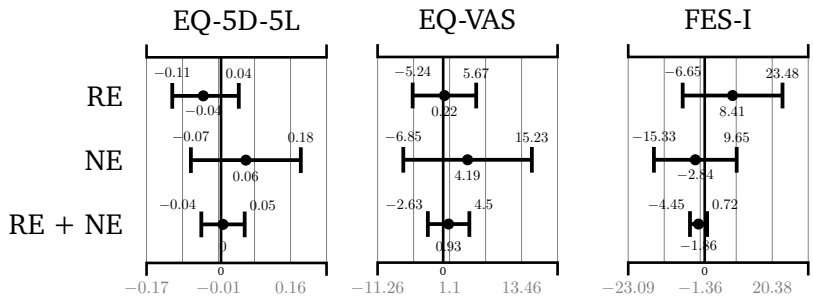


Figure 2.1: The estimated ATE for the primary and secondary outcomes, including 95% confidence intervals, for the RE, NE and the combined datasets. A positive ATE favors DLA over PLA for EQ-5D-5L and EQ-VAS, and PLA over DLA for FES-I. Tol et al. (2024, eTable 3, Figure 4).

separate estimates were obtained by estimating a linear model of the outcome on the treatment indicator, and accounting for additional variables including X and for the hospital where the treatment was performed. In Figure 2.1, it is clearly observed that the RE-only confidence intervals are narrower than the NE-only ones, as expected for a larger number of RE patients. Combining the RE and the NE data further improves accuracy as evidenced by even narrower confidence intervals. Notwithstanding the improved accuracy, the results in Figure 2.1 suggest no superiority of PLA or DLA for any of the studied outcomes.

2.4. Discussion

There are some limitations to our analysis, revolving around sample size and uncertainty quantification. First, missing data were prevalent in the primary outcome of interest and in some covariates used to account for confounding bias. Due to the predominantly elderly study population, missingness was attributed to nonresponse to the survey partly explained by dementia. To account for the missingness, we performed the analysis imputing the missing outcomes, only to find, again, no significant differences between PLA and DLA. We also imputed the missing covariates, resulting in no change in our conclusions, even though we did not include the outcome in the prediction model for the miss-

ing covariates at the time of the analysis — which appears to be a mistake in retrospect (see the discussion in Section 2.2.1). Nonetheless, our conclusions continued to hold when BMI, the covariate with the most missing measurements was excluded from the covariates.

Second, even in the absence of missing data, the sample size in the natural experiment was rather moderate, which could hinder the convergence of the flexible estimation methods we employed for the propensity score and the outcome regression, which are building blocks of the AIPW estimator. Nonetheless, our conclusions were robust to different estimation strategies of the propensity score and the outcome regression.

A potential approach to further mitigate these two issues could be the inclusion of further patients into the analysis. In addition to patients from the Netherlands, one could consider the combination of data sets from various countries; for example, data from studies listed in Van der Sijp et al. (2018, Table 1). In this case, special attention should be paid to the proper definition of the study population, and, thus, the estimand, as well as the conditions for combinability. For instance, methods capable of transporting the treatment effects from a source population to a target population (Colnet et al., 2023) could be employed to 'transfer' treatment effects from another country to the Netherlands. These combinations typically require conditions stipulating the similarity of the source and target populations along some attributes, such as complete conditional independence (2.2), or, weaker, the sameness of the conditional average function of the outcome given the covariates. Proceeding so would, however, necessitate further testing based on the estimates of these conditional mean functions, or testing for conditional independence. This, in turn, leads us to the third limitation of our approach concerning uncertainty quantification.

Namely, there are limitations regarding the testing for combinability (2.2) of the randomised and natural experiment data. Our test was a rather simple test of mean-independence as opposed to complete statistical independence stipulated by the condition (2.2). However, let us point out that mean-independence is, in fact, sufficient for the identification (2.3) in the light of (2.4). The limitation, then, consists in our assumption, following Lu et al.

(2019), of the linear model (2.7) for the conditional mean function. The linear model (2.4) is appealing because it lends itself for a straightforward testing of mean-independence, based on the estimated coefficient of the RE/NE indicator. It is, however, overly simplistic. In addition, although the inference of the treatment effect took into account the uncertainty in the imputation of missing data, it was not adjusted for the uncertainty in the combinability testing.[8] It seems unclear from the literature how to test (2.2), or its weaker counterpart of mean-independence without the linearity (2.7) restriction, and how uncertainty from this testing could be rigorously accounted for in our setting — let alone simultaneously accounting for uncertainty from the imputation. Even so, combining the RE and the NE data does not change our conclusion: the combined result of no superiority of any of PLA or DLA also applies separately in each of the RE and NE setting.

These limitations in combinability testing indicate potential directions for future research. Specifically, the development of methods to test (2.2) or mean-independence without the linearity (2.7) restriction, and to quantify the (additional) uncertainty from the testing of (2.2). For instance, one could adapt the unconfoundedness testing of Yang et al. (2022) to the testing of combinability. We conjecture that the limiting distribution of the estimated treatment effect would be a mixture of two distributions, corresponding to (2.2) holding and not holding, similarly to Yang et al. (2022).

Another potential direction for future research is indicated by the discussion in Section 2.2.1, on the imputation of missing covariates. Namely, we argued (informally) that unconfoundedness conditional on the imputed, instead of the observed, covariates does not hold in general. It may be worth to investigate whether this biases causal inference based on imputed covariates.

Finally, we may propose some recommendations for future empirical studies wishing to combine data from randomised and natural experiments. Suppose that the above issues with the pretesting of combinability is resolved. Even

[8] As indicated in Section 2.2.1, the method of Yang et al. (2022) takes into account the uncertainty of such pretesting, but they test for the violation of unconfoundedness (2.1) in the natural experiment as opposed to the combinability condition (2.2).

then, it seems rather unavoidable for empirical studies to have a sufficiently large sample size in both the RE and the NE arms, because combinability testing would likely involve mean-independence testing via the estimation of conditional mean (regression) functions. Unless one is willing to make parametric assumptions about the regressions, their flexible, nonparametric estimation requires large sample sizes. More specific to our current setting of hip fracture treatments is the recommendation to account for, in advance, the expected large extent of missingness deriving from the old age of the study population.

2.5. Conclusion

In this chapter, we inferred the causal effects of two hip fracture treatments, PLA and DLA, on health outcomes. The inference was based on data from the APOLLO trial (Tol et al., 2022, 2024): a randomised experiment and a 'natural experiment', which was considered as good as randomised conditional on baseline characteristics of patients. First, we estimated the average treatment effect of DLA versus PLA separately for the randomised and the natural experiments, utilising an asymptotically efficient Augmented Inverse Propensity Weighted (AIPW) estimator. Then we tested for systematic differences between patients in the randomised versus the natural experiment. Having found no evidence of such differences, we combined data from the randomised and the natural experiment to improve the accuracy of the treatment effect estimates by increasing the number of data points. We found that no treatment was superior to the other in terms of the investigated outcomes. This conclusion applies uniformly to the randomised and the natural experiment, and also to the combined dataset, despite the improved accuracy. While there are limitations to our analysis regarding sample size — stemming from the design of the trial and missing data — and uncertainty quantification — stemming from pretesting —, we believe that our conclusions are rather robust to these as supported by sensitivity analyses.

Chapter 3

Private Double Robust Inference

Abstract

Privacy mechanisms preserve the privacy of individuals in a sample by injecting noise into their sensitive data in a controlled manner, revealing only the noisy, privatised data to the statistician for inference purposes. The inference of a parameter exhibits a rate double robustness property when the large-sample bias of an estimator of the parameter is characterised by the product of the estimation errors of two other, auxiliary (or nuisance), often infinite-dimensional, parameters. We propose a novel class of rate double robust parameters whose novelty lies in the potentially nonlinear but smooth dependence on a low-dimensional regression parameter. Among others, this includes average treatment effects. We show that the properties of the sensitive-data model carry over to the privatised-data model by a suitable choice of the privacy mechanism, which, in general, means a total-variationally private mechanism. In particular, the double robustness property is retained, enabling efficient estimation from the privatised sample. We also find that the estimation of the nuisance parameters is not harder, albeit possibly less efficient and computationally more demanding, from the privatised sample compared to the sensitive sample for a given, suitable privacy mechanism. Indeed, if the estimation is feasible from the sensitive sample by some procedure, we can directly transport that procedure to the privatised setting. Lastly, we develop a private method of moments estimator for parametric models. This shows that in the private setting a parametric assumption about one nuisance parameter affords more flexible modelling and slower estimation of the other one, as is the well-known case in the nonprivate setting.

3.1. Introduction

The aim of statistics is to learn about unobservable quantities, called parameters, from observable data. In contrast, the aim of privacy-preserving methods is to hide the observable data in order to protect the units to whom the data belong. Balancing these two opposing aims is the content of this chapter.

For example, consider inferring effects of a treatment as causal parameters from observational data. To account for differences between units with different treatment status — that is, confounding —, one collects a set of variables, called covariates, which render the units comparable and thereby facilitate causal inference. However, the covariates may contain sensitive data or, based on a large enough number of covariates, specific units in the data set may become individually identifiable. It is therefore desirable to protect the covariates on the level of each unit. This is accomplished by what is called local privacy mechanism.

A local privacy mechanism deliberately injects noise into the covariates of each unit and reveals only the noisy, privatised data to the statistician, with the original, nonprivate, sensitive covariates remaining undisclosed in the hold of the respective units.[1] The task of the statistician is then to infer the parameters of interest of the sensitive-data distribution solely based on the noisy data. This is the task we address in this chapter: the inference of real-valued parameters, where parts of the data, to which we henceforth refer as covariates, are protected by a local, unit-level privacy mechanism. Inference remains feasible because the noise is introduced in a controlled manner and the privacy mechanism itself, but not the sensitive covariates, are known to the statistician.

Specifically, we focus on the inference of parameters which admit what is referred to as rate double robustness property. A parameter has the rate double robustness property if the asymptotic bias of an estimator thereof is

[1]The terminology 'privatised', 'private' and 'nonprivate' is somewhat ambiguous. By *privatised* (or *private*) data we mean the data which result from the injection of noise into the original data, which we refer to as *nonprivate data*. Thus, the distinction is understood with respect to the privacy mechanism, not the privacy-right claimentship.

characterised by the product of the estimation error of two, often infinite-dimensional, parameters. A well-known example is the average treatment effect. There, the bias is the product of errors in the estimation of the conditional mean of the outcome of interest given the covariates (outcome regression) and of the conditional probability of treatment given the covariates (propensity score). The rate double robustness property is attractive due to the product structure: a poorly estimated parameter may be compensated by a well-estimated other.

Our contribution is threefold. First, we propose a novel class of rate double robust parameters in the nonparametric model, building on the work of Rotnitzky et al. (2021) and Chernozhukov et al. (2022). The class comprises parameters that depend linearly on an infinite-dimensional regression function, as in Rotnitzky et al. (2021) and Chernozhukov et al. (2022), but also involve a possibly nonlinear but smooth dependence on a *low-dimensional* regression function evaluated at a point of its domain. Parameters in our class are double robust in a sense that their asymptotic bias is characterised by the product of the estimation errors of the infinite-dimensional regression and another infinite-dimensional nuisance parameter, called Riesz representer. The inclusion of a nonlinear dependence on the low-dimensional regression is a novelty compared to Rotnitzky et al. (2021), who consider only linear dependencies. How our class relates to that of Chernozhukov et al. (2022) is more nuanced, hence we defer its in-detail discussion to Section 3.2. Briefly, our proposed class intersects with those of Rotnitzky et al. (2021) and Chernozhukov et al. (2022), but there is no strict inclusion either way for either of them in terms of the double robust property and the nonlinear dependence.

Besides double robustness, another advantage of our class is that parameters therein naturally lend themselves for inference in the privacy-preserving setting. This follows from the fact that regressions play a vital role in connecting the nonprivate and the private data.

Second, we provide conditions for the privacy mechanism to maintain identifiability of the parameters of the unobserved, nonprivate-data distribution from the observed, private-data distribution. When the covariates are distributed on a finite set, a larger class of privacy mechanisms is admissible in general,

compared to generic but absolutely continuous covariates. Further, we connect the semiparametric properties of the private-data model to those of the nonprivate-data model. We show that if a parameter in the nonprivate-data model enjoys the rate double robustness property, then it continues to do so in the private-data model, provided the privacy mechanism is admissible.

Third, we study private estimation. We show that for admissible privacy mechanisms, the estimation of the parameters, including the infinite-dimensional ones in the double robust product structure, is, albeit possibly less efficient and computationally more demanding, not particularly harder in the private compared to the nonprivate setting. More precisely, the estimation of the infinite-dimensional parameters can be recast as optimisation problems. Provided these problems can be solved in the nonprivate setting, these solutions are exactly transferable to the private setting. We do not discuss any specific estimator in the infinite-dimensional case; nonetheless, we study the role of functional-form (parametric) assumptions in the double robust property. To this end, we develop a private method of moments estimator for *any* $\mathbb{R}^K$-valued parameter which is identified from regular-enough moment conditions in the nonprivate-data model. With this at our disposal, a smooth functional-form assumption for one infinite-dimensional parameter can be traded off for more flexible modelling, thus slower converging estimation, of the other one. Thus, the parametric double robustness is retained in the private setting.

In summary, to the best of our knowledge, our work is the first achieving private nonparametric rate double robust inference for parameters as general as the ones in our proposed class, and with data taking values in generic spaces.

Our results are illustrated with the average treatment effect and the average treatment effect on the treated. We show how they emerge as special cases of our proposed double robust class, ultimately arriving at the usual double robustness condition, featuring the outcome regression and the propensity score.

The rest of the chapter is organised as follows. In Section 3.2, we situate our work in the literature. In Section 3.3, we present our general framework and introduce some notation. Section 3.4 describes the proposed double robust class, and its semiparametric and inferential properties without privacy.

Section 3.5 introduces the privacy framework, and shows how the properties of the double robust class carry over to the private setting. Section 3.6 discusses the private estimation of the double robust parameter of interest. Section 3.7 focuses on the private estimation of parameters auxiliary to the estimation of the parameter of interest. Section 3.8 concludes.

3.2. Literature

Section 3.2.1 contains an in-detail discussion of how our results speak to the double robust literature. Section 3.2.2 describes advancements in the field of privacy-preserving inference and how our results relate to them.

3.2.1. Double Robust Inference

Two of the perhaps most encompassing contributions to double robust inference are due to Rotnitzky et al. (2021) and Chernozhukov et al. (2022). Compared to Rotnitzky et al. (2021), our class is novel since it includes a nonlinear dependence on a low-dimensional regression function. Rotnitzky et al. (2021) allow for linear dependence on parameters more general than regression functions. However, their sufficient conditions for a product-form bias (Rotnitzky et al., 2021, Proposition 3) stipulate a parameter structure where both factors in the product are ratios of two regressions, with the same denominator in both. As modelling regressions separately is arguably more intuitive than modelling their ratios, as is nonetheless performed in Smucler et al. (2019), this structure does not translate into a 'natural' rate double robustness property in general with the variationally dependent ratios in the product (common denominator). An exception, applying, for instance, to the average treatment effect, is when the denominator and the nominator are chosen so that the resulting ratio in each factor is itself a regression function. But then these parameters are strictly included in our class.

On the other hand, Chernozhukov et al. (2022) allow for nonlinear dependencies on multiple infinite-dimensional regressions. However, to achieve asymptotic efficiency, they require a convergence rate faster than $n^{-1/4}$ for

the estimation of *each* of these regressions, where n is the sample size (Chernozhukov et al., 2022, Assumption 14, Theorem 9). As they also state, this nonlinearity comes at the price of their parameter class ceasing to be double robust. By having one infinite-dimensional regression and one low-dimensional one, we still retain double robustness because we can estimate the low-dimensional parameter at root-n rate. Indeed, it shall be seen that a merely consistent estimation of the infinite-dimensional regression can suffice in our class as long as it is compensated for by a fast-enough estimation of the Riesz representer, and the other way around.

One example of interest included in our parameter class is the average treatment effect on the treated, which was already shown to be double robust by Chernozhukov et al. (2022, Example 6). Yet, they do not seem to cover the whole class of parameters with the linear infinite-dimensional and the nonlinear low-dimensional regressions in a doubly robust manner we do.

3.2.2. Privacy-Preserving Inference

As regards privacy protection, our chosen approach is *statistical privacy*. Generally, a distinction can be drawn between *cryptographic* and statistical privacy. Cryptographic privacy, resting on the premise that certain mathematical problems are very difficult to solve, is stronger than statistical privacy, which only delivers probabilistic guarantees. Before the breakthrough of Gentry (2009), it was impossible to perform statistical inference under cryptographic guarantees, and even since then it has remained challenging due to the limited number and type of operations that are supported on encrypted data (see Yang et al. (2019) for a survey). Partly because of this, and partly owing to the rather difficult mathematics involved in cryptographic constructions, attention to privacy-preserving statistical inference has mostly been focused on the less stringent but much more flexible statistical privacy paradigm. Although this paradigm dates back to Warner (1965), it is only since the work of Evfimievski et al. (2003) and Dwork et al. (2006) that it has attracted much attention in the literature. *Ibid* provide a mathematically more rigorous treatment of privacy, formulating the notion of the by-now-widespread differential privacy, defined

in terms of conditional likelihood ratios. Following them, a vast amount of formal definitions was introduced. In a major survey, Desfontaines and Pejó (2022) describe a daunting 255 types of privacy mechanisms, and how they relate to each other.

Initially, statistical privacy took the form of *central privacy*. Central (or global) privacy protection[2] is concerned with privacy at the level of sample and aggregate statistics, and corresponds to the notion of plausible deniability. For example, noising the sample mean to guarantee within probabilistic bounds that the presence of any given unit in the sample is deniable, especially units with outlier covariate values. Various topics have been addressed in the central paradigm. Smith (2008) addresses asymptotic efficiency in parametric models. Chaudhuri and Hsu (2012) establish finite sample guarantees in the form of convergence rates of nonparametric estimators, relating them to influence functions in robust statistics. Karwa and Vadhan (2017) construct finite sample confidence intervals for the mean of a normally distributed variable. Sheffet (2017), Alabi et al. (2020), and Jiang et al. (2024) study private estimation of parameters, t-values, and confidence intervals in linear regression models. Kamath et al. (2020) calculate the minimum sample size needed for certain accuracy to nonparametrically estimate the mean subject to central differential privacy. Drechsler et al. (2021) construct differentially private nonparametric confidence intervals for the median. Golowich (2021) studies learnability of nonparametric regression function classes subject to central differential privacy. Bun et al. (2021) analyse private hypothesis selection.

Later, attention has shifted towards *local privacy*. As opposed to central privacy, local privacy is concerned with privacy at the level of units, noising the covariates of individual units in the sample. In the local paradigm, Loh and Wainwright (2012) establish convergence rates for high-dimensional linear models with noisy covariates, also accommodating local differential privacy. Kairouz et al. (2015) investigate the trade-off between the privacy level of differentially private mechanisms and various measures of statistical util-

[2]Sometimes it is, somewhat misleadingly, also called unit-level privacy to contrast it with other notions such as *element-level privacy* in Asi et al. (2019).

ity, assuming a finitely distributed covariate. Acharya et al. (2019) study local differentially private estimation of densities on a finite set. Barnes et al. (2020) obtain data processing inequalities bounding the Fisher information of the private-data model, working with a parametric $\theta \in \mathbb{R}^K$ model. Berrett et al. (2021) derive almost sure L_2-consistency and convergence rates for the expected L_2-loss of nonparametric regression estimators in a privacy paradigm close to, but not entirely equivalent to, the local differential privacy paradigm.

Causal estimands falling into our proposed class, such as the average treatment effect and the average treatment effect on the treated are parameters of frequent practical interest. Yet, studies combining causal inference with privacy are not particularly common. Battistin and Chesher (2014) quantify the bias in average treatment effects under covariates measured with additive Gaussian noise (such as the Gaussian mechanism for differential privacy), but they do not consider estimation. Zhu et al. (2022) consider measurement error in the treatment, not the covariates, in a nonparametric model. Ohnishi and Awan (2023) study local differentially private estimation of the average treatment effect in a randomised experiment, whereas our parameter class is more geared towards observational, nonrandomised studies. Niu et al. (2022) propose meta-algorithms for the estimation of the conditional average treatment effect in a nonparametric model subject to central privacy constraints. Agarwal and Singh (2024) address various types of data corruption, neither limited to privacy protection nor to identically distributed data, and derive finite sample bounds. However, they assume that the covariates admit a low-rank representation and that the privacy mechanism is additive. We do not rely on such assumptions, especially that additive privacy mechanism is not compatible with categorical, i.e. finitely distributed, covariates; indeed we allow for generic covariates. Moreover, our class of parameters is not limited to causal ones. Li et al. (2024), also assuming a finitely distributed covariate, study the centrally private estimation of the conditional average treatment effect in an adaptive experiment, that is, in a contextual bandit setting.

A recent contribution to the combination of statistical efficiency, which double robustness is a means to, and local privacy is the work of Steinberger (2023). He considers efficient private estimation when also the privacy mech-

anism to use is optimised over in the class of all possible differentially private mechanisms. This is in contrast to our approach, wherein we consider efficient private estimation for a *given* privacy mechanism. In this respect, our work is more limited. Yet, we allow for infinite-dimensional models, whereas the results of Steinberger (2023) are limited to low-dimensional models parametrised by $\theta \in \mathbb{R}^K$, or even $\theta \in \mathbb{R}$ is some cases. Another contribution is due to Duchi et al. (2018) and Duchi and Ruan (2024), providing minimax and local minimax rates, respectively, also optimising over a class of differentially private mechanisms. However, while they provide minimax results only up to numerical constants, our results are asymptotically exact.

In Section 3.7, we propose a locally private version of the method of moments estimator (Hansen, 1982; Newey and McFadden, 1994) which fits into the framework of empirical risk minimisers, M-estimators. This framework has been studied in the literature for parametric models $\theta \in \mathbb{R}^K$. Chaudhuri et al. (2011), Kifer et al. (2012), and Bassily et al. (2014) propose centrally private approximation method to the *sample* M-criterion. Fukuchi et al. (2017) analyse empirical risk minimisation in the local privacy paradigm, establishing convergence rates. Lei (2011) and Slavkovic and Molinari (2021), also building on Newey and McFadden (1994), design a centrally differentially private M-estimation framework. Mangold et al. (2023) construct centrally differentially private empirical risk minimisers and derive rates for the obtained minimum in a high-dimensional parametric model. These results prove convergence and minimax rates, unconcerned about the (asymptotic) distribution of the error. Rather surprisingly, only two studies appear to address the question of the limiting distribution. Asi et al. (2019) establish an asymptotic normality result in a privacy paradigm less stringent than central privacy (hence, in turn, less stringent than local privacy), and the estimator centred at a quantity arising from discretisation (which need not be the true parameter). Asi and Duchi (2020) in the central privacy paradigm show the asymptotic normality of a private estimator of the median. To the best of our knowledge, our private estimator in Section 3.7 seems to be the first to obtain asymptotic normality in a local privacy paradigm, centred at the true parameter of interest. This is possible because we not only approximate the sample analogue of the M-criterion pri-

vately, as most authors, but we have exact identification on the population level, and we build thereon to show asymptotic normality.

3.3. Preliminaries

Let (V, X) be a random element defined on the probability space $(\Omega, \mathscr{F}_\Omega, \mathbb{P})$, with distribution P_{VX} belonging to $\mathcal{P}_{VX} \subseteq \bar{\mathcal{P}}_{VZ}$, where $\bar{\mathcal{P}}_{VZ}$ is the set of all possible distributions on the measurable space $(\mathfrak{V} \times \mathfrak{X}, \mathscr{F}_{\mathfrak{V}\times\mathfrak{X}})$. We assume that P_{VX} admits a density p_{VX} with respect to a known dominating product measure $\nu_V \times \nu_X$, thus $p_{VX} \in \mathrm{d}\mathcal{P}_{VX} \coloneqq \left\{ \frac{\mathrm{d}P}{\mathrm{d}(\nu_V \times \nu_X)} : P \in \mathcal{P}_{VX} \right\}$. By not making finite-dimensional parametric assumptions about p_{VX}, our interest is in the nonparametric model, wherein $\mathrm{d}\mathcal{P}_{VX} = \mathrm{d}\bar{\mathcal{P}}_{VX}$, the set of all possible densities:

$$\mathrm{d}\bar{\mathcal{P}}_{VX} \coloneqq \left\{ p : \mathfrak{V} \times \mathfrak{X} \to \mathbb{R}_+, \int_{\mathfrak{V}\times\mathfrak{X}} p \,\mathrm{d}(\nu_V \times \nu_X) = 1 \right\}. \tag{3.1}$$

Our aim is to infer $\mathbb{R}$-valued parameters $\chi(P_{VX})$ of P_{VX} when X, which we refer to as the covariate(s), is protected by a local privacy mechanism. We shall be more specific in Section 3.4, after we introduce our notation.

Notation. For a possibly random $h : \mathfrak{V} \times \mathfrak{X} \to \mathbb{R}$, we denote with $P_{VX}h$ the integral $\int_{\mathfrak{V}\times\mathfrak{X}} h(v,x) \,\mathrm{d}P_{VX}(v,x)$, which remains a random variable if and only if h is a random function with a particular realisation $(v,x) \mapsto h(\omega; v, x)$ for an $\omega \in \Omega$. We may write $P_{VX}h(V,X)$ for the same integral to indicate the variables of integration. We adopt this notation especially when h also takes a third argument, say, θ in some set Θ, so we write $P_{VX}h(V,X,\theta)$ for the integral $\int_{\mathfrak{V}\times\mathfrak{X}} h(v,x,\theta) \,\mathrm{d}P_{VX}(v,x)$, which depends on θ. For $p \in [1,\infty)$ and h with codomain $\mathbb{R}$, we write $\|h\|_{L_p(P_{VX})} \coloneqq \left(\int_{\mathfrak{V}\times\mathfrak{X}} |h(v,x)|^p \,\mathrm{d}P_{VX}(v,x) \right)^{\frac{1}{p}}$, and we let $\|h\|_\infty = \sup_{(v,x)\in\mathfrak{V}\times\mathfrak{X}} |h(v,x)|$. We write $L_p(P_{VX})$ to denote the class of functions h: $\|h\|^p_{L_p(P_{VX})} < \infty$, and write $L_p^0(P_{VX})$ for a class of functions h that are in $L_p(P_{VX})$ and $P_{VX}h = 0$. Similar notation applies for integration with respect to the empirical measure $\mathbb{P}_n \coloneqq \frac{1}{n}\sum_{i\in[n]} \delta_{(V_i,X_i)}$ and other (random) probability measures.

3.4. Double Robust Inference

We confine ourselves to parameters $\chi(P_{VX})$ in a novel class that exhibits a rate double robustness property. We characterise this class in Section 3.4.1. In Section 3.4.2 and Section 3.4.3, we analyse the semiparametric properties and study estimation of $\chi(P_{VX})$ in the setting where X is *not* privatised and is therefore observed.

3.4.1. Double Robust Parameter Class

Consider the following class of parameters. Let $m, g : \mathfrak{V} \times \mathfrak{X} \to \mathbb{R}$ be arbitrary, fixed and known functions which do not depend on P_{VX}. Let V_1 and V_2 be arbitrary, potentially empty, subsets of V, taking values in the sets $\mathfrak{V}_1$ and $\mathfrak{V}_2$, respectively, where, importantly, $\mathfrak{V}_2$ is finite. Define the conditional expectation (regression) functions

$$\mu_{\mathcal{X}}(v_1, x) \coloneqq \mathbb{E}\left[m(V, X) \mid V_1 = v_1, X = x\right], \quad (v_1, x) \in \mathfrak{V}_1 \times \mathfrak{X}, \tag{3.2}$$

$$\gamma_{\mathcal{V}}(v_2) \coloneqq \mathbb{E}\left[g(V, X) \mid V_2 = v_2\right], \quad v_2 \in \mathfrak{V}_2, \tag{3.3}$$

assuming that $\mu_{\mathcal{X}} \in L_2(P_{V_1X})$ and $\gamma_{\mathcal{V}} \in L_2(P_{V_2})$, where P_{V_1X} and P_{V_2} are the marginal distributions of (V_1, X) and V_2, respectively, with densities $p_{V_1X} \coloneqq \frac{\mathrm{d}P_{V_1X}}{\mathrm{d}\nu_{V_1X}}$ and $p_{V_2} \coloneqq \frac{\mathrm{d}P_{V_2}}{\mathrm{d}\nu_{V_2}}$ with respect to known dominating measures ν_{V_1X}, ν_{V_2}. Then the proposed parameters are of the form

$$\chi(P_{VX}) \coloneqq \mathbb{E} f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) \tag{3.4}$$

for a fixed $c \in \mathfrak{V}_2$ and an arbitrary, known function

$$f : \mathfrak{V} \times \mathfrak{X} \times L_2(P_{V_1X}) \times \Gamma \ni (v, x, \mu, \gamma) \mapsto f(v, x, \mu, \gamma) \in \mathbb{R}$$

that does not not depend on P_{VX}, where Γ is an arbitrary fixed subset of $\mathbb{R}$ which includes $g(\mathfrak{V}, \mathfrak{X})$, the image of g. We constrain f, requiring that

$$L_2(P_{V_1X}) \ni \mu \mapsto \mathbb{E} f(V, X, \mu, \gamma) \text{ is continuous for all } \gamma \in \Gamma, \tag{C.C}$$

$$L_2(P_{V_1X}) \ni \mu \mapsto f(V, X, \mu, \gamma) \text{ is linear } P_{VX}\text{-a.s. for all } \gamma \in \Gamma, \tag{C.L}$$

$$\Gamma \ni \gamma \mapsto f(V, X, \mu, \gamma) \text{ is twice continuously differentiable } P_{VX}\text{-a.s. for all } \mu \in L_2(P_{V_1X}). \tag{C.D}$$

Notice that $\chi(P_{VX})$ depends on the whole function of the infinite-dimensional regression $\mu_{\mathcal{X}}$, but depends on the low-dimensional regression $\gamma_{\mathcal{V}}$ evaluated at a single point $c \in \mathfrak{V}_2$. The latter is without much loss of generality — but with great notational convenience — as our results could be extended to f taking the whole regression function $(\gamma_{\mathcal{V}}(c))_{c\in\mathfrak{V}_2}$ as its input.

For brevity, we denote with $\partial_\gamma f$ the first and with $\partial^2_\gamma f$ the second derivative in (C.D), and we note that, importantly, the linearity (C.L) of f implies the P_{VX}-a.s. linearity of $\mu \mapsto \frac{\partial^j f}{\partial \gamma^j}(V, X, \mu, \bar{\gamma})$ for any integer $j \geq 1$ for which the derivative exists.[3] For estimation purposes we also require that

$$\mathbb{E}\left[\{f(V, X, \mu, \gamma) - f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))\}^2\right] \to 0 \text{ as } (\mu, \gamma) \to (\mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)), \tag{C.S}$$

$$\mathbb{E}\left[\{\partial_\gamma f(V, X, \mu, \gamma) - \partial_\gamma f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))\}^2\right] \to 0 \text{ as } (\mu, \gamma) \to (\mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)). \tag{C.DS}$$

In (C.C), continuity is understood with respect to the $L_2(P_{V_1X})$-norm; in (C.S) and (C.DS), the convergence of (μ, γ) is understood with respect to the product metric ρ induced by the $L_2(P_{V_1X})$-norm on $L_2(P_{V_1X})$ and by any norm on $\mathbb{R}$.[4] The condition (C.S) states that f is smooth enough in the two regression parameters for $f(\cdot, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))$ to be well approximated on average by $f(\cdot, \mu, \gamma)$ as long as $(\mu, \gamma) \to (\mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))$. The condition (C.DS) has the same intuition for $\partial_\gamma f$.

It is assumed that f, m, g, h are such that the integrals converge; in particular

$$\mathbb{E} f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))^2 < \infty, \tag{3.5}$$

$$\mathbb{E}\mathbb{1}_{V_2=c} g(V, X)^2 < \infty, \tag{3.6}$$

$$\mathbb{E}\left[m(V, X)^2 \,\middle|\, V_1, X\right] < \infty \quad P_{V_1X}\text{-a.s.}. \tag{3.7}$$

[3] This follows directly from the definition of the derivative. For $j = 1$, $\frac{\partial f}{\partial \gamma}(v, x, \mu, \bar{\gamma}) = \lim_{\epsilon\to 0} \frac{f(v,x,\mu,\bar{\gamma}+\epsilon) - f(v,x,\mu,\bar{\gamma})}{\epsilon}$, which is linear μ by (C.L). For $j \geq 2$, it follows from the linearity of the $(j-1)$th derivative.

[4] Norms on finite dimensional linear spaces such as $\mathbb{R}$ are equivalent (see e.g. Kress (2014, Theorem 1.6)).

There are various parameters of the form (3.4). A trivial example is the mean of $h(V, X)$ for some known, real-valued function h, which is included by setting

$$f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) \coloneqq h(V, X).$$

Rotnitzky et al. (2021) and Chernozhukov et al. (2022) present a list of more complicated parameters belonging to this class. For exposition, we illustrate how the average treatment effect is included, for which we adopt the potential outcome framework of Neyman (1924) and Rubin (1974) as in the previous chapters.

Example 3.1 (Average Treatment Effect)**.** *Let $V \coloneqq (Y, D)$ for a treatment $D \in \{0, 1\}$ and an outcome $Y = DY^1 + (1 - D)Y^0$ for partially unobserved potential outcomes Y^0, Y^1 with values in $\mathbb{R}$. Take $V_1 \coloneqq D$, $V_2 \coloneqq \varnothing$ empty, $m(V, X) \coloneqq Y$, $\mu_{\mathcal{X}}(d, x) \coloneqq \mathbb{E}[Y \mid D = d, X = x]$, and $g, \gamma_{\mathcal{V}}$ anything. If the covariates X are such that $Y^d \perp\!\!\!\perp D \mid X$, then $\mathbb{E}Y^d = \mathbb{E}\mu_{\mathcal{X}}(d, X)$ for $d \in \{0, 1\}$. Then $f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) \coloneqq \mu_{\mathcal{X}}(d, X)$ gives*

$$\mathbb{E}f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) = \mathbb{E}Y^d,$$

while $f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) \coloneqq \mu_{\mathcal{X}}(1, X) - \mu_{\mathcal{X}}(0, X)$ gives

$$\mathbb{E}f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) = \mathbb{E}Y^1 - \mathbb{E}Y^0,$$

with (C.C), (C.L), (C.D), (C.S) (C.DS) holding.

If we did not allow for the dependence on the parameter $\gamma_{\mathcal{V}}$, our class of parameters (3.4) and conditions (C.C), (C.L) would be a strict subset of those of Rotnitzky et al. (2021). Allowing such a nonlinear but smooth dependence enables us to capture more parameters, such as the average treatment effect on the treated, which was indeed already shown to be double robust by Chernozhukov et al. (2022, Example 6).

Example 3.2 (Average Treatment Effect on the Treated)**.** *Consider the setting of Example 3.1, but now take $g(V, X) \coloneqq D$, and $\gamma_{\mathcal{V}}(c) \coloneqq p_1$ for $p_1 \coloneqq \mathbb{E}D$ (so that V_2 is again empty). Then $\mathbb{E}\left[Y^0 \,\middle|\, D = 1\right] = \mathbb{E}D\mu_{\mathcal{X}}(0, X)/p_1$. Hence $f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}) \coloneqq D\mu_{\mathcal{X}}(0, X)/p_1$ gives*

$$\mathbb{E}f(Y, D, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}) = \mathbb{E}\left[Y^0 \,\middle|\, D = 1\right],$$

while $f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}) \coloneqq D(\mu_{\mathcal{X}}(1, X) - \mu_{\mathcal{X}}(0, X))/p_1$ *gives*

$$\mathbb{E}f(Y, D, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}) = \mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D = 1\right],$$

with (C.C)*,* (C.L)*,* (C.D)*,* (C.S)*,* (C.DS) *holding.*

3.4.2. Efficient Influence Function

The efficient influence function $\tilde{\chi}$ of the parameter $\chi(P_{VX})$ enables the construction of asymptotically efficient estimators of $\chi(P_{VX})$ whose limiting variance is $P_{VX}\tilde{\chi}^2$. To this end, we derive the efficient influence function of $\chi(P_{VX})$ under the nonparametric model (3.1). To begin with, note that the continuity and linearity conditions (C.C), (C.L) imply that for all $\gamma \in \Gamma$, there exists a unique function $r_{P_{VX},\gamma} : \mathfrak{V}_1 \times \mathfrak{X} \to \mathbb{R}$, $r_{P_{VX},\gamma} \in L_2(P_{V_1X})$, such that

$$\mathbb{E}f(V, X, \mu, \gamma) = \mathbb{E}r_{P_{VX},\gamma}(V_1, X)\mu(V_1, X) \quad \text{for all } \mu \in L_2(P_{V_1X}),$$

where the expectations are taken with respect to P_{VX}. This is the consequence of the Riesz representation theorem, and $r_{\gamma,P_{VX}}$ is called the Riesz representer of $\mu \mapsto \mathbb{E}f(V, X, \mu, \gamma)$.[5] As we mostly deal with the case when $\gamma = \gamma_{\mathcal{V}}(c)$ for a given $c \in \mathfrak{V}_2$, we denote with r the Riesz representer of $\mu \mapsto \mathbb{E}f(V, X, \mu, \gamma_{\mathcal{V}}(c))$ satisfying

$$\mathbb{E}f(V, X, \mu, \gamma_{\mathcal{V}}(c)) = \mathbb{E}r(V_1, X)\mu(V_1, X) \quad \text{for all } \mu \in L_2(P_{V_1X}), \tag{3.8}$$

with the expectations again taken with respect to P_{VX}, and we suppress its dependence on $(P_{VX}, \gamma_{\mathcal{V}}(c))$ for notational convenience. Typically, one can find the representer r by the tower property of expectations, conditioning on (V_1, X).

Example 3.1 (Average Treatment Effect, continued)**.** *Let*

$$\pi_{\mathcal{X}}(d|\, x) \coloneqq \mathbb{E}\left[\mathbb{1}_{D=d} \,|\, X = x\right]$$

denote the propensity score. Then the Riesz representer for the case of $\mathbb{E}Y^d$ *is*

$$r(d', x) = \frac{\mathbb{1}_{d'=d}}{\pi_{\mathcal{X}}(d|\, x)},$$

[5] Named after the Hungarian mathematician Riesz Frigyes.

and for the case of $\mathbb{E}Y^1 - \mathbb{E}Y^0$*, it is*

$$r(d,x) = \frac{d}{\pi_{\mathcal{X}}(1|\,x)} - \frac{1-d}{1-\pi_{\mathcal{X}}(1|\,x)}.$$

None of these representers r *depends on* $\gamma_{\mathcal{V}}(c)$*.*

Example 3.2 (Average Treatment Effect on the Treated, continued)**.** *The Riesz representer for the case of* $\mathbb{E}\left[Y^0 \,\middle|\, D=1\right]$ *is*

$$r(d,x) = \frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|\,x)}{1-\pi_{\mathcal{X}}(1|\,x)},$$

and for the case of $\mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D=1\right]$*, it is*

$$r(d,x) = \frac{d}{p_1} - \frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|\,x)}{1-\pi_{\mathcal{X}}(1|\,x)}.$$

Both representers r *depend on* $\gamma_{\mathcal{V}}(c) = p_1$*.*

Using the representation (3.8), we can now derive the efficient influence function of $\chi(P_{VX})$. Proposition 3.1 clearly shows how the dependence on the regression functions $\mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)$ manifests itself in the efficient influence function $\tilde{\chi}$. Indeed, if $\chi(P_{VX})$ did not depend on $\mu_{\mathcal{X}}$, the first term in (3.9) would be zero; similarly, if $\chi(P_{VX})$ did not depend on $\gamma_{\mathcal{V}}(c)$, so that $\partial_\gamma f = 0$, the second term in (3.9) would be zero.

Proposition 3.1 (Efficient Influence Function of $\chi(P_{VX})$)**.** *In the nonparametric model* (3.1) *for* P_{VX}*, the efficient influence function* $\tilde{\chi} : \mathfrak{V}\times\mathfrak{X} \to \mathbb{R}$ *of* $\chi(P_{VX})$ *in* (3.4) *is, at* P_{VX}*,*

$$\begin{aligned}\tilde{\chi}(v,x) \coloneqq\, & r(v_1,x)(m(v,x) - \mu_{\mathcal{X}}(v_1,x)) \\ & + \frac{\mathbb{1}_{v_2=c}}{p_{V_2}(c)}(g(v,x) - \gamma_{\mathcal{V}}(c))\mathbb{E}\partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) \\ & + f(v,x,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) - \chi(P_{VX}),\end{aligned} \tag{3.9}$$

where we denote by v_1, v_2 *the subsets of coordinates of* v *that correspond to* V_1, V_2 *of* V*. In the case of* $V_2 = \varnothing$*, it is understood that* $\frac{\mathbb{1}_{v_2=c}}{p_{V_2}(c)}(g(v,x) - \gamma_{\mathcal{V}}(c)) = g(v,x) - \mathbb{E}g(V,X)$*.*

Proof. All proofs are presented in the appendices Sections 3.A to 3.C. ■

For the average treatment effects, Proposition 3.1 recovers the familiar efficient influence functions under the nonparametric model with unknown propensity score.

Example 3.1 (Average Treatment Effect, continued)**.** *Recall that* $m(V,X)=Y$*,* $\mu_{\mathcal{X}}(d,x)=\mathbb{E}\left[Y \mid D=d, X=x\right]$*, and that for the propensity score* $\pi_{\mathcal{X}}(d|\,x)=\mathbb{E}\left[\mathbb{1}_{D=d}|\,X=x\right]$*, the Riesz representer for* $\mathbb{E}Y^d$ *is* $r(d',x)=\frac{\mathbb{1}_{d'=d}}{\pi_{\mathcal{X}}(d|\,x)}$*, and for* $\mathbb{E}Y^1-\mathbb{E}Y^0$*, it is* $r(d,x)=\frac{d}{\pi_{\mathcal{X}}(1|\,x)}-\frac{1-d}{1-\pi_{\mathcal{X}}(1|\,x)}$*. Then* $\chi(P_{VX})=\mathbb{E}Y^d$ *with* $f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))=\mu_{\mathcal{X}}(d,X)$ *and* $\partial_\gamma f=0$*, has efficient influence function* (3.9) *equal to*

$$\begin{aligned}\tilde{\chi}(\tilde{v},\tilde{x})&=\frac{\mathbb{1}_{\tilde{d}=d}}{\pi_{\mathcal{X}}(d\mid\tilde{x})}(\tilde{y}-\mu_{\mathcal{X}}(\tilde{d},\tilde{x}))+\mu_{\mathcal{X}}(d,\tilde{x})-\chi(P_{VX})\\&=\frac{\mathbb{1}_{\tilde{d}=d}}{\pi_{\mathcal{X}}(d\mid\tilde{x})}(\tilde{y}-\mu_{\mathcal{X}}(d,\tilde{x}))+\mu_{\mathcal{X}}(d,\tilde{x})-\chi(P_{VX}),\end{aligned}$$

since $\mathbb{1}_{\tilde{d}=d}\mu_{\mathcal{X}}(\tilde{d},\tilde{x})=\mathbb{1}_{\tilde{d}=d}\mu_{\mathcal{X}}(d,\tilde{x})$*; while* $\chi(P_{VX})=\mathbb{E}Y^1-\mathbb{E}Y^0$ *with*

$$f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))=\mu_{\mathcal{X}}(1,X)-\mu_{\mathcal{X}}(0,X)$$

and $\partial_\gamma f=0$*, has efficient influence function* (3.9) *equal to*

$$\begin{aligned}\tilde{\chi}(v,x)&=\left(\frac{d}{\pi_{\mathcal{X}}(1|\,x)}-\frac{1-d}{1-\pi_{\mathcal{X}}(1|\,x)}\right)(y-\mu_{\mathcal{X}}(d,x))\\&\quad+\mu_{\mathcal{X}}(1,X)-\mu_{\mathcal{X}}(0,X)-\chi(P_{VX})\\&=\frac{d}{\pi_{\mathcal{X}}(1|\,x)}(y-\mu_{\mathcal{X}}(1,x))-\frac{1-d}{1-\pi_{\mathcal{X}}(1|\,x)}(y-\mu_{\mathcal{X}}(0,x))\\&\quad+\mu_{\mathcal{X}}(1,X)-\mu_{\mathcal{X}}(0,X)-\chi(P_{VX}),\end{aligned}$$

as in Hahn (1998, Proof of Theorem 1). Note that while the model for (Y^0,Y^1,D,X) *is* not *nonparametric, because it is constrained by* $Y^d\perp\!\!\!\perp D|\,X$*, the model for* (Y,D,X) *is nonparametric if the models for* $Y^0|\,X$*,* $Y^1|\,X$*,* $D|\,X$ *and* X *are all nonparametric.*

Example 3.2 (Average Treatment Effect on the Treated, continued)**.** *Recall that* $p_1=\mathbb{E}D$*,* $V_2=\varnothing$ *and* $\tilde{\gamma}_{\mathcal{V}}(c)=p_1$*, and that the Riesz representer for* $\mathbb{E}\left[Y^0\,\middle|\,D=1\right]$ *is* $r(d,x)=\frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|\,x)}{1-\pi_{\mathcal{X}}(1|\,x)}$ *and for* $\mathbb{E}\left[Y^1-Y^0\,\middle|\,D=1\right]$*, it is*

$r(d,x) = \frac{d}{p_1} - \frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|x)}{1-\pi_{\mathcal{X}}(1|x)}$. *Then* $\chi(P_{VX}) = \mathbb{E}\left[Y^0 \,\middle|\, D=1\right]$ *with*

$$\begin{aligned} f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) &= D\mu_{\mathcal{X}}(0,X)/p_1, \\ \partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) &= -D\mu_{\mathcal{X}}(0,X)/p_1^2 = -f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))/p_1 \end{aligned}$$

has efficient influence function (3.9) *equal to*

$$\begin{aligned} \tilde{\chi}(v,x) =& \frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|x)}{1-\pi_{\mathcal{X}}(1|x)}(y-\mu_{\mathcal{X}}(d,x)) - \frac{d-p_1}{p_1}\chi(P_{VX}) + D\mu_{\mathcal{X}}(0,X)/p_1 \\ & - \chi(P_{VX}) \\ =& \frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|x)}{1-\pi_{\mathcal{X}}(1|x)}(y-\mu_{\mathcal{X}}(0,x)) + \frac{d}{p_1}\mu_{\mathcal{X}}(0,X) - \frac{d}{p_1}\chi(P_{VX}), \end{aligned}$$

since $(1-d)\mu_{\mathcal{X}}(d,x) = (1-d)\mu_{\mathcal{X}}(0,x)$*; while* $\chi(P_{VX}) = \mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D=1\right]$ *with*

$$\begin{aligned} f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) &= D(\mu_{\mathcal{X}}(1,X) - \mu_{\mathcal{X}}(0,X))/p_1, \\ \partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) &= -D(\mu_{\mathcal{X}}(1,X) - \mu_{\mathcal{X}}(0,X))/p_1^2 \\ &= -f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))/p_1 \end{aligned}$$

has efficient influence function (3.9) *equal to*

$$\begin{aligned} \tilde{\chi}(v,x) =& \left(\frac{d}{p_1} - \frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|x)}{1-\pi_{\mathcal{X}}(1|x)}\right)(y-\mu_{\mathcal{X}}(d,x)) - \frac{d-p_1}{p_1}\chi(P_{VX}) \\ & + \frac{d}{p_1}(\mu_{\mathcal{X}}(1,x) - \mu_{\mathcal{X}}(0,x)) - \chi(P_{VX}) \\ =& \frac{d}{p_1}(y-\mu_{\mathcal{X}}(1,x)) - \frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|x)}{1-\pi_{\mathcal{X}}(1|x)}(y-\mu_{\mathcal{X}}(0,x)) \\ & + \frac{d}{p_1}(\mu_{\mathcal{X}}(1,x) - \mu_{\mathcal{X}}(0,x)) + \frac{d}{p_1}\chi(P_{VX}) \end{aligned}$$

as in Hahn (1998, Proof of Theorem 1). The same remark on the nonparametric nature of the (Y,D,X)*-model in Example 3.1 applies here too.*

3.4.3. Double Robust Estimation

With the efficient influence function at hand, we can construct an asymptotically efficient estimator of $\chi(P_{VX})$ from a random sample $((V_i,X_i))_{i\in[n]}$ from P_{VX} by taking an initial estimator $\chi(\hat{P}_{VX})$ and correcting it as

$$\hat{\chi}_n \coloneqq \chi(\hat{P}_{VX}) + \mathbb{P}_n\hat{\tilde{\chi}} = \chi(\hat{P}_{VX}) + \frac{1}{n}\sum_{i\in[n]}\hat{\tilde{\chi}}(V_i,X_i). \tag{3.10}$$

Here, the estimator of the influence function (3.9) is

$$\begin{aligned}\hat{\tilde{\chi}}(v,x) \coloneqq \hat{r}(v_1,x)(m(v,x)-\hat{\mu}_{\mathcal{X}}(v_1,x)) + \frac{\mathbb{1}_{v_2=c}}{\hat{p}_{V_2}(c)}(g(v,x)-\hat{\gamma}_{\mathcal{V}}(c))\hat{e} \\ + f(v,x,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - \chi(\hat{P}_{VX}),\end{aligned} \tag{3.11}$$

where $\hat{r},\hat{\mu}_{\mathcal{X}}$, taking values in $L_2(P_{V_1X})$, are some estimators of $r,\mu_{\mathcal{X}}$, respectively; $\hat{p}_{V_2}(c)$, taking values in $\mathbb{R}$, is some estimator of $p_{V_2}(c)$; and $\hat{e}$, taking values in $\mathbb{R}$, is some estimator of

$$e \coloneqq \mathbb{E}\partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)). \tag{3.12}$$

Note that we use the same initial estimator $\chi(\hat{P}_{VX})$ in $\hat{\tilde{\chi}}$, which we may set to

$$\chi(\hat{P}_{VX}) = \mathbb{P}_n f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) = \frac{1}{n}\sum_{i\in[n]} f(V_i,X_i,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)). \tag{3.13}$$

Therefore, we have

$$\begin{aligned}\hat{\chi}_n &= \chi(\hat{P}_{VX}) + \mathbb{P}_n\hat{\tilde{\chi}} \\ &= \frac{1}{n}\sum_{i\in[n]}\left\{\hat{r}(v_{1i},X_i)(m(V_i,X_i)-\hat{\mu}_{\mathcal{X}}(v_{1i},X_i)) + \frac{\mathbb{1}_{V_{2i}=c}}{\hat{p}_{V_2}(c)}(g(V_i,X_i)-\hat{\gamma}_{\mathcal{V}}(c))\hat{e}\right. \\ &\quad \left. + f(V_i,X_i,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c))\right\}.\end{aligned}$$

We assume that the estimators

$$\begin{aligned}&\hat{\eta} \coloneqq (\hat{r},\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c),\hat{p}_{V_2}(c),\hat{e}) \in L_2(P_{V_1X})^2\times\Gamma\times\mathbb{R}^2 \text{ of} \\ &\eta \coloneqq (r,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c),p_{V_2}(c),e)\end{aligned} \tag{3.14}$$

are computed from random samples from P_{VX} which are independent[6] of

$$\mathcal{S} \coloneqq ((V_i,X_i))_{i\in[n]}.$$

Specifically, we assume that there are two more random samples

$$\mathcal{S}' \coloneqq ((V_i',X_i'))_{i\in[n]} \text{ and } \mathcal{S}'' \coloneqq ((V_i'',X_i''))_{i\in[n]}$$

[6]In practice, cross-fitting may be applied to the same effect; see e.g. Kennedy (2023) for an in-detail description.

from P_{VX}, with $\mathcal{S}, \mathcal{S}', \mathcal{S}''$ pairwise independent, where $\mathcal{S}'$ is used for the estimation of $(r, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c), p_{V_2}(c))$, and $\mathcal{S}', \mathcal{S}''$ are used for the estimation of e in (3.12) as

$$\hat{e} \coloneqq \mathbb{P}_n'' \partial_\gamma f(V, X, \hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c)) \coloneqq \frac{1}{n} \sum_{i \in [n]} \partial_\gamma f(V_i'', X_i'', \hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c)). \tag{3.15}$$

With this estimation strategy, we can establish the consistency of $\hat{e}$, and hence of $\hat{\tilde{\chi}}$ and $\hat{\chi}_n$ in turn, without additional regularity conditions. Indeed, decompose

$$\sqrt{n}(\hat{\chi}_n - \chi(P_{VX})) = \sqrt{n}\mathbb{P}_n \tilde{\chi} + \sqrt{n}(\mathbb{P}_n - P_{VX})(\hat{\tilde{\chi}} - \tilde{\chi}) + \sqrt{n} R_n, \tag{3.16}$$

$$R_n \coloneqq \chi(\hat{P}_{VX}) - \chi(P_{VX}) + P_{VX} \hat{\tilde{\chi}}, \tag{3.17}$$

where we used that $P_{VX} \tilde{\chi} = 0$ by $\tilde{\chi}$ being the influence function. The term $\sqrt{n}\mathbb{P}_n \tilde{\chi} \overset{P_{VX}}{\rightsquigarrow} \mathcal{N}(0, P_{VX} \tilde{\chi}^2)$ by the standard central limit theorem. The second term in (3.16) is called the empirical process term and is vanishing as $o_{P_{VX}}(1)$ under consistent estimators $\hat{\eta}$ and stochastic boundedness conditions.

Assumption 3.1 (Consistent Estimators)**.** *It holds that*

$$\|\hat{r} - r\|_{L_2(P_{VX})} = o_{P_{VX}}(1), \tag{3.18}$$

$$\hat{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c) = o_{P_{VX}}(1), \tag{3.19}$$

$$\hat{p}_{V_2}(c) - p_{V_2}(c) = o_{P_{VX}}(1). \tag{3.20}$$

Further, it either holds that

$$\|m - \mu_{\mathcal{X}}\|_\infty = O(1), \tag{3.21}$$

$$\|\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}\|_\infty = o_{P_{VX}}(1), \tag{3.22}$$

or that

$$\|m - \hat{\mu}_{\mathcal{X}}\|_\infty = O_{P_{VX}}(1), \tag{3.23}$$

$$\|\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}\|_{L_2(P_{VX})} = o_{P_{VX}}(1), \tag{3.24}$$

$$P_{VX}\left(\left\{(V_1, X) \in \mathfrak{V}_1 \times \mathfrak{X} : |r(V_1, X)| > \bar{R}\right\}\right) = 0 \tag{3.25}$$

for some constant $\bar{R} < \infty$. Above, we may replace the $\|\cdot\|_{L_2(P_{VX})}$ norms by $\|.\|_\infty$ norms. Likewise, (3.25) may be replaced by $\|r\|_\infty < \infty$.

Lemma 3.1 (Vanishing Empirical Process Term)**.** *Assume that η is estimated from $\mathcal{S}', \mathcal{S}''$ as described above and that Assumption 3.1 holds. Then $\hat{e} - e = o_{P_{VX}}(1)$ and $(\mathbb{P}_n - P_{VX})(\hat{\tilde{\chi}} - \tilde{\chi}) = o_{P_{VX}}\left(n^{-1/2}\right)$.*

Hence, if the nuisance parameters in η are consistently estimated and bound conditions apply, the behaviour of $\sqrt{n}(\hat{\chi}_n - \chi(P_{VX}))$ is governed by the second-order bias term R_n in (3.17). We show that our class of parameters (3.4) enjoys a rate double robustness property, therefore R_n exhibits a product-structure of estimation errors.

Theorem 3.1 (Double Robustness)**.** *Let $\mu'_{\mathcal{X}}, r'$ be arbitrary, possibly random, elements taking values in $L_2(P_{V_1X})$. Let $p'_{V_2}(c), \gamma'_{\mathcal{V}}(c), \chi'$ be arbitrary, possibly random, variables taking values in $\mathbb{R}$. For the true distribution P_{VX} and for the $\mu'_{\mathcal{X}}, \gamma'_{\mathcal{V}}(c)$, let*

$$\chi_0 \coloneqq \chi(P_{VX}), \quad e' \coloneqq P_{VX}\partial_\gamma f(V, X, \mu'_{\mathcal{X}}, \gamma'_{\mathcal{V}}(c))$$

be $\mathbb{R}$-valued, possibly random, variables. Let e'' be an arbitrary $\mathbb{R}$-valued, possibly random, variable. In the spirit of (3.9), set

$$\begin{aligned}\tilde{\chi}'(v,x) \coloneqq & r'(v_1,x)(m(v,x) - \mu'_{\mathcal{X}}(v_1,x)) + \frac{\mathbb{1}_{v_2=c}}{p'_{V_2}(c)}(g(v,x) - \gamma'_{\mathcal{V}}(c))e'' \\ & + f(v,x,\mu'_{\mathcal{X}}, \gamma'_{\mathcal{V}}(c)) - \chi'.\end{aligned} \tag{3.26}$$

Suppose that $r' - r, \mu'_{\mathcal{X}} - \mu_{\mathcal{X}} \in L_2(P_{V_1X})$. Then

$$\begin{aligned}\chi' - \chi_0 + P_{VX}\tilde{\chi}' = & - P_{VX}(r - r')(\mu_{\mathcal{X}} - \mu'_{\mathcal{X}}) \\ & + (\gamma_{\mathcal{V}}(c) - \gamma'_{\mathcal{V}}(c))\left(\frac{p_{V_2}(c)}{p'_{V_2}(c)}e'' - e'\right) \\ & - (\gamma_{\mathcal{V}}(c) - \gamma'_{\mathcal{V}}(c))^2 \frac{P_{VX}\partial^2_\gamma f(V, X, \mu'_{\mathcal{X}}, \widetilde{\gamma_{\mathcal{V}}(c)})}{2}\end{aligned} \tag{3.27}$$

for some $\widetilde{\gamma_{\mathcal{V}}(c)}$ between $\gamma_{\mathcal{V}}(c)$ and $\gamma'_{\mathcal{V}}(c)$.

Theorem 3.1 implies that the bias in (3.17) is

$$\begin{aligned}R_n &= \chi(\hat{P}_{VX}) - \chi(P_{VX}) + P_{VX}\hat{\tilde{\chi}} \\ &= - P_{VX}(r - \hat{r})(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}) + (\gamma_{\mathcal{V}}(c) - \hat{\gamma}_{\mathcal{V}}(c))\left(\frac{p_{V_2}(c)}{\hat{p}_{V_2}(c)}\hat{e} - e'\right) \\ &\quad - (\gamma_{\mathcal{V}}(c) - \hat{\gamma}_{\mathcal{V}}(c))^2 \frac{P_{VX}\partial^2_\gamma f(V, X, \hat{\mu}_{\mathcal{X}}, \tilde{\gamma}_{\mathcal{V}}(c))}{2},\end{aligned} \tag{3.28}$$

for some $\tilde{\gamma}_{\mathcal{V}}(c)$ between $\gamma_{\mathcal{V}}(c)$ and $\hat{\gamma}_{\mathcal{V}}(c)$, and

$$e' \coloneqq P_{VX}\partial_\gamma f(V, X, \hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c)). \tag{3.29}$$

Because V_2 is distributed on a finite set, any reasonable estimators of $\gamma_{\mathcal{V}}(c)$, $p_{V_2}(c)$ are $\sqrt{n}$-consistent; for instance

$$\hat{\gamma}_{\mathcal{V}}(c) \coloneqq \frac{1}{N_c}\sum_{i\in[n]} \mathbb{1}_{V'_{2i}=c} g(V'_i, X'_i), \quad \hat{p}_{V_2}(c) \coloneqq N_c/n, \quad N_c \coloneqq \sum_{i\in[n]} \mathbb{1}_{V'_{2i}=c} \tag{3.30}$$

satisfy $\hat{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c) = O_{P_{VX}}\left(n^{-1/2}\right)$, $\hat{p}_{V_2}(c) - p_{V_2}(c) = O_{P_{VX}}\left(n^{-1/2}\right)$ by the standard central limit theorem. Suppose that $\hat{e} - e' = o_{P_{VX}}(1)$ and that $P_{VX}\partial_\gamma^2 f(V, X, \hat{\mu}_{\mathcal{X}}, \tilde{\gamma}_{\mathcal{V}}(c)) = O_{P_{VX}}(1)$. It follows from (3.28) that the bias is then

$$R_n = -P_{VX}((r - \hat{r})(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}})) + o_{P_{VX}}\left(n^{-1/2}\right). \tag{3.31}$$

Hence, R_n is ultimately determined by the product of the estimation errors of the Riesz representer r and the regression function $\mu_{\mathcal{X}}$. A faster rate of one estimator can be traded off for a slower rate of the other. For instance, if a parametric model for r is correctly specified and estimated at $\sqrt{n}$-rate, then it suffices that $\mu_{\mathcal{X}}$ is consistent, without rate requirements, and the other way around too. For the average treatment effect (on the treated) the bias takes the following form.

Example 3.1 (Average Treatment Effect, continued)**.** *Recall that the Riesz representer for $\mathbb{E}Y^d$ is $r(d', x) = \frac{\mathbb{1}_{d'=d}}{\pi_{\mathcal{X}}(d|x)}$, and for $\mathbb{E}Y^1 - \mathbb{E}Y^0$, it is $r(d, x) = \frac{d}{\pi_{\mathcal{X}}(1|x)} - \frac{1-d}{1-\pi_{\mathcal{X}}(1|x)}$. For $\mathbb{E}Y^d$, the bias is*

$$\begin{aligned} R_n =& -P_{VX}(r - \hat{r})(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}) \\ =& P_{DX}\left[\mathbb{1}_{D=d}\frac{\hat{\pi}_{\mathcal{X}}(d|X) - \pi_{\mathcal{X}}(d|X)}{\hat{\pi}_{\mathcal{X}}(d|X)\pi_{\mathcal{X}}(d|X)}\left(\hat{\mu}_{\mathcal{X}}(D, X) - \mu_{\mathcal{X}}(D, X)\right)\right], \end{aligned}$$

while for $\mathbb{E}Y^1 - \mathbb{E}Y^0$ it is

$$\begin{aligned} R_n =& -P_{VX}(r - \hat{r})(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}) \\ =& P_{DX}\left[D\frac{\hat{\pi}_{\mathcal{X}}(1|X) - \pi_{\mathcal{X}}(1|X)}{\hat{\pi}_{\mathcal{X}}(1|X)\pi_{\mathcal{X}}(1|X)}\left(\hat{\mu}_{\mathcal{X}}(D, X) - \mu_{\mathcal{X}}(D, X)\right)\right] \\ &+P_{DX}\left[(1-D)\frac{\hat{\pi}_{\mathcal{X}}(1|X) - \pi_{\mathcal{X}}(1|X)}{(1-\hat{\pi}_{\mathcal{X}}(1|X))(1-\pi_{\mathcal{X}}(1|X))}\left(\hat{\mu}_{\mathcal{X}}(D, X) - \mu_{\mathcal{X}}(D, X)\right)\right]. \end{aligned}$$

Noting that $\mathbb{1}_{D=d}\hat{\mu}_{\mathcal{X}}(D,X) = \mathbb{1}_{D=d}\hat{\mu}_{\mathcal{X}}(d,X)$ *and similarly for* $\mu_{\mathcal{X}}$*, we arrive, by the definition of* $\pi_{\mathcal{X}}$ *and the tower property of expectation, to the usual bias formulae*

$$
\begin{aligned}
R_n &= P_X\left[\frac{\hat{\pi}_{\mathcal{X}}(d|\,X)-\pi_{\mathcal{X}}(d|\,X)}{\hat{\pi}_{\mathcal{X}}(d|\,X)}\big(\hat{\mu}_{\mathcal{X}}(d,X)-\mu_{\mathcal{X}}(d,X)\big)\right] && \text{for } \mathbb{E}Y^d,\\
R_n &= P_X\left[\frac{\hat{\pi}_{\mathcal{X}}(1|\,X)-\pi_{\mathcal{X}}(1|\,X)}{\hat{\pi}_{\mathcal{X}}(1|\,X)}\big(\hat{\mu}_{\mathcal{X}}(1,X)-\mu_{\mathcal{X}}(1,X)\big)\right]\\
&\quad + P_X\left[\frac{\hat{\pi}_{\mathcal{X}}(1|\,X)-\pi_{\mathcal{X}}(1|\,X)}{1-\hat{\pi}_{\mathcal{X}}(1|\,X)}\big(\hat{\mu}_{\mathcal{X}}(0,X)-\mu_{\mathcal{X}}(0,X)\big)\right] && \text{for } \mathbb{E}Y^1-\mathbb{E}Y^0.
\end{aligned}
$$

Example 3.2 (Average Treatment Effect on the Treated, continued)**.** *Recall that the Riesz representer for* $\mathbb{E}\left[Y^0 \,\middle|\, D=1\right]$ *is* $r(d,x) = \frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|\,x)}{1-\pi_{\mathcal{X}}(1|\,x)}$*, and for* $\mathbb{E}\left[Y^1-Y^0 \,\middle|\, D=1\right]$*, it is* $r(d,x) = \frac{d}{p_1} - \frac{1-d}{p_1}\frac{\pi_{\mathcal{X}}(1|\,x)}{1-\pi_{\mathcal{X}}(1|\,x)}$*. Let* $\hat{p}_1 := \frac{1}{n}\sum_{i\in[n]} D_i'$ *denote the estimator of* $p_1 = \mathbb{E}D$*. For* $\mathbb{E}\left[Y^0 \,\middle|\, D=1\right]$*, the first term of the bias* R_n *in* (3.28) *is*

$$
\begin{aligned}
&\qquad\qquad -P_{VX}(r-\hat{r})(\mu_{\mathcal{X}}-\hat{\mu}_{\mathcal{X}})\\
&= P_{DX}\left[(1-D)\left(\frac{1}{p_1}\frac{\pi_{\mathcal{X}}(1|\,X)}{1-\pi_{\mathcal{X}}(1|\,X)} - \frac{1}{\hat{p}_1}\frac{\hat{\pi}_{\mathcal{X}}(1|\,X)}{1-\hat{\pi}_{\mathcal{X}}(1|\,X)}\right)\right.\\
&\qquad\qquad \left.\times\,(\hat{\mu}_{\mathcal{X}}(D,X)-\mu_{\mathcal{X}}(D,X))\right]\\
&= P_{DX}\left[(1-D)\left(\frac{1}{p_1}\frac{\pi_{\mathcal{X}}(1|\,X)}{1-\pi_{\mathcal{X}}(1|\,X)} - \frac{1}{\hat{p}_1}\frac{\hat{\pi}_{\mathcal{X}}(1|\,X)}{1-\hat{\pi}_{\mathcal{X}}(1|\,X)}\right)\right.\\
&\qquad\qquad \left.\times\,(\hat{\mu}_{\mathcal{X}}(0,X)-\mu_{\mathcal{X}}(0,X))\right]\\
&= P_X\left[(1-\pi_{\mathcal{X}}(1|\,X))\left(\frac{1}{p_1}\frac{\pi_{\mathcal{X}}(1|\,X)}{1-\pi_{\mathcal{X}}(1|\,X)} - \frac{1}{\hat{p}_1}\frac{\hat{\pi}_{\mathcal{X}}(1|\,X)}{1-\hat{\pi}_{\mathcal{X}}(1|\,X)}\right)\right.\\
&\qquad\qquad \left.\times\,(\hat{\mu}_{\mathcal{X}}(0,X)-\mu_{\mathcal{X}}(0,X))\right]\\
&= \frac{p_1-\hat{p}_1}{\hat{p}_1 p_1}P_X\left[\frac{(\hat{\pi}_{\mathcal{X}}(1|\,X)-1)\pi_{\mathcal{X}}(1|\,X)}{1-\hat{\pi}_{\mathcal{X}}(1|\,X)}(\hat{\mu}_{\mathcal{X}}(0,X)-\mu_{\mathcal{X}}(0,X))\right]\\
&\quad + \frac{p_1-\hat{p}_1}{\hat{p}_1}P_X\left[\frac{(\pi_{\mathcal{X}}(1|\,X)-\hat{\pi}_{\mathcal{X}}(1|\,X))(\hat{\mu}_{\mathcal{X}}(0,X)-\mu_{\mathcal{X}}(0,X))}{1-\hat{\pi}_{\mathcal{X}}(1|\,X)}\right]
\end{aligned}
$$

where the second equality follows similarly to the average treatment effect; while for

$\mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D = 1\right]$*, the first term of the bias* R_n *in* (3.28) *is then*

$$\begin{aligned}-P_{VX}(r - \hat{r})(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}) = \frac{\hat{p}_1 - p_1}{\hat{p}_1 p_1} P_X \left[\pi_{\mathcal{X}}(1|\,X)(\hat{\mu}_{\mathcal{X}}(1, X) - \mu_{\mathcal{X}}(1, X))\right] \\ + \frac{\hat{p}_1 - p_1}{\hat{p}_1 p_1} P_X \left[\frac{(\hat{\pi}_{\mathcal{X}}(1|\,X) - 1)\pi_{\mathcal{X}}(1|\,X)}{1 - \hat{\pi}_{\mathcal{X}}(1|\,X)}(\hat{\mu}_{\mathcal{X}}(0, X) - \mu_{\mathcal{X}}(0, X))\right] \\ + \frac{\hat{p}_1 - \hat{p}_1}{\hat{p}_1} P_X \left[\frac{(\pi_{\mathcal{X}}(1|\,X) - \hat{\pi}_{\mathcal{X}}(1|\,X))(\hat{\mu}_{\mathcal{X}}(0, X) - \mu_{\mathcal{X}}(0, X))}{1 - \hat{\pi}_{\mathcal{X}}(1|\,X)}\right].\end{aligned}$$

Suppose $\hat{e} - e' = o_{P_{VX}}(1)$ *and that* $P_{YDX}\partial_\gamma^2 f(Y, D, X, \hat{\mu}_{\mathcal{X}}, \tilde{p}_1) = O_{P_{VX}}(1)$*. Then for the bias to vanish at speed* $o_{P_{VX}}\left(n^{-1/2}\right)$*, it suffices that* $\|\Delta\|_{L_1(P_{VX})} = o_{P_{VX}}\left(n^{-1/2}\right)$*, where* $\Delta(x) \coloneqq (\hat{\mu}_{\mathcal{X}}(0|\,x) - \mu_{\mathcal{X}}(0|\,x))(\hat{\pi}_{\mathcal{X}}(1|\,x) - \pi_{\mathcal{X}}(1|\,x))$*, and* $\hat{\mu}_{\mathcal{X}}$ *is consistent because* $\hat{p}_1 - p_1 = O_{P_{VX}}\left(n^{-1/2}\right)$*, provided* $1 - \hat{\pi}_{\mathcal{X}}(1|\,x)$ *is bounded away from zero.*

Thus, we find that the average treatment effect on the treated is double robust. This is aligned with Chernozhukov et al. (2022, Example 6), and is an improvement on Rotnitzky et al. (2021, Example 12), who too, establish asymptotic normality, but not efficiency, as this parameter is not natively included in their class.

Our results amounts to the asymptotic efficiency of $\hat{\chi}_n$ under fast enough estimation rates, boundedness conditions, and the estimation strategy using independent samples.

Assumption 3.2 (Rates of Estimators)**.** *It holds that*

$$\begin{aligned}P_{VX}((r - \hat{r})(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}})) &= o_{P_{VX}}\left(n^{-1/2}\right), \\ \hat{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c) &= O_{P_{VX}}\left(n^{-1/2}\right), \\ \hat{p}_{V_2}(c) - p_{V_2}(c) &= O_{P_{VX}}\left(n^{-1/2}\right), \\ P_{VX}\partial_\gamma^2 f(V, X, \hat{\mu}_{\mathcal{X}}, \tilde{\gamma}_{\mathcal{V}}(c)) &= O_{P_{VX}}(1).\end{aligned}$$

Corollary 3.1 (Asymptotic Efficiency of $\hat{\chi}_n$)**.** *If the estimators are constructed from the independent random samples* $\mathcal{S}, \mathcal{S}', \mathcal{S}''$ *as described, and Assumptions 3.1 and 3.2 hold, then* $\sqrt{n}(\hat{\chi}_n - \chi(P_{VX})) \overset{P_{VX}}{\rightsquigarrow} \mathcal{N}(0, P_{VX}\tilde{\chi}^2)$ *as* $n \to \infty$*.*

We close this section by noting that, conveniently and also expectedly, the use of independent samples $\mathcal{S}, \mathcal{S}', \mathcal{S}''$ for the estimation is not necessary when

the image space $\mathfrak{X}$ of the covariates and $\mathfrak{V}_1$ are finite. Indeed, the plug-in estimator $\chi(\hat{P}_{VX})$ is asymptotically efficient when we compute the estimators $(\hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}})$ and $\chi(\hat{P}_{VX})$ on the same sample $\mathcal{S}$, provided some boundedness conditions hold; see Proposition 3.4 in Section 3.A.1 for a formal statement.

3.5. Privacy

Recall that our aim is to make inference about $\chi(P_{VX})$ based on a random sample $((V_i, X_i))_{i\in[n]}$ from P_{VX} such that the sensitive covariates X_i are privacy protected for all units $i \in [n]$. Accordingly, we introduce the privacy concepts in Section 3.5.1, followed by the discussion of how they allow identification in Section 3.5.2, and of their semiparametric properties in Section 3.5.3.

3.5.1. Privacy Concepts

To achieve privacy protection, a noisy version Z_i of X_i is generated with the help of a local privacy mechanism. The mechanism is called 'local', because the noising process happens at the level of each unit i, and therefore does not require a trusted third party. Correspondingly, each unit i keeps their own covariate X_i to themselves, and only reveals (V_i, Z_i) for inference purposes. We consider a setting where only X_i — as opposed to the covariates of all units, $(X_i)_{i\in[n]}$ — is used to generate Z_i — a subclass of privacy regimes that are known in the literature as noninteractive (Steinberger, 2023). Moreover, we assume that each unit i uses the same privacy mechanism, and consider inference *given* that privacy mechanism. As discussed in Section 3.2.2, this stands in contrast to the work of Steinberger (2023) and Duchi and Ruan (2024), who also optimise over possible privacy mechanisms in the whole class of differentially private mechnanisms. In some aspects, nevertheless, our results are more general; namely, they are valid for the nonparametric model (3.1) and are asymptotically exact. As is usual, we assume that the privacy mechanism is common knowledge, and hence can be used for inference purposes. To summarise, our aim is to make inference about $\chi(P_{VX})$ using only the sample

$((V_i, Z_i))_{i\in[n]}$, where Z_i is generated based solely on X_i via a known and fixed privacy mechanism, which is the same for all units $i \in [n]$.

Now we formalise the notion of a privacy mechanism. Let Z take values in the measurable space $(\mathfrak{Z}, \mathscr{F}_\mathfrak{Z})$, and conditionally on X, be drawn from some $Q \in \mathcal{Q}(\mathfrak{X} \to \mathfrak{Z})$, where $\mathcal{Q}(\mathfrak{X} \to \mathfrak{Z})$ is the set of all Markov kernels $Q : \mathscr{F}_\mathfrak{Z} \times \mathfrak{X} \to [0,1]$, so that $B \mapsto Q(B|\, x)$ is a probability measure for all $x \in \mathfrak{X}$, and $x \mapsto Q(B|\, x)$ is measurable for every $B \in \mathscr{F}_\mathfrak{Z}$. The Markov kernel Q is called the privacy mechanism. By construction, the privacy mechanism does not utilise any information in V, because it would only create a more involved dependency structure, and would clearly be without any increase in the level of privacy. Thus, Z is independent of V given the covariates X:

$$Z \perp\!\!\!\perp V \mid X,$$

and the conditional distribution of Z given (V, X) is $Q(\cdot|\, X)$. As we use an identical and noninteractive mechanism for all units $i \in [n]$, the sequence $((V_i, X_i, Z_i))_{i\in[n]}$ is an independent and identically distributed (i.e. random) sample from the distribution of the partly unobserved data (V, X, Z),

$$\begin{aligned} P_{VXZ}(B_\mathfrak{v}, B_\mathfrak{x}, B_\mathfrak{z}) &= \int \mathbb{1}_{(v,x)\in B_\mathfrak{v}\times B_\mathfrak{x}} Q(B_\mathfrak{z}|\, x)\, \mathrm{d}P_{VX}(v,x) \\ &= \int_{B_\mathfrak{v}} \int_{B_\mathfrak{x}} Q(B_\mathfrak{z}|\, x) p_{VX}(v,x)\, \mathrm{d}\nu_X(x)\, \mathrm{d}\nu_V(v), \end{aligned} \tag{3.32}$$

for $B_\mathfrak{v} \in \mathscr{F}_\mathfrak{V}, B_\mathfrak{x} \in \mathscr{F}_\mathfrak{X}, B_\mathfrak{z} \in \mathscr{F}_\mathfrak{Z}$. Therefore, $((V_i, Z_i))_{i\in[n]}$ is a random sample from the distribution of the observed data (V, Z), the mixture

$$P_{VZ}(B_\mathfrak{v}, B_\mathfrak{z}) = \int_{B_\mathfrak{v}} \int_\mathfrak{X} Q(B_\mathfrak{z}|\, x) p_{VX}(v,x)\, \mathrm{d}\nu_X(x)\, \mathrm{d}\nu_V(v), \quad B_\mathfrak{v} \in \mathscr{F}_\mathfrak{V}, B_\mathfrak{z} \in \mathscr{F}_\mathfrak{Z}. \tag{3.33}$$

The relation (3.33) provides an intuitive view on the role of the privacy mechanism Q as the carrier of information between the unobserved and observed data distribution. The amount of information carried determines the degree of privacy and the extent to which inference about $\chi(P_{VX})$ is possible. Indeed, if the mechanism Q does not depend on X, i.e. $Q(B|\, x) = \bar{Q}(B)$ for some probability measure $\bar{Q}$ on $\mathfrak{Z}$, then Z carries absolutely no information about X, and (3.33) reduces to $P_{VZ}(B_\mathfrak{v}, B_\mathfrak{z}) = P_V(B_\mathfrak{v})\bar{Q}(B_\mathfrak{z})$. This constitutes

maximal privacy, but it precludes inference about parameters $\chi(P_{VX})$ based on P_{VZ} (except those parameters that are only a functional of P_V, so that $\chi(P_{VX}) = \bar{\chi}(P_V)$ for some known functional $\bar{\chi}$). In the other extreme, if Q concentrates completely on X, i.e. $Q(B|\,x) = \delta_x(B)$ for the Dirac measure δ_x, then Z is equal to X and (3.33) reduces to $P_{VZ}(B_{\mathsf{v}}, B_{\mathsf{z}}) = P_{VX}(B_{\mathsf{v}}, B_{\mathsf{z}})$. This constitutes no privacy whatsoever, but it readily lends itself for inference about $\chi(P_{VX})$ based on $P_{VZ} = P_{VX}$. To understand this trade-off better, we first study privacy in more detail, and we turn to inference considerations in Section 3.5.2.

Intuitively, a privacy mechanism should introduce a sufficient amount of noise into the covariate X, so that the noisy version Z should not be too informative about the value of X. One of the most well-known definitions capturing this idea is α-differential privacy which is included in the more general definition of (α, δ)-differential privacy, for $\delta := 0$, formalised by Dwork et al. (2006).

Definition 3.1 (Local (α, δ)-Differential Privacy: (α, δ)-LDP)**.** *For $\alpha, \delta \geq 0$, the privacy mechanism $Q \in \mathcal{Q}(\mathfrak{X} \to \mathfrak{Z})$ is locally (α, δ)-differentially private if $Q(B|\,x) \leq e^{\alpha} Q(B|\,x') + \delta$ for all $B \in \mathscr{F}_{\mathfrak{Z}}$ and for all $x, x' \in \mathfrak{X}$.*

Definition 3.1 states that however much X varies, the extent to which Z varies is to be limited. As (α, δ) approaches zero, this variation is restricted to the degree such that $Q(B|\,x) \approx Q(B|\,x')$, so that the value of X does not matter for Z; this corresponds to maximal privacy. Conversely, (α, δ) approaching infinity means no privacy, because the inequality limiting this variation becomes vacuous.

While (α, δ)-LDP is presumably the most widespread privacy definition in use (Desfontaines and Pejó, 2022), we shall see in Section 3.5.2 that it is not particularly well suited for (our) inference purposes in general. Therefore, let us also recall the definition of local α-total variation privacy (Barber and Duchi (2014, Definition 4)), which shall be shown to be more suitable for inference purposes.

Definition 3.2 (Local α-Total Variation Privacy (α-LTVP))**.** *For $0 \leq \alpha \leq 1$, the privacy mechanism $Q \in \mathcal{Q}(\mathfrak{X} \to \mathfrak{Z})$ is locally α-differentially private if $\|Q(\cdot|\,x) - Q(\cdot|\,x')\|_{\mathrm{TV}} := \sup_{B \in \mathscr{F}_{\mathfrak{Z}}} |Q(B|\,x) - Q(B|\,x')| \leq \alpha$ for all $x, x' \in \mathfrak{X}$.*

Similarly to (α, δ)-LDP, α approaching zero means that α-LTVP provides maximal privacy, while α approaching one means no privacy. Lemma 3.2 shows that (α, δ)-LDP and α-LTVP are related, but there is no strict inclusion in either way for *arbitrary* (α, δ) values. However, when (α, δ) are small enough — so that privacy is strict —, (α, δ)-LDP translates into $(e^\alpha - 1 + \delta)$-LTVP. For example, an α-LDP, which is an $(\alpha, \delta = 0)$-LDP, is also an $(e^\alpha - 1)$-LTVP for $\alpha \leq \log(2) \approx 0.69$.

Lemma 3.2 ((α, δ)-LDP and α-LTVP)**.** *For all $0 \leq \alpha \leq 1$, every α-LTVP mechanism is $(\tilde{\alpha}, \alpha)$-LDP for any $\tilde{\alpha} \geq 0$. For all $\alpha, \delta \geq 0$ such that $e^\alpha - 1 + \delta \leq 1$, every (α, δ)-LDP mechanism is $(e^\alpha - 1 + \delta)$-LTVP.*

3.5.2. Identification

We proceed by studying identification of $\chi(P_{VX})$ under privacy protection, and by exhibiting particular subsets of privacy mechanisms, for instance, in the class of (α, δ)-LDP and α-LTVP mechanisms, which guarantee identification.

For some subset $\mathrm{d}\mathcal{P} \subset \mathrm{d}\bar{\mathcal{P}}_{VX}$ of all possible densities of (V, X), let

$$\mathcal{P}_{VZ}(\mathrm{d}\mathcal{P}, Q) \coloneqq \Bigg\{ P : P(B_{\mathsf{v}}, B_{\mathsf{z}}) = \int_{B_{\mathsf{v}}} \int_{\mathfrak{X}} Q(B_{\mathsf{z}} | \, x) p(v, x) \, \mathrm{d}\nu_X(x) \, \mathrm{d}\nu_V(v) \\ \text{holds for all } \; B_{\mathsf{v}} \in \mathscr{F}_{\mathfrak{V}}, B_{\mathsf{z}} \in \mathscr{F}_{\mathfrak{Z}}, \text{ as } p \text{ runs through } \mathrm{d}\mathcal{P} \Bigg\} \tag{3.34}$$

be the set of all possible distributions of (V, Z) generated by a Markov-kernel Q as the density of the distribution of (V, X) varies across $\mathrm{d}\mathcal{P}$. A sufficient and necessary condition for the identification of *every* parameter $\chi : \mathcal{P}_{VX} \to \mathbb{R}, P_{VX} \mapsto \chi(P_{VX})$ from P_{VZ} is the existence of a map $L_Q : \mathcal{P}_{VZ}(\mathrm{d}\mathcal{P}_{VX}, Q) \to \mathcal{P}_{VX}$ such that

$$P_{VX} = L_Q(P_{VZ}). \tag{3.35}$$

We may think of L_Q as a map inverting (3.33) to recover P_{VX} from every P_{VZ} generated by a given Q. If L_Q existed, then we could identify *every* parameter of interest $\chi(P_{VX})$ of the partly unobserved data distribution from the

observed data distribution via

$$\psi(P_{VZ}) \coloneqq \chi(L_Q(P_{VZ})) = \chi(P_{VX}). \tag{3.36}$$

We emphasise 'every', because for some parameters, the existence of an L_Q satisfying (3.35) is not necessary (but clearly sufficient). For example, when χ is only the functional of the marginal P_V and not of P_{VX} — that is, $\chi(P_{VX}) = \bar{\chi}(P_V)$ as above —, then it is not necessary to recover the full P_{VX} from P_{VZ}.

Further, we also require that the image space of Z, $\mathfrak{Z}$, coincide with that of X, so that $\mathfrak{Z} \coloneqq \mathfrak{X}$. Formally arguing about the necessity of this requirement is beyond the scope of this chapter, but we can set out the following intuitive arguments. On one hand, we conjecture that for models parameterised by $\theta \in \mathbb{R}^K$, identification may still be possible even when $\mathfrak{X}$ is a 'richer' set than $\mathfrak{Z}$, provided sufficient amount of information about θ is retained in P_{VZ}; but we do expect that $\mathfrak{Z}$ being at least as 'rich' as $\mathfrak{X}$ is rather necessary in nonparametric models (3.1).[7] This is because a 'poorer' space of Z than that of X could mean a loss of information about X in some region of $\mathfrak{X}$ carried in Z. Now, in models parametrised by $\theta \in \mathbb{R}^K$, regions of $\mathfrak{X}$ where no loss is suffered might still be informative of θ, so that inference is possible. But, by the local nature of nonparametric models, an information loss occurring in a specific region of $\mathfrak{X}$ may not be offset by suffering no loss in other regions of $\mathfrak{X}$.

On the other hand, choosing a 'richer' space for Z than that of X would, intuitively, dilute the signal about X carried in Z, decreasing the precision of inference. Thus, as privacy requirements can be met regardless of $\mathfrak{Z}, \mathfrak{X}$, we require that $\mathfrak{Z} \coloneqq \mathfrak{X}$, which simplifies our exposition.

To illustrate our reasoning, it is instructive to consider the case of a covariate distributed on a finite set. Example 3.3 shows that the existence of the

[7] Considering the privacy mechanism of Hucke (2019) presented in Steinberger (2023, Section 2.2, Display (2.1)) for a binomial model parametrised by $\theta \in (0,1)$, one sees that $|\mathfrak{Z}| = 2 < 3 = |\mathfrak{X}|$. However, a conclusion that $|\mathfrak{Z}| < |\mathfrak{X}|$ could then also suffice for identification would be incorrect. This is because the mechanism in Display (2.1) is a result of an optimisation over all mechanisms Q in order to minimise the asymptotic variance of the resulting estimator of θ; for the identification of θ itself, one has to consider an initial privacy mechanism that ensures the identification of θ, and hence it does not depend thereon. (Steinberger, 2023).

inverse of the Markov kernel viewed as a matrix is sufficient for the identification (3.36) of $\chi(P_{VX})$ from P_{VZ}.

Example 3.3 (Covariate Distributed on a Finite Set)**.** *Suppose that $\mathfrak{X}$ and $\mathfrak{Z}$ are finite sets with elements $x_1, \ldots, x_{|\mathfrak{X}|}$ and $z_1, \ldots, z_{|\mathfrak{Z}|}$, respectively. For the counting measure ν_X, P_{VZ} in (3.33) admits a $\nu_V \times \nu_X$-density*

$$p_{VZ}(v, z) = \sum_{x \in \mathfrak{X}} Q(\{z\} \mid x) p_{VX}(v, x), \quad (v, z) \in \mathfrak{V} \times \mathfrak{Z}. \tag{3.37}$$

Representing Q as the $|\mathfrak{Z}|$-by-$|\mathfrak{X}|$ matrix

$$Q = \begin{bmatrix} Q(\{z_1\} \mid x_1) & Q(\{z_1\} \mid x_2) & \cdots & Q(\{z_1\} \mid x_{|\mathfrak{X}|}) \\ Q(\{z_2\} \mid x_1) & Q(\{z_2\} \mid x_2) & \cdots & Q(\{z_2\} \mid x_{|\mathfrak{X}|}) \\ \vdots & \vdots & \ddots & \vdots \\ Q(\{z_{|\mathfrak{Z}|}\} \mid x_1) & Q(\{z_{|\mathfrak{Z}|}\} \mid x_2) & \cdots & Q(\{z_{|\mathfrak{Z}|}\} \mid x_{|\mathfrak{X}|}) \end{bmatrix}, \tag{3.38}$$

the display (3.37) is equivalent to

$$\bar{p}_{VZ}(v) := \begin{bmatrix} p_{VZ}(v, z_1) \\ p_{VZ}(v, z_2) \\ \vdots \\ p_{VZ}(v, z_{|\mathfrak{Z}|}) \end{bmatrix} = Q \begin{bmatrix} p_{VX}(v, x_1) \\ p_{VX}(v, x_2) \\ \vdots \\ p_{VX}(v, x_{|\mathfrak{X}|}) \end{bmatrix} =: Q \bar{p}_{VX}(v), \quad v \in \mathfrak{V}. \tag{3.39}$$

Suppose that $|\mathfrak{Z}| = |\mathfrak{X}| =: J$, and, for $z, x \in \mathbb{R}^{J \times 1}$, consider

$$z = Qx, \tag{3.40}$$

the system of linear equations in x. The system (3.40) has a unique solution for all $z \in \mathbb{R}^{J \times 1}$ if and only if the matrix Q is invertible, in which case the solution is $x = Q^{-1}z$, where Q^{-1} is the inverse of Q (e.g. Piziak and Odell (2007, Theorems 1.3 and 1.4)). Conclude that if Q is invertible, then $\bar{p}_{VX}(v)$ can be recovered from (3.39) as $\bar{p}_{VX}(v) = Q^{-1}\bar{p}_{VZ}(v)$ for all $v \in \mathfrak{V}$. Hence, if Q is invertible, L_Q in (3.35) exists, is unique, and is completely determined by the matrix Q^{-1}.

It is easy to verify that for the discrete covariate in Example 3.3, the mechanism in Steinberger (2023, Section 2.2) given by the matrix representation

$$Q = \frac{1}{e^\alpha + J - 1} \begin{bmatrix} e^\alpha & 1 & \cdots & 1 \\ 1 & e^\alpha & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & e^\alpha \end{bmatrix}$$

is α-LDP with inverse

$$Q^{-1} = \frac{e^\alpha + J - 1}{e^{2\alpha} + (J-2)e^\alpha - J + 1} \begin{bmatrix} e^\alpha + J - 2 & -1 & \cdots & -1 \\ -1 & e^\alpha + J - 2 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & e^\alpha + J - 2 \end{bmatrix}$$

for $\alpha > 0$. By Lemma 3.2 this translates into $(e^\alpha - 1)$-LTVP for small enough α. Hence, we demonstrated that when the covariates are distributed on a finite set, we can exhibit privacy mechanisms in the class of (α, δ)-LDP and α-LTVP that imply identifiability (3.36) of $\chi(P_{VX})$ from P_{VZ}. In fact, for this finite case, *any* privacy mechanism can be deployed, not restricted to (α, δ)-LDP or α-LTVP, as long as the kernel matrix Q is invertible. This result gives freedom for practitioners to choose whichever privacy mechanism Q is deemed suitable, provided the matrix Q is invertible.

While achieving identification (3.36) through the inversion of (3.33) is relatively straightforward for covariates distributed on a finite set, the inversion for a generic (but absolutely continuous) covariate calls for more creativity. However, recall our discussion following (3.33): with the Dirac measure δ_x we could recover P_{VX} at the expense of no privacy. If we could make sure that privacy is also guaranteed, we would achieve both of our aims of inference and privacy. This motivates the construction of the mechanism

$$Q(B \mid x) \coloneqq \alpha\delta_x(B) + (1-\alpha)\bar{Q}(B), \quad 0 < \alpha < 1, \tag{3.41}$$

where $\bar{Q}$ is a fixed and known probability measure on $\mathfrak{Z} = \mathfrak{X}$ admitting ν_X-density $\bar{q}$ with $\sup_{z\in\mathfrak{X}} \bar{q}(z) < \infty$. It easily follows that (3.41) is an α-LTVP mechanism. The Z drawn from mechanism (3.41) for a unit with covariate

$X = x$ is equal to the covariate x itself with probability α, and it is equal to pure noise drawn from $\bar{Q}$ with probability $1 - \alpha$. Hence, the smaller α, the stricter the privacy.

Lemma 3.8 in Section 3.B establishes the distributional implications of deploying (3.41) as our privacy mechanism. In particular, Lemma 3.8 (ii) and (iii) yield

$$P_{VX}(B_{\mathsf{v}}, B_{\mathsf{x}}) = \frac{1}{\alpha} P_{VZ}(B_{\mathsf{v}}, B_{\mathsf{x}}) - \frac{1-\alpha}{\alpha} \bar{Q}(B_{\mathsf{x}}) P_V(B_{\mathsf{v}}), \quad B_{\mathsf{v}} \in \mathscr{F}_{\mathfrak{V}}, B_{\mathsf{z}} \in \mathscr{F}_{\mathfrak{X}},$$
$$p_{VX}(v, x) = \frac{1}{\alpha} p_{VZ}(v, x) - \frac{1-\alpha}{\alpha} \bar{q}(z) p_V(v), \quad (v, x) \in \mathfrak{V} \times \mathfrak{X},$$

whereby the identification (3.36) of $\chi(P_{VX})$ from P_{VZ} readily follows. To complete our formalism in (3.36), we define the map L_Q as

$$(L_Q P_{VZ})(B_{\mathsf{v}}, B_{\mathsf{x}}) \coloneqq \frac{1}{\alpha} P_{VZ}(B_{\mathsf{v}}, B_{\mathsf{x}}) - \frac{1-\alpha}{\alpha} \bar{Q}(B_{\mathsf{x}}) P_V(B_{\mathsf{v}})$$
$$= \frac{1}{\alpha} P_{VZ}(B_{\mathsf{v}}, B_{\mathsf{x}}) - \frac{1-\alpha}{\alpha} \bar{Q}(B_{\mathsf{x}}) P_{VZ}(B_{\mathsf{v}}, \mathcal{X}), \quad B_{\mathsf{v}} \in \mathscr{F}_{\mathfrak{V}}, B_{\mathsf{z}} \in \mathscr{F}_{\mathfrak{X}},$$

to find that $\chi(P_{VX}) = \chi(L_Q P_{VZ}) = \psi(P_{VZ})$. Note that, suggestive of the notation, L_Q is a linear map, and that it depends on and only on Q in (3.41).

It may seem unsatisfactory that while for the finitely distributed covariate we allow for any invertible privacy mechanism, for the generic covariate we restrict ourselves to (3.41). However, apart from some special cases to be discussed next, it appears difficult to obtain invertibility of (3.33) and thus identification without having an additive Dirac-measure δ_x in $Q(\cdot \,|\, x)$. The special cases arise when $Q(\cdot \,|\, x) = \bar{Q}_{\mathrm{c}}(\cdot \,|\, x)$, where $\bar{Q}_{\mathrm{c}}(\cdot \,|\, x)$ admits a ν_X-density $\bar{q}_{\mathrm{c}}(\cdot \,|\, x)$ (as is also the case for the finitely distributed covariates). Then inverting (3.33) is equivalent to solving a Fredholm integral equation of the first kind (see e.g. Polânin and Manzhirov (1998)) where $\bar{q}_{\mathrm{c}}$ induces a linear integral operator. Such equations are usually ill-posed (Polânin and Manzhirov, 1998), but in a few favourable cases they do admit a unique solution which also retains the nonparametric model class.

One such favourable case occurs when $\mathfrak{Z} = \mathfrak{X} = \mathbb{R}$ and $\bar{q}_{\mathrm{c}}$ corresponds to, for example, the Laplace mechanism satisfying α-LDP, adding Laplace noise from the Laplace density p_ε. Then for the Fourier transform $\mathcal{F}$ and its inverse $\mathcal{F}^{-1}$, we have $p_{VX}(v, x) = (\mathcal{F}^{-1}(w \mapsto \frac{(\mathcal{F} p_{VZ}(v, \cdot))(w)}{(\mathcal{F} p_\varepsilon)(w)}))(x)$ by the convolution theorem. Another favourable case occurs when $\mathfrak{Z} = \mathfrak{X}$ and $\bar{q}_{\mathrm{c}}$ induces

a compact integral operator. Then the equation admits a unique and smooth solution (Kress (2014, Theorem 3.4)). One class of compact linear operators are Hilbert-Schmidt operators (Reed and Simon (1972, Chapter VI.6), Kress (2014, Chapter 3)). The $\bar{q}_c$ induces a Hilbert-Schmidt operator if

$$\int_{\mathfrak{X}\times\mathfrak{X}} \bar{q}_c(z|\,x)^2 \,\mathrm{d}\nu_X(z)\,\mathrm{d}\nu_X(x) < \infty$$

(Reed and Simon (1972, Theorem VI.23)), or if $(z,x) \mapsto \bar{q}_c(z|\,x)$ was continuous and the domain $\mathfrak{Z} = \mathfrak{X} \subset \mathbb{R}^K$ was compact (Kress (2014, Theorem 2.27)).

The first, convolution case is specific to privacy achieved by *additive* noise and image space $\mathbb{R}^K$, which is not suitable for a generic covariate under our consideration, for example when X also contains coordinates distributed on a finite set. The second, compact case also places restrictions on the domain, or require $\int_{\mathfrak{X}\times\mathfrak{X}} \bar{q}_c(z|\,x)^2 \,\mathrm{d}\nu_X(z)\,\mathrm{d}\nu_X(x) < \infty$, which we do not expect to hold unless $\mathfrak{X}$ is compact (for example, when $\mathfrak{X} = \mathbb{R}$ and the mechanism is the additive Laplace noise, then the last integral is infinite). In contrast to these, our mechanism (3.41) allows for more generic covariate types and space $\mathfrak{Z} = \mathfrak{X}$ handled smoothly by a single mechanism.

Summarising our results in this section, we showed that when the covariates are distributed on a finite set, then any privacy mechanism whose matrix representation is invertible is sufficient for the identification of every parameter $\chi(P_{VX})$ from P_{VZ}. If we insist that $|\mathfrak{Z}| = |\mathfrak{X}|$, then this also a necessary condition. Regarding generic covariates, we established that the requirement $\mathfrak{Z} = \mathfrak{X}$ and the adoption of the α-TVPL privacy mechanism (3.41) are sufficient — but potentially not necessary — conditions for the identification of every parameter $\chi(P_{VX})$ from P_{VZ}. For later use, we collect the mechanisms ensuring identification in the set

$$\mathcal{Q}_\psi := \left\{ Q \in \mathcal{Q}(\mathfrak{X} \to \mathfrak{X}) : \begin{cases} |\mathfrak{X}| = J < \infty \text{ and (3.38) is invertible; or} \\ \text{Q is (3.41)} \end{cases} \right\}, \tag{3.42}$$

so that a unique $L_Q : \mathcal{P}_{VZ}(\mathrm{d}\mathcal{P}_{VX}, Q) \to \mathcal{P}_{VX}$ in (3.35) exists if $Q \in \mathcal{Q}_\psi$, thus yielding identification (3.36).

3.5.3. Semiparametric Properties

In this section, we deduce the semiparametric properties of any parameter $\psi(P_{VZ}) = \chi(P_{VX})$ under the model $\mathcal{P}_{VZ}(Q, \mathrm{d}\mathcal{P}_{VX})$ in (3.34), that is, the model for P_{VZ} generated by the model $\mathrm{d}\mathcal{P}_{VX}$ for p_{VX} and a given privacy mechanism $Q \in \mathcal{Q}$. Specifically, we derive the tangent set $\mathcal{T}_{VZ}(Q, \mathrm{d}\mathcal{P}_{VX})$ for the model $\mathcal{P}_{VZ}(Q, \mathrm{d}\mathcal{P}_{VX})$, and the efficient influence function $\tilde{\psi}$ of $\psi(P_{VZ}) = \chi(P_{VX})$ at P_{VZ}, building on the identification results in Section 3.5.2.

Our analysis is general and not limited to the doubly robust class of parameters (3.4) in Section 3.4.1. We only require that the parameter $\chi(P_{VX})$ be $\mathbb{R}$-valued and differentiable at P_{VX} with respect to some tangent set $\mathcal{T}_{VX}$ (Bolthausen et al. (2002, Definition 1.10)). Although some of our results hold for any model $\mathrm{d}\mathcal{P}_{VX}$, our main results are obtained for the nonparametric model $\mathrm{d}\bar{\mathcal{P}}_{VX}$ of (3.1) with tangent set $\mathcal{T}_{VX} = L_2^0(P_{VX})$.

At the core of our results is the linear operator $Q_{\mathcal{X}} : L_2(P_{VZ}) \to L_2(P_{VX})$ defined for a given Markov kernel $Q \in \mathcal{Q}(\mathfrak{X} \to \mathfrak{Z})$ as

$$(Q_{\mathcal{X}} k)(v, x) := \int_{\mathfrak{Z}} k(v, z) Q(\mathrm{d}z \,|\, x). \tag{3.43}$$

The following properties of $Q_{\mathcal{X}}$ play an essential role in the derivation of the tangent set and the efficient influence function.

Lemma 3.3 (Properties of $Q_{\mathcal{X}}$ in (3.43))**.** *The following assertions hold true.*

(i) The operator $Q_{\mathcal{X}}$ is the conditional expectation operator

$$(Q_{\mathcal{X}} k)(v, x) = \mathbb{E}\left[k(V, Z) \,|\, V = v, X = x\right].$$

(ii) The operator $Q_{\mathcal{X}}$ has adjoint $Q_{\mathcal{X}}^* : L_2(P_{VX}) \to L_2(P_{VZ})$,

$$(Q_{\mathcal{X}}^* h)(v, z) = \mathbb{E}\left[h(V, X) \,|\, V = v, Z = z\right].$$

(iii) Change of measure: $P_{VZ} k = P_{VX} Q_{\mathcal{X}} k$ for all $k \in L_2(P_{VZ})$. In particular, if $Q_{\mathcal{X}} : L_2(P_{VZ}) \to L_2(P_{VX})$ has a right-inverse $Q_{\mathcal{X}}^{-\mathrm{R}} : L_2(P_{VX}) \to L_2(P_{VZ})$ so that $Q_{\mathcal{X}}(Q_{\mathcal{X}}^{-\mathrm{R}} h) = h$ for all $h \in L_2(P_{VX})$, then $k := Q_{\mathcal{X}}^{-\mathrm{R}} h$ yields

$$P_{VZ}(Q_{\mathcal{X}}^{-\mathrm{R}}) h = P_{VX} Q_{\mathcal{X}} Q_{\mathcal{X}}^{-\mathrm{R}} h = P_{VX} h.$$

(iv) *If X is distributed on a finite set with $|\mathfrak{Z}| = |\mathfrak{X}| = J$, then the operator $Q_{\mathcal{X}} : L_2(P_{VZ}) \to L_2(P_{VX})$ of (3.43) can be represented in the matrix notation (3.38) as, for all $v \in \mathfrak{V}$,*

$$\begin{bmatrix} (Q_{\mathcal{X}}k)(v,x_1) \\ (Q_{\mathcal{X}}k)(v,x_2) \\ \vdots \\ (Q_{\mathcal{X}}k)(v,x_J) \end{bmatrix} = Q^{\intercal} \begin{bmatrix} k(v,z_1) \\ k(v,z_2) \\ \vdots \\ k(v,z_J) \end{bmatrix},$$

and it has inverse $Q_{\mathcal{X}}^{-1} : L_2(P_{VX}) \to L_2(P_{VZ})$ if and only if Q is invertible, given by, for all $v \in \mathfrak{V}$,

$$\begin{bmatrix} k(v,z_1) \\ k(v,z_2) \\ \vdots \\ k(v,z_J) \end{bmatrix} = (Q^{\intercal})^{-1} \begin{bmatrix} h(v,x_1) \\ h(v,x_2) \\ \vdots \\ h(v,x_J) \end{bmatrix}$$

with $(Q^{\intercal})^{-1} = (Q^{-1})^{\intercal}$.

(v) *Under (3.41), $(Q_{\mathcal{X}}k)(v,x) = \alpha k(v,x) + (1-\alpha)\int_{\mathcal{X}} k(v,z)\bar{Q}(\mathrm{d}z)$. Moreover, $Q_{\mathcal{X}} : L_2(P_{VZ}) \to L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})$ is a bounded, hence continuous, linear operator for the norm*

$$\|h\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})} := \|h\|_{L_2(P_{VX})} + \|h\|_{L_2(P_V\otimes\bar{Q})} \tag{3.44}$$

on $L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})$, where $P_V \otimes \bar{Q}$ is the distribution of a random element $(V, \bar{Z})$ with independent coordinates $V \sim P_V$ and $\bar{Z} \sim \bar{Q}$.

(vi) *Under (3.41), the inverse of $Q_{\mathcal{X}} : L_2(P_{VZ}) \to L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})$ exists, and is, as in Polânin and Manzhirov (1998, Section 4.9-1., Equation 1),*

$$(Q_{\mathcal{X}}^{-1}h)(v,z) = \frac{1}{\alpha}h(v,z) - \frac{1-\alpha}{\alpha}\int_{\mathfrak{X}} h(v,x)\bar{Q}(\mathrm{d}x)$$

for all $h \in L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})$. That is, $Q_{\mathcal{X}}(Q_{\mathcal{X}}^{-1}h) = h$ and $Q_{\mathcal{X}}^{-1}(Q_{\mathcal{X}}k) = k$ for all $h \in L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})$ and all $k \in L_2(P_{VZ})$. Moreover $Q_{\mathcal{X}}^{-1} : L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q}) \to L_2(P_{VZ})$ is a bounded, hence continuous, linear operator for the norm (3.44).

With the help of Lemma 3.3, we can derive the tangent set $\mathcal{T}_{VZ}(Q, \mathrm{d}\mathcal{P}_{VX})$ of the model $\mathcal{P}_{VZ}(Q, \mathrm{d}\mathcal{P}_{VX})$. Lemma 3.4 shows that the tangent set of the model $\mathcal{P}_{VZ}(Q, \mathrm{d}\mathcal{P}_{VX})$ is $Q_{\mathcal{X}}^{*}\mathcal{T}_{VX}$. By Lemma 3.3, this means that

$$\mathcal{T}_{VZ}(Q, \mathrm{d}\mathcal{P}_{VX}) = Q_{\mathcal{X}}^{*}\mathcal{T}_{VX} = \{(v, z) \mapsto \mathbb{E}\left[s(V, X) \mid V = v, Z = z\right] : s \in \mathcal{T}_{VX}\},$$

which is typical of mixture models such as P_{VZ} in (3.33) (see e.g. Van der Vaart (1998, Chapter 25.5)), and is also in agreement with Steinberger (2023, Lemma 3.1).

Lemma 3.4 also shows that $\mathcal{P}_{VZ}(Q, \mathrm{d}\bar{\mathcal{P}}_{VX})$ remains nonparametric if $Q_{\mathcal{X}}$ is invertible.

Lemma 3.4 (Tangent Set of $\mathcal{P}_{VZ}(Q)$)**.** *The following assertions are true.*

(i) Suppose either that X is distributed on a finite set and $Q \in \mathcal{Q}(\mathfrak{X} \to \mathfrak{Z})$ is an arbitrary mechanism whose matrix representation may or may not be invertible, or that the mechanism Q is (3.41). Then $\mathcal{T}_{VZ}(Q, \mathrm{d}\mathcal{P}_{VX}) = \{Q_{\mathcal{X}}^{*}s : s \in \mathcal{T}_{VX}\}$ for $\mathcal{T}_{VX} \subset L_2^0(P_{VX})$ with respect to any model $\mathrm{d}\mathcal{P}_{VX}$.

(ii) Suppose that $\mathcal{T}_{VX} = L_2^0(P_{VX})$ and X is distributed on a finite set with $|\mathfrak{Z}| = |\mathfrak{X}|$. Then $\mathcal{T}_{VZ}(Q, \mathrm{d}\bar{\mathcal{P}}_{VX}) = L_2^0(P_{VZ})$ if and only if $Q \in \mathcal{Q}_{\psi}$.

(iii) Suppose that $\mathcal{T}_{VX} = L_2^0(P_{VX})$ and the privacy mechanism Q is (3.41). Then the closure of $\mathcal{T}_{VZ}(Q, \mathrm{d}\bar{\mathcal{P}}_{VX})$ in $L_2(P_{VZ})$ is $L_2^0(P_{VZ})$.

Next, we use the properties of $Q_{\mathcal{X}}$ in Lemma 3.3 and the tangent set $\mathcal{T}_{VZ}(Q, \mathrm{d}\bar{\mathcal{P}}_{VX}))$ in Lemma 3.4 to derive the efficient influence function $\tilde{\psi}$ of $\psi(P_{VZ})$. We find that, concisely, $\tilde{\psi} = Q_{\mathcal{X}}^{-1}\tilde{\chi}$.

Proposition 3.2 (Efficient Influence Function of $\psi(P_{VZ})$)**.** *Suppose that P_{VX} belongs to the nonparametric model $\bar{\mathcal{P}}_{VX}$ characterised by $\mathrm{d}\bar{\mathcal{P}}_{VX}$ in (3.1), and $\chi(P_{VX})$ has the efficient influence function $\tilde{\chi} : \mathfrak{V} \times \mathfrak{X} \to \mathbb{R}$ at P_{VX}. Assume that $Q \in \mathcal{Q}_{\psi}$, and that $\tilde{\chi} \in L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})$ if Q is (3.41). Then the efficient influence function $\tilde{\psi} : \mathfrak{V} \times \mathfrak{X} \to \mathbb{R}$ in the model $\mathcal{P}_{VZ}(Q, \mathrm{d}\bar{\mathcal{P}}_{VX})$ for P_{VZ}, at P_{VZ}, is*

$$\tilde{\psi} = Q_{\mathcal{X}}^{-1}\tilde{\chi}. \tag{3.45}$$

For covariates distributed on a finite set, the results of Proposition 3.2 are as follows.

Example 3.3 (Covariate Distributed on a Finite Set, continued)**.** *Suppose that $Q \in \mathcal{Q}_\psi$ and $|\mathfrak{Z}| = |\mathfrak{X}| = J$, so that the privacy mechanism can be represented by a J-by-J invertible matrix Q. Then*

$$\begin{bmatrix} \tilde{\psi}(v, z_1) \\ \tilde{\psi}(v, z_2) \\ \vdots \\ \tilde{\psi}(v, z_J) \end{bmatrix} = (Q^\intercal)^{-1} \begin{bmatrix} \tilde{\chi}(v, x_1) \\ \tilde{\chi}(v, x_2) \\ \vdots \\ \tilde{\chi}(v, x_J) \end{bmatrix}$$

for all $v \in \mathfrak{V}$. Because Q is the (3.38) representation of a Markov kernel, each of its columns sums to one. Together with the existence of Q^{-1}, this implies that $(Q^{-1})^\intercal$ acts on a constant vector $(c, c, c, \ldots, c)^\intercal \in \mathbb{R}^{J\times 1}$ as the identity:

$$(Q^{-1})^\intercal (c, c, c, \ldots, c)^\intercal = (c, c, c, \ldots, c)^\intercal. \tag{3.46}$$

Indeed, right multiply $(c, c, c, \ldots, c)Q = (c, c, c, \ldots, c)$ by Q^{-1} and take transpose to see this.

An important implication of Proposition 3.2 is that an asymptotically efficient estimator of $\psi(P_{VZ})$ based on a random sample from $P_{VZ} \in \mathcal{P}_{VZ}(Q, \mathrm{d}\bar{\mathcal{P}}_{VX})$ with a given $Q \in \mathcal{Q}_\psi$ has limiting variance $P_{VZ}\tilde{\psi}^2$, which is equal to

$$P_{VX}[Q_{\mathcal{X}}(\tilde{\psi}^2)] = P_{VX}[Q_{\mathcal{X}}[(Q_{\mathcal{X}}^{-1}\tilde{\chi})(Q_{\mathcal{X}}^{-1}\tilde{\chi})]]$$

by Lemma 3.3(iii) and Proposition 3.2. Note that if we had $Q_{\mathcal{X}} = I$ for the identity operator I, i.e. no privacy $Q(B|\, x) = \delta_x(B)$, then

$$P_{VX}[Q_{\mathcal{X}}[(Q_{\mathcal{X}}^{-1}\tilde{\chi})(Q_{\mathcal{X}}^{-1}\tilde{\chi})]] = P_{VX}\tilde{\chi}^2,$$

the familiar asymptotically efficient variance in the model $\mathcal{P}_{VX}$. Therefore, if we employ an α-LTVP privacy mechanism, we expect that if $\alpha \to 1$, then $P_{VX}[Q_{\mathcal{X}}[(Q_{\mathcal{X}}^{-1}\tilde{\chi})(Q_{\mathcal{X}}^{-1}\tilde{\chi})]] \to P_{VX}\tilde{\chi}^2$, thereby regaining nonprivate efficiency.

3.6. Estimation

In this section, we combine results from Section 3.4 and Section 3.5 to construct efficient estimators of parameters belonging to the rate double robust

class (3.4) when the covariates X are protected by a local privacy mechanism. More specifically, our aim is the construction of an asymptotically efficient estimator of $\psi(P_{VZ})$ based on random samples from P_{VZ}, which belongs to the model $\mathcal{P}_{VZ}(Q, \mathrm{d}\bar{\mathcal{P}}_{VX})$ generated by the nonparametric model $\mathrm{d}\bar{\mathcal{P}}_{VX}$ for p_{VX} and a fixed and known privacy mechanism $Q \in \mathcal{Q}_\psi$. We consider parameters $\psi(P_{VZ}) = \chi(P_{VX})$ in the rate double robust class of (3.4) in Section 3.4.1.

In Section 3.6.1, we present our estimation strategy based on privatised samples and on results in Section 3.5. In Section 3.6.2, building on the results in Section 3.4, we show that the rate double robustness property of $\chi(P_{VX})$ estimated from P_{VX}-samples directly carries over to the estimation of $\psi(P_{VZ})$ from P_{VZ}-samples.

3.6.1. Estimation Strategy

As in Section 3.4.3, we assume that three, pairwise independent, random samples $\bar{\mathcal{S}} = ((V_i, Z_i))_{i\in[n]}$, $\bar{\mathcal{S}}' = ((V_i', Z_i'))_{i\in[n]}$, $\bar{\mathcal{S}}'' = ((V_i'', Z_i''))_{i\in[n]}$ from P_{VZ} are available for inference. The sample $\bar{\mathcal{S}}$ is the privacy protected version of $\mathcal{S} = ((V_i, X_i))_{i\in[n]}$, generated from $\mathcal{S} = ((V_i, X_i))_{i\in[n]}$ by replacing the X_i with a random draw $Z_i \mid X_i \sim Q(\cdot \mid X_i)$, and likewise for $\bar{\mathcal{S}}', \bar{\mathcal{S}}''$. The privacy mechanism Q belongs to $\mathcal{Q}_\psi$ in order to guarantee identification (Section 3.5.2). Analogously to (3.10), we begin with an initial estimator $\psi(\hat{P}_{VZ})$ and correct it as

$$\hat{\psi}_n \coloneqq \psi(\hat{P}_{VZ}) + \bar{\mathbb{P}}_n \hat{\tilde{\psi}} = \psi(\hat{P}_{VZ}) + \frac{1}{n}\sum_{i\in[n]} \hat{\tilde{\psi}}(V_i, Z_i), \tag{3.47}$$

where $\bar{\mathbb{P}}_n \coloneqq \frac{1}{n}\sum_{i\in[n]} \delta_{(V_i,Z_i)}$ is the empirical measure constructed from $\bar{\mathcal{S}}$, and

$$\hat{\tilde{\psi}} \coloneqq Q_{\mathcal{X}}^{-1} \check{\tilde{\chi}}, \tag{3.48}$$

with $\check{\tilde{\chi}}$ being an estimate of $\tilde{\chi}$ in Proposition 3.1:

$$\begin{aligned}\check{\tilde{\chi}}(v,x) \coloneqq{}& \check{r}(v_1,x)(m(v,x) - \check{\mu}_{\mathcal{X}}(v_1,x)) + \frac{\mathbb{1}_{v_2=c}}{\check{p}_{V_2}(c)}(g(v,x) - \check{\gamma}_{\mathcal{V}}(c))\check{e} \\ &+ f(v,x,\check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c)) - \psi(\hat{P}_{VZ}) \\ =:{}& \check{\tilde{\chi}}_0(v,x) - \psi(\hat{P}_{VZ}).\end{aligned} \tag{3.49}$$

Note that $Q_{\mathcal{X}}^{-1}$ is known by construction so it need not be estimated. Here, all the estimates $\psi(\hat{P}_{VZ})$ of $\psi(P_{VZ})$ and

$$\begin{aligned}&\check{\eta} \coloneqq (\check{r}, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c), \check{p}_{V_2}(c), \check{e}) \in L_2(P_{V_1X})^2 \times \Gamma \times \mathbb{R}^2 \text{ of} \\ &\eta = (r, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c), p_{V_2}(c), e)\end{aligned} \tag{3.50}$$

are based on the privatised samples $\bar{\mathcal{S}}, \bar{\mathcal{S}}', \bar{\mathcal{S}}''$ from P_{VZ}. Because $Q \in \mathcal{Q}_\psi$ implies that $P_{VX} = L_Q P_{VZ}$ for a unique and known linear map L_Q, every parameter in η and $\psi(P_{VZ}) = \chi(P_{VX})$ is identified and can be estimated in any case by the plug-in strategy $\check{P}_{VX} \coloneqq L_Q \hat{P}_{VZ}$ where $\hat{P}_{VZ}$ is some estimate of P_{VZ} constructed from random P_{VZ}-samples.

In fact, for each parameter $\eta_j \in \eta$, we could pick a different estimator $\check{P}_{VX,j} \coloneqq L_Q \hat{P}_{VZ,j}$. However, to estimate expectations that are not conditional, that is, $\psi(P_{VZ}) = \chi(P_{VX})$ and e, we restrict the corresponding $\hat{P}_{VZ,j}$ to be the empirical measure, because that gives a characterisation of the estimators similar to the nonprivate setting. Specifically, consider our main estimand of interest in (3.4), $\chi(P_{VX}) = \mathbb{E} f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))$. Let

$$\begin{aligned}&\bar{f}(v, z, \mu, \gamma) \coloneqq (Q_{\mathcal{X}}^{-1}(v, x) \mapsto f(v, x, \mu, \gamma))(v, z) \\ &= \mathbb{E}\left[f(V, X, \mu, \gamma) \,\middle|\, V = v, Z = z \right], \quad (v, z, \mu, \gamma) \in \mathfrak{V} \times \mathfrak{Z} \times L_2(P_{V_1X}) \times \Gamma.\end{aligned}$$

By Lemma 3.3, $\mathbb{E}\left[\bar{f}(V, Z, \mu, \gamma) \,\middle|\, V = v, X = x \right] = f(v, x, \mu, \gamma)$ for all $(v, x, \mu, \gamma) \in \mathfrak{V} \times \mathfrak{X} \times L_2(P_{V_1X}) \times \Gamma$, whence

$$\mathbb{E}\bar{f}(V, Z, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) = \mathbb{E} f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) = \chi(P_{VX})$$

by the tower property of expectations. Now consider $e = \mathbb{E}\partial_\gamma f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))$. Because $Q_{\mathcal{X}}^{-1}$ and differentiation commute, we see that

$$\begin{aligned}&\partial_\gamma \bar{f}(v, z, \mu, \bar{\gamma}) \coloneqq \frac{\partial \bar{f}}{\partial \gamma}(v, z, \mu, \bar{\gamma}) = \frac{\partial}{\partial \gamma}(Q_{\mathcal{X}}^{-1}(v, x) \mapsto f(v, x, \mu, \bar{\gamma}))(v, z) \\ &= (Q_{\mathcal{X}}^{-1}(v, x) \mapsto \partial_\gamma f(v, x, \mu, \bar{\gamma}))(v, z), \quad (v, z, \mu, \bar{\gamma}) \in \mathfrak{V} \times \mathfrak{Z} \times L_2(P_{V_1X}) \times \Gamma.\end{aligned}$$

By Lemma 3.3, $\mathbb{E}\left[\partial_\gamma \bar{f}(V, Z, \mu, \bar{\gamma}) \,\middle|\, V = v, X = x \right] = \partial_\gamma f(v, x, \mu, \bar{\gamma})$ for all $(v, x, \mu, \bar{\gamma}) \in \mathfrak{V} \times \mathfrak{X} \times L_2(P_{V_1X}) \times \Gamma$, whence

$$\mathbb{E}\partial_\gamma \bar{f}(V, Z, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) = \mathbb{E}\partial_\gamma f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) = e$$

Table 3.1: Use of Samples for Estimation

Estimators	Samples		
	$\bar{\mathcal{S}}$	$\bar{\mathcal{S}}'$	$\bar{\mathcal{S}}''$
$\hat{\psi}_n$	✓	✓	✓
$\psi(\hat{P}_{VZ})$	✓	✓	
$\check{e}$		✓	✓
$\check{r}, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c), \check{p}_{V_2}(c)$		✓	

A given sample is used in the construction of a given estimator if and only if ✓ is present in their corresponding cell.

by the tower property of expectations. Since $Q_{\mathcal{X}}^{-1}$ need not be estimated, analogously to (3.13) and (3.15), we estimate $\psi(P_{VZ})$ and e, respectively, as

$$\psi(\hat{P}_{VZ}) \coloneqq \bar{\mathbb{P}}_n \bar{f}(V, Z, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c)) = \frac{1}{n} \sum_{i \in [n]} \bar{f}(V_i, Z_i, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c)), \tag{3.51}$$

$$\check{e} \coloneqq \bar{\mathbb{P}}''_n \partial_\gamma \bar{f}(V, Z, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c)) \coloneqq \frac{1}{n} \sum_{i \in [n]} \partial_\gamma \bar{f}(V''_i, Z''_i, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c)). \tag{3.52}$$

We require that only $\bar{\mathcal{S}}'$ is used to compute the estimates $(\check{r}, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c), \check{p}_{V_2}(c))$. For clarity, Table 3.1 summarises which samples are used for which estimates.

Notice that $Q_{\mathcal{X}}^{-1} c = c$ for a constant function c. As $\psi(\hat{P}_{VZ})$ does not depend on (v, x), we have $\hat{\tilde{\psi}}(v, z) = (Q_{\mathcal{X}}^{-1} \check{\tilde{\chi}}_0)(v, z) - \psi(\hat{P}_{VZ})$, hence

$$\hat{\psi}_n = \psi(\hat{P}_{VZ}) + \bar{\mathbb{P}}_n \hat{\tilde{\psi}} = \bar{\mathbb{P}}_n Q_{\mathcal{X}}^{-1} \check{\tilde{\chi}}_0 = \frac{1}{n} \sum_{i \in [n]} (Q_{\mathcal{X}}^{-1} \check{\tilde{\chi}}_0)(V_i, Z_i)$$

for $\check{\tilde{\chi}}_0$ in (3.49).

3.6.2. Double Robustness

In this section, we repeat the analysis of Section 3.4.3 to understand the behaviour of $\hat{\psi}_n$, the privacy-preserving estimator constructed in Section 3.6.1. Following the decomposition (3.16), (3.17) in Section 3.4.3, we write

$$\sqrt{n}(\hat{\psi}_n - \psi(P_{VZ})) = \sqrt{n} \bar{\mathbb{P}}_n \tilde{\psi} + \sqrt{n}(\bar{\mathbb{P}}_n - P_{VZ})(\hat{\tilde{\psi}} - \tilde{\psi}) + \sqrt{n} \bar{R}_n, \tag{3.53}$$

$$\bar{R}_n \coloneqq \psi(\hat{P}_{VZ}) - \psi(P_{VZ}) + P_{VZ} \hat{\tilde{\psi}}, \tag{3.54}$$

since $P_{VZ}\tilde{\psi} = 0$ as $\tilde{\psi}$ is an influence function. The first term $\sqrt{n}\bar{\mathbb{P}}_n\tilde{\psi} \overset{P_{VZ}}{\rightsquigarrow} \mathcal{N}(0, P_{VZ}\tilde{\psi}^2)$ by the standard central limit theorem.

To control the empirical process term in (3.53), Lemma 3.9 in Section 3.B establishes some technical results concerning norms under P_{VX} and P_{VZ} and the continuity of the operators $Q_{\mathcal{X}}, Q_{\mathcal{X}}^{-1}$. In the light of these technical results, under the consistency and boundedness conditions of Assumption 3.3, Lemma 3.5 shows that the empirical process term in (3.53) vanishes as $o_{P_{VZ}}(1)$. It rests on the same arguments as its nonprivate counterpart, Lemma 3.1, but it relies on the continuity of the operator $Q_{\mathcal{X}}^{-1}$. It is solely because of this that Assumption 3.3 is more involved than its nonprivate counterpart, Assumption 3.1, and that it requires that the whole (V, X) be distributed on a finite set, as opposed to only X be finitely distributed.

Assumption 3.3 (Consistent Privatised Estimators)**.** *Either the privacy mechanism Q is equal to (3.41); or (V, X) is distributed on a finite set with $\inf_{(v,x)\in\mathfrak{V}\times\mathfrak{X}} p_{VX}(v, x) > 0$ holding and $Q \in \mathcal{Q}_\psi$. Let $\iota := 1$ in the former case and $\iota := 0$ in the latter case. Define the norm and the measure*

$$\|\cdot\|_{\mathrm{L}_2} := \|\cdot\|_{L_2(P_{VX})} + \iota\|\cdot\|_{L_2(P_V\otimes\bar{Q})}, \quad P_{\mathrm{L}} := P_{VX} + \iota P_V \otimes \bar{Q},$$

and let $\mathrm{L}_2 := \{f : \mathfrak{V}\times\mathfrak{X} \to \mathbb{R} : \|f\|_{\mathrm{L}_2} < \infty\}$. It holds that $\tilde{\chi}, \tilde{\tilde{\chi}}, r \in \mathrm{L}_2$ and

$$\|f(\cdot, \mu, \gamma) - f(\cdot, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))\|_{\mathrm{L}_2} \to 0 \text{ as } \rho((\mu, \gamma), (\mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))) \to 0, \quad (3.55)$$

$$\|\partial_\gamma f(\cdot, \mu, \gamma) - \partial_\gamma f(\cdot, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))\|_{\mathrm{L}_2} \to 0 \text{ as } \rho((\mu, \gamma), (\mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))) \to 0, \quad (3.56)$$

and

$$\|\check{r} - r\|_{\mathrm{L}_2} = o_{P_{VZ}}(1), \quad (3.57)$$

$$\check{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c) = o_{P_{VZ}}(1), \quad (3.58)$$

$$\check{p}_{V_2}(c) - p_{V_2}(c) = o_{P_{VZ}}(1). \quad (3.59)$$

Further, it either holds that

$$\|m - \mu_{\mathcal{X}}\|_\infty = O(1), \quad (3.60)$$

$$\|\mu_{\mathcal{X}} - \check{\mu}_{\mathcal{X}}\|_\infty = o_{P_{VZ}}(1), \quad (3.61)$$

or that

$$\|m - \check{\mu}_{\mathcal{X}}\|_{\infty} = O_{P_{VZ}}(1), \tag{3.62}$$

$$\|\mu_{\mathcal{X}} - \check{\mu}_{\mathcal{X}}\|_{\mathrm{L}_2} = o_{P_{VZ}}(1), \tag{3.63}$$

$$P_{\mathrm{L}}\left(\left\{(V_1, X) \in \mathfrak{V}_1 \times \mathfrak{X} : |r(V_1, X)| > \bar{R}\right\}\right) = 0 \tag{3.64}$$

for some constant $\bar{R} < \infty$. Above, we may replace $\|\cdot\|_{\mathrm{L}_2}$ norms by $\|.\|_{\infty}$ norms. Likewise, (3.64) *may be replaced by $\|r\|_{\infty} < \infty$.*

Lemma 3.5 (Vanishing Empirical Process Term — Privatised Estimators)**.** *Assume that η is estimated from $\bar{\mathcal{S}}', \bar{\mathcal{S}}''$ as described above and that Assumption 3.3 holds. Then $\check{e} - e = o_{P_{VZ}}(1)$ and $(\bar{\mathbb{P}}_n - P_{VZ})(\hat{\tilde{\psi}} - \tilde{\psi}) = o_{P_{VZ}}\left(n^{-1/2}\right)$.*

Having analysed the first two terms of the decomposition of $\hat{\psi}_n$ in (3.53), we now turn our attention to third, bias term $\bar{R}_n = \psi(\hat{P}_{VZ}) - \psi(P_{VZ}) + P_{VZ}\hat{\tilde{\psi}}$ in (3.54). Theorem 3.1 and Lemma 3.3 give an attractive characterisation of $\bar{R}_n$. Specifically, by (3.48) and Lemma 3.3 (iii), we can change the measure:

$$P_{VZ}\hat{\tilde{\psi}} = P_{VZ}Q_{\mathcal{X}}^{-1}\check{\tilde{\chi}} = P_{VX}\check{\tilde{\chi}}.$$

Therefore, by the identification $\psi(P_{VZ}) = \chi(P_{VX})$ for $Q \in \mathcal{Q}_{\psi}$,

$$\bar{R}_n = \psi(\hat{P}_{VZ}) - \psi(P_{VZ}) + P_{VZ}\hat{\tilde{\psi}} = \psi(\hat{P}_{VZ}) - \chi(P_{VX}) + P_{VX}\check{\tilde{\chi}}.$$

Notice that the expectation of $\check{\tilde{\chi}}$ is taken with respect to P_{VX}. Then Theorem 3.1 implies that

$$\begin{aligned}\bar{R}_n &= \psi(\hat{P}_{VZ}) - \chi(P_{VX}) + P_{VX}\check{\tilde{\chi}} \\ &= -P_{VX}(r - \check{r})(\mu_{\mathcal{X}} - \check{\mu}_{\mathcal{X}}) + (\gamma_{\mathcal{V}}(c) - \check{\gamma}_{\mathcal{V}}(c))\left(\frac{p_{V_2}(c)}{\check{p}_{V_2}(c)}\check{e} - \bar{e}'\right) \\ &\quad - (\gamma_{\mathcal{V}}(c) - \check{\gamma}_{\mathcal{V}}(c))^2\frac{P_{VX}\partial_{\gamma}^2 f(V, X, \check{\mu}_{\mathcal{X}}, \tilde{\gamma}_{\mathcal{V}}(c))}{2},\end{aligned} \tag{3.65}$$

for some $\tilde{\gamma}_{\mathcal{V}}(c)$ between $\gamma_{\mathcal{V}}(c)$ and $\check{\gamma}_{\mathcal{V}}(c)$, and

$$\bar{e}' \coloneqq P_{VX}\partial_{\gamma} f(V, X, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c)) = P_{VZ}\partial_{\gamma}\bar{f}(V, Z, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c)), \tag{3.66}$$

where the last equality is by the construction of $\bar{f}$. Suppose that $\check{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c) = O_{P_{VZ}}\left(n^{-1/2}\right)$, $\check{p}_{V_2}(c) - p_{V_2}(c) = O_{P_{VZ}}\left(n^{-1/2}\right)$, and that $\check{e} - \bar{e}' = o_{P_{VZ}}(1)$ and $P_{VX}\partial_{\gamma}^2 f(V, X, \check{\mu}_{\mathcal{X}}, \check{\gamma}_{\mathcal{V}}(c)) = O_{P_{VZ}}(1)$. Then

$$\bar{R}_n = -P_{VX}(r - \check{r})(\mu_{\mathcal{X}} - \check{\mu}_{\mathcal{X}}) + o_{P_{VZ}}\left(n^{-1/2}\right), \tag{3.67}$$

so that the privacy-preserving estimator $\hat{\psi}_n$ exhibits the same rate double robustness property as the nonprivate $\hat{\chi}_n$. We may summarise our results as follows.

Assumption 3.4 (Rates of Private Estimators)**.** *It holds that*

$$\begin{aligned} P_{VX}((r-\check{r})(\mu_{\mathcal{X}}-\check{\mu}_{\mathcal{X}})) &= o_{P_{VZ}}\left(n^{-1/2}\right), \\ \check{\gamma}_{\mathcal{V}}(c)-\gamma_{\mathcal{V}}(c) &= O_{P_{VZ}}\left(n^{-1/2}\right), \\ \check{p}_{V_2}(c)-p_{V_2}(c) &= O_{P_{VZ}}\left(n^{-1/2}\right), \\ P_{VX}\partial_{\gamma}^{2}f(V,X,\check{\mu}_{\mathcal{X}},\tilde{\gamma}_{\mathcal{V}}(c)) &= O_{P_{VZ}}(1). \end{aligned}$$

Corollary 3.2 (Asymptotic Efficiency of $\hat{\psi}_n$)**.** *Suppose that the privacy mechanism $Q \in \mathcal{Q}_{\psi}$. If the estimators are constructed from the independent random samples $\bar{\mathcal{S}}, \bar{\mathcal{S}}', \bar{\mathcal{S}}''$ as described, and Assumptions 3.3 and 3.4 hold, then $\sqrt{n}(\hat{\psi}_n - \psi(P_{VZ})) \overset{P_{VZ}}{\rightsquigarrow} \mathcal{N}(0, P_{VZ}\tilde{\psi}^2)$ as $n \to \infty$.*

3.7. Estimation of Nuisance Parameters

Section 3.6.2 shows that the privacy-preserving estimator $\hat{\psi}_n$ is asymptotically efficient provided that the nuisance parameters

$$\eta_{\setminus e} \coloneqq (\mu_{\mathcal{X}}, r, \gamma_{\mathcal{V}}(c), p_{V_2}(c))$$

are estimable at appropriate rates described in Assumption 3.4. In particular, if the finite-dimensional parameters $\gamma_{\mathcal{V}}(c), p_{V_2}(c)$ are estimable within an error of $O_{P_{VZ}}\left(n^{-1/2}\right)$, then $\hat{\psi}_n$ continues to enjoy the same rate double robustness property (3.67) as $\hat{\chi}_n$ in (3.31), namely,

$$\bar{R}_n = -P_{VX}(r-\check{r})(\mu_{\mathcal{X}}-\check{\mu}_{\mathcal{X}}) + o_{P_{VZ}}\left(n^{-1/2}\right).$$

In this section, we consider the estimation of $\eta_{\setminus e}$ from the private sample

$$\bar{\mathcal{S}}' = ((V_i', Z_i'))_{i\in[n]},$$

a random sample from P_{VZ} which belongs to the model $\mathcal{P}_{VZ}(Q, \mathrm{d}\bar{\mathcal{P}}_{VX})$ generated by the nonparametric model $\mathrm{d}\bar{\mathcal{P}}_{VX}$ for p_{VX} and a fixed and known privacy mechanism $Q \in \mathcal{Q}_{\psi}$.

In Section 3.7.1, we show that the estimation of the low-dimensional parameters $\gamma_{\mathcal{V}}(c), p_{V_2}(c)$ is trivial. However, the estimation of the infinite-dimensional parameters $\mu_{\mathcal{X}}, r$ is not trivial, even from the nonprivate sample $\mathcal{S}' = ((V_i', X_i'))_{i\in[n]}$; estimation from the private sample $\bar{\mathcal{S}}'$ calls for even more creativity. It is beyond the scope of our current work to completely solve this estimation problem. Nonetheless, in Sections 3.7.2 and 3.7.3, we shed some light on the estimation of $\mu_{\mathcal{X}}$ and r, showing that they can be recast as optimisation problems, which, in turn, lend themselves for a 'natural' estimation strategy from $\bar{\mathcal{S}}'$. In particular, for when one is willing to make functional form assumptions about $\mu_{\mathcal{X}}, r$, we design a privatised generalised method of moments estimator and show that root-n rates are attainable, thereby retaining rate double robustness in the private setting.

3.7.1. Estimation of Low-Dimensional Parameters

First, we focus on the finite-dimensional parameters $p_{V_2}(c), \gamma_{\mathcal{V}}(c)$. Because V is not privatised, $p_{V_2}(c)$ is estimable in the same manner as in the nonprivate case, with

$$\check{p}_{V_2}(c) \coloneqq \hat{p}_{V_2}(c) = N_c/n, \quad N_c \coloneqq \sum_{i\in[n]} \mathbb{1}_{V_{2i}'=c}$$

satisfying $\check{p}_{V_2}(c) - p_{V_2}(c) = O_{P_{VZ}}\left(n^{-1/2}\right)$ of Assumption 3.4 by the standard central limit theorem and Slutsky's lemma. To estimate $\gamma_{\mathcal{V}}(c)$, note that, by definition,

$$\gamma_{\mathcal{V}}(c) = \mathbb{E}\left[g(V,X) \mid V_2 = c\right] = \frac{1}{p_{V_2}(c)} \mathbb{E}\mathbb{1}_{V_2=c} g(V,X),$$

because V_2 is discretely distributed. We just saw that $p_{V_2}(c)$ is well-estimable; to estimate $\mathbb{E}\mathbb{1}_{V_2=c}g(V,X)$, we employ the same strategy as we did for $\psi(P_{VZ})$. Define $g_c(v,x) \coloneqq \mathbb{1}_{v_2=c} g(v,x)$ and $\bar{g}_c(v,z) \coloneqq (Q_{\mathcal{X}}^{-1} g_c)(v,z)$. By Lemma 3.3 (iii), $\mathbb{E}\bar{g}_c(V,Z) = \mathbb{E}g_c(V,X)$. Whence,

$$\check{\gamma}_{\mathcal{V}}(c) \coloneqq \frac{1}{\check{p}_{V_2}(c)} \bar{\mathbb{P}}_n' \bar{g}_c(V,Z) = \frac{1}{\check{p}_{V_2}(c)} \frac{1}{n} \sum_{i\in[n]} \bar{g}_c(V_i', Z_i') \tag{3.68}$$

satisfies $\check{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c) = O_{P_{VZ}}\left(n^{-1/2}\right)$ of Assumption 3.4 by the standard central limit theorem and Slutsky's lemma.

3.7.2. Estimation of the Regression Function

In this section, we study the estimation of

$$\mu_{\mathcal{X}}(v_1, x) = \mathbb{E}\left[m(V, X) \mid V_1 = v_1, X = x\right].$$

As expected, when (V_1, X) is distributed on a finite set, the estimation problem becomes low-dimensional, easily meeting the rate requirements of Assumption 3.4. Indeed, then

$$\mu_{\mathcal{X}}(v_1, x) = \frac{1}{p_{V_1 X}(v_1, x)} \mathbb{E}\mathbb{1}_{V_1 = v_1, X = x} m(V, X).$$

For a given (v_1, x), we can estimate $\mathbb{E}\mathbb{1}_{V_1 = v_1, X = x} m(V, X)$ at root-n rate as we estimated $\mathbb{E}\mathbb{1}_{V_2 = c} g(V, X)$ above for $\gamma_{\mathcal{V}}(c)$. We estimate $p_{V_1 X}(v_1, x)$ by

$$\check{p}_{V_1, X}(v_1, x) \coloneqq \sum_{z \in \mathcal{X}} (Q^{-1})_{z,x} \hat{p}_{V_1, Z}(v_1, z),$$

where $(Q^{-1})_{z,x}$ is the element in the z-th row and x-th column of Q^{-1}, based on the identification of p_{VX} from p_{VZ} in Example 3.3, integrating out $V \setminus V_1$. If (V_1, X) is distributed on a finite set, this automatically translates into uniform convergence across (v_1, x): $\|\check{\mu}_{\mathcal{X}} - \mu_{\mathcal{X}}\|_\infty = O_{P_{VZ}}\left(n^{-1/2}\right)$.

When X is generic, estimation becomes challenging, even when V_1 is distributed on a finite set. Suppose we could make a connection between $\mu_{\mathcal{X}}$ and regressions natively arising as functionals of P_{VZ}, such as $\mu_{\mathcal{Z}}(v_1, z) \coloneqq \mathbb{E}\left[m(V, Z) \mid V_1 = v_1, Z = z\right]$. In this case, we could make modelling assumptions about $\mu_{\mathcal{Z}}$, which is a regression in the observable private data, and capitalise on them, obtaining the required rates for the estimation of $\mu_{\mathcal{X}}$. By the identification $P_{VX} = L_Q P_{VZ}$ we have, of course, such a connection. For the mechanism (3.41), one can show that if V_1 is distributed on a finite set, then

$$\begin{aligned}\mu_{\mathcal{X}}(v_1, x) = &\frac{1}{p_{V_1 X}(v_1, x)} \\ &\times \left\{\frac{1}{\alpha} p_{V_1 Z}(v_1, x) \mu_{\mathcal{Z}}(v_1, x) - \frac{1 - \alpha}{\alpha} \bar{q}(x) P_V \mathbb{1}_{V_1 = v_1} m(V, x)\right\},\end{aligned}$$

where notice that $P_V \mathbb{1}_{V_1 = v_1} m(V, x)$ is a function of x. But then, even if we model $\mu_{\mathcal{Z}}$ correctly, $p_{V_1 Z}$ and $x \mapsto P_V \mathbb{1}_{V_1 = v_1} m(V, x)$ still need to be estimated. Hence, we cannot impose modelling assumption on $\mu_{\mathcal{Z}}$ and then estimate $\mu_{\mathcal{X}}$

fast enough unless we also model $p_{V_1 Z}$ and estimate $x \mapsto P_V \mathbb{1}_{V_1=v_1} m(V,x)$ fast enough.

Alternatively, we can cast the estimation problem as an optimisation one as follows. It is well-known that

$$\mu_{\mathcal{X}} = \arg \min_{\mu \in L_2(P_{V_1} X)} P_{VX}\left[\Delta_\mu^2(V,X)\right], \quad \Delta_\mu^2(v,x) \coloneqq (m(v,x) - \mu(v_1,x))^2. \tag{3.69}$$

If we had access to $\mathcal{S}'$, we could estimate $\mu_{\mathcal{X}}$ as a regularised minimiser of the $\mathcal{S}'$-based empirical loss $\mathbb{P}'_n\left[\Delta_\mu^2(V,X)\right]$ over $L_2(P_{V_1} X)$.

But we have no access to $\mathcal{S}'$. Notwithstanding, Lemma 3.3 (iii) enables us to rewrite (3.69) as

$$\mu_{\mathcal{X}} = \arg \min_{\mu \in L_2(P_{V_1} X)} P_{VZ}\left[\overline{\Delta_\mu^2}(V,Z)\right], \quad \overline{\Delta_\mu^2}(v,z) \coloneqq (Q_{\mathcal{X}}^{-1}(\Delta_\mu^2))(v,z). \tag{3.70}$$

Then we could estimate $\mu_{\mathcal{X}}$ as a regularised minimiser of the $\bar{\mathcal{S}}'$-based empirical loss

$$\bar{\mathbb{P}}'_n\left[\Delta_\mu^2(V,Z)\right] = \frac{1}{n}\sum_{i\in[n]} \overline{\Delta_\mu^2}(V'_i, Z'_i)$$

over $\mu \in L_2(P_{V_1} X)$.

To address the general case of optimising over the whole space $L_2(P_{V_1} X)$ is beyond our current scope. However, we can derive convergence rates for the low-dimensional parametric case. Specifically, suppose that $\mathfrak{V} \times \mathfrak{X} = \mathbb{R}^{K_{\mathcal{V}}} \times \mathbb{R}^{K_{\mathcal{X}}}$ for finite positive integers $K_{\mathcal{V}}, K_{\mathcal{X}}$, and $\mu_{\mathcal{X}} = \mu_{\theta_0}$ belonging to the model

$$\mathcal{M}(\Theta) \coloneqq \left\{\mu_\theta : \theta \in \Theta \subset \mathbb{R}^K\right\} \tag{3.71}$$

for some finite positive integer K. For instance, if X is distributed on a finite set, then the model assumption $\mathcal{M}(\Theta)$ is without loss of generality, for a function with finitely many values can always be represented as a simple function in $\mathcal{M}(\Theta)$.

The identification (3.69) suggests the use of estimators $\hat{\theta}$ arising from empirical loss minimisation. We deploy the method of moment estimator developed by Hansen (1982) and Newey and McFadden (1994), because, as we shall see, it conveniently extends to the private setting. In the nonprivate setting, it has the following properties.

Lemma 3.6 (Nonprivate Generalised Method of Moments (Hansen (1982) and Newey and McFadden (1994))). *Suppose that $\theta_0 \in \Theta \subset \mathbb{R}^K$ is the unique minimiser*

$$\theta_0 = \arg\min_{\theta\in\Theta} P_{VX}\Xi_\theta,$$

for a fixed and known $\Xi_\theta : \mathfrak{V}\times\mathfrak{X} \to \mathbb{R}$, $\theta \in \Theta$. Further suppose that the derivative $\mathrm{D}_\theta\Xi_{\tilde\theta}(v,x)$ of $\theta \mapsto \Xi_\theta(v,x)$ at $\tilde\theta$ exists for all $\tilde\theta \in \Theta$ and all $(v,x) \in \mathfrak{V}\times\mathfrak{X}$. Assume that

(i) The value θ_0 is in the interior of the compact Θ.

(ii) Let

$$\phi_{\tilde\theta}(v,x) \coloneqq \mathrm{D}_\theta\Xi_{\tilde\theta}(v,x)^\intercal, \quad (v,x) \in \mathfrak{V}\times\mathfrak{X},$$

be the $\mathbb{R}^{K\times 1}$-valued derivative of $\theta \mapsto \Xi_\theta(v,x)$ at $\tilde\theta$. The ϕ_θ is measurable for all $\theta \in \Theta$, and

$$\left\| \|\phi_{\theta_0}\|_2^2 \right\|_{L_1(P_{VX})} < \infty. \tag{3.72}$$

(iii) The $\mathbb{R}^{K\times K}$-valued derivative $\dot\phi_{\tilde\theta}(v,x) \coloneqq \mathrm{D}_\theta\phi_{\tilde\theta}(v,x)$ of $\theta \mapsto \phi_\theta(v,x)$ at $\tilde\theta$ exists for all $\tilde\theta \in \Theta$ and all $(v,x) \in \mathfrak{V}\times\mathfrak{X}$. The $\dot\phi_\theta$ is measurable for all $\theta \in \Theta$, and $\theta \mapsto \dot\phi_\theta(v,x)$ is continuous for all $\theta \in \Theta$ and all $(v,x) \in \mathfrak{V}\times\mathfrak{X}$. The expectation of $\dot\phi_{\theta_0}$ exists, and the matrix $P_{VX}\dot\phi_{\theta_0}$ is invertible. Furthermore, in a neighbourhood $\mathrm{Nb}(\theta_0)$ of θ_0,

$$\left\| \sup_{\theta\in\mathrm{Nb}(\theta_0)} \left\|\dot\phi_\theta\right\|_1 \right\|_{L_1(P_{VX})} < \infty, \tag{3.73}$$

where $\left\|\dot\phi_\theta\right\|_1$ is the sum of the absolute values of the entries of $\dot\phi_\theta$.

Let $A_n \in \mathbb{R}^{K\times K}$ be an arbitrary sequence of (possibly P_{VX}-random) matrices with $A_n \overset{P_{VX}}{\to} A_0$ as $n \to \infty$ for a symmetric positive definite A_0. Then the solution $\hat\theta$ to

$$\tilde\theta \mapsto \Lambda_n(\tilde\theta) \coloneqq \left(\mathbb{P}'_n\phi_{\tilde\theta}^\intercal\right) A_n \left(\mathbb{P}'_n\phi_{\tilde\theta}\right) \equiv 0$$

up to a $o_{P_{VX}}\left(n^{-1/2}\right)$ term, that is, $\Lambda_n(\hat\theta) = o_{P_{VX}}\left(n^{-1/2}\right)$, satisfies $\sqrt{n}(\hat\theta - \theta_0) \overset{P_{VX}}{\rightsquigarrow} \mathcal{N}(0,\Sigma)$ as $n \to \infty$, where

$$\Sigma \coloneqq (\dot\Phi^\intercal A_0\dot\Phi)\dot\Phi^\intercal A_0\Phi A_0\dot\Phi(\dot\Phi^\intercal A_0\dot\Phi)^{-1}, \quad \dot\Phi \coloneqq P_{VX}\dot\phi_{\theta_0}, \quad \Phi \coloneqq P_{VX}\phi_{\theta_0}\phi_{\theta_0}^\intercal.$$

Further, let $\xi_\theta : \mathfrak{V} \times \mathfrak{X} \to \mathbb{R}$, $\theta \in \Theta$, *be possibly* P_{VX}*-random functions with the derivative of* $\theta \mapsto \xi_\theta(v,x)$ *at* $\tilde{\theta}$, $\mathrm{D}_\theta \xi_{\tilde{\theta}}(v,x)$, *existent for all* $\tilde{\theta} \in \Theta$ *and all* $(v,x) \in \mathfrak{V} \times \mathfrak{X}$, P_{VX}*-a.s.. If*

$$\left\| \sup_{\tilde{\theta} \in \mathrm{Nb}(\theta_0)} \left\| \mathrm{D}_\theta \xi_{\tilde{\theta}} \right\|_2 \right\|_{L_2(P_{VX})} = O_{P_{VX}}(1), \tag{3.74}$$

where $\left\| \mathrm{D}_\theta \xi_{\tilde{\theta}} \right\|_2$ *is the sum of squared entries of the vector* $\mathrm{D}_\theta \xi_{\tilde{\theta}}$*, then*

$$\left\| \xi_{\hat{\theta}} - \xi_{\theta_0} \right\|_{L_2(P_{VX})} = O_{P_{VX}}\left(n^{-1/2}\right).$$

With the help of Lemma 3.6, we can easily attain the parametric root-n rate in the nonprivate setting.

Corollary 3.3 (Regression — Rate of Nonprivate Estimator $\mu_{\hat{\theta}}$)**.** *Assume that* $\mu_{\mathcal{X}} = \mu_{\theta_0}$*, belonging to the model* $\mathcal{M}(\Theta)$ *of* (3.71)*. Let* A_n, $\Xi_\theta \coloneqq \Delta^2_{\mu_\theta}$,

$$\phi_{\tilde{\theta}}(v,x) \coloneqq \mathrm{D}_\theta \Delta^2_{\mu_{\tilde{\theta}}}(v,x)^\intercal = 2(m(v,x) - \mu_{\tilde{\theta}}(v_1,x))\mathrm{D}_\theta \mu_{\tilde{\theta}}(v_1,x)^\intercal, \quad \xi_\theta \coloneqq \mu_\theta$$

satisfy the conditions of Lemma 3.6 pertaining to A_n, Ξ_θ, $\phi_{\tilde{\theta}}$, ξ_θ *therein, and let* $\hat{\theta}$ *be the solution to the estimating equation*

$$\tilde{\theta} \mapsto \Lambda_n(\tilde{\theta}) \coloneqq \left(\mathbb{P}'_n \phi^\intercal_{\tilde{\theta}}\right) A_n \left(\mathbb{P}'_n \phi_{\tilde{\theta}}\right) \equiv 0$$

satisfying $\Lambda_n(\hat{\theta}) = o_{P_{VX}}\left(n^{-1/2}\right)$*. Then* $\sqrt{n}(\hat{\theta} - \theta_0) \overset{P_{VX}}{\rightsquigarrow} \mathcal{N}(0,\Sigma)$ *as* $n \to \infty$ *for some* Σ*, and*

$$\left\| \hat{\mu}_{\mathcal{X}} - \mu_{\mathcal{X}} \right\|_{L_2(P_{VX})} = O_{P_{VX}}\left(n^{-1/2}\right),$$

where $\hat{\mu}_{\mathcal{X}} \coloneqq \mu_{\hat{\theta}}$*.*

Now we devise a privacy-preserving variant of the method of moments estimator in Lemma 3.6. In fact, Lemma 3.3 (iii) almost immediately gives us a privacy-preserving construction from the sample $\bar{\mathcal{S}}'$.

Proposition 3.3 (Private Generalised Method of Moments)**.** *Consider the context of the nonprivate generalised method of moments in Lemma 3.6 with* Ξ_θ, θ_0, ϕ_θ, $\dot{\phi}_\theta$ *satisfying the conditions therein, but the* $\|\cdot\|_{L_1(P_{VX})}$ *replaced with* $\|\cdot\|_{L_1(P_{VZ})}$ *in* (3.72) *and* (3.73)*. Assume that* P_{VZ} *belongs to the model* $\mathcal{P}_{VZ}(Q,$

$\mathrm{d}\mathcal{P}_{VX}$) *generated by a model* $\mathrm{d}\mathcal{P}_{VX}$ *for* p_{VX} *subject to the parametric assumptions of Lemma 3.6, and by a fixed and known privacy mechanism* $Q \in \mathcal{Q}_\psi$. *Let* $\bar{A}_n \in \mathbb{R}^{K\times K}$ *be an arbitrary sequence of (possibly* P_{VZ}*-random) matrices with* $\bar{A}_n \overset{P_{VZ}}{\to} \bar{A}_0$ *as* $n \to \infty$ *for a symmetric positive definite* $\bar{A}_0$. *Let* $\check{\theta}$ *be the solution to*

$$\tilde{\theta} \mapsto \bar{\Lambda}_n(\tilde{\theta}) := \big(\bar{\mathbb{P}}'_n \bar{\phi}^\intercal_{\tilde{\theta}}\big)\bar{A}_n\big(\bar{\mathbb{P}}'_n \bar{\phi}_{\tilde{\theta}}\big) \equiv 0,$$

up to a $o_{P_{VZ}}\left(n^{-1/2}\right)$ *term, that is,* $\bar{\Lambda}_n(\check{\theta}) = o_{P_{VZ}}\left(n^{-1/2}\right)$, *where* $\bar{\phi}_{\tilde{\theta}} := \mathrm{D}_\theta \bar{\Xi}_{\tilde{\theta}}$ *with* $\bar{\Xi}_\theta := Q_{\mathcal{X}}^{-1}\Xi_\theta$. *Then* $\sqrt{n}(\check{\theta} - \theta_0) \overset{P_{VZ}}{\rightsquigarrow} \mathcal{N}(0, \bar{\Sigma})$ *as* $n \to \infty$, *where*

$$\bar{\Sigma} := (\dot{\Phi}^\intercal \bar{A}_0 \dot{\Phi})\dot{\Phi}^\intercal \bar{A}_0 \bar{\Phi} \bar{A}_0 \dot{\Phi}(\dot{\Phi}^\intercal \bar{A}_0 \dot{\Phi})^{-1}, \quad \dot{\Phi} := P_{VX}\dot{\phi}_{\theta_0}, \quad \bar{\Phi} := P_{VZ}\bar{\phi}_{\theta_0}\bar{\phi}^\intercal_{\theta_0},$$

with the same $\dot{\Phi} = P_{VX}\dot{\phi}_{\theta_0} = P_{VX}\mathrm{D}_\theta \phi_{\theta_0}$ *and* $\phi_{\theta_0} = \mathrm{D}_\theta \Xi^\intercal_{\theta_0}$ *as in Lemma 3.6.*

Further, let $\xi_\theta : \mathfrak{V} \times \mathfrak{X} \to \mathbb{R}$, $\theta \in \Theta$, *be possibly* P_{VZ}*-random functions with the derivative of* $\theta \mapsto \xi_\theta(v, x)$ *at* $\tilde{\theta}$, $\mathrm{D}_\theta \xi_{\tilde{\theta}}(v, x)$, *existent for all* $\tilde{\theta} \in \Theta$ *and all* $(v, x) \in \mathfrak{V} \times \mathfrak{X}$, P_{VX}*-a.s.. If*

$$\left\| \sup_{\tilde{\theta} \in \mathrm{Nb}(\theta_0)} \left\| \mathrm{D}_\theta \xi_{\tilde{\theta}} \right\|_2 \right\|_{L_2(P_{VX})} = O_{P_{VZ}}(1), \tag{3.75}$$

then, $\left\|\xi_{\check{\theta}} - \xi_{\theta_0}\right\|_{L_2(P_{VX})} = O_{P_{VZ}}\left(n^{-1/2}\right)$.

Proposition 3.3 shows that the asymptotic variance of the nonprivate and private generalised method of moments only differ in $A_0, \bar{A}_0$ and $\Phi, \bar{\Phi}$. For instance, if $A_0 = \bar{A}_0 = I_K$ for the $K \times K$ identity matrix I_K, then

$$\begin{aligned}\bar{\Sigma} - \Sigma &= (\dot{\Phi}^\intercal \dot{\Phi})\dot{\Phi}^\intercal(\bar{\Phi} - \Phi)\dot{\Phi}(\dot{\Phi}^\intercal \dot{\Phi})^{-1} \\ &= (\dot{\Phi}^\intercal \dot{\Phi})\dot{\Phi}^\intercal\big(P_{VZ}\bar{\phi}_{\theta_0}\bar{\phi}^\intercal_{\theta_0} - P_{VX}\phi_{\theta_0}\phi^\intercal_{\theta_0}\big)\dot{\Phi}(\dot{\Phi}^\intercal \dot{\Phi})^{-1},\end{aligned}$$

the definiteness of which, however, is, regrettably, not clear.

In Proposition 3.3, the restriction of $\|\cdot\|_{L_1(P_{VZ})}$-bounds in (3.72), and (3.73) instead of the $\|\cdot\|_{L_1(P_{VX})}$-bounds in the nonprivate case of Lemma 3.6 appears to be unavoidable in our construction: while Lemma 3.3 (iv) shows that $\|h\|_{L_p(P_{VX})} \lesssim \|h\|_{L_p(P_{VZ})}$ when the mechanism Q is (3.41), the converse $\|h\|_{L_p(P_{VZ})} \lesssim \|h\|_{L_p(P_{VX})}$ — which is required for the proof — may not hold. In turn, one can replace $\|\cdot\|_{L_p(P_{VZ})}$ with $\|\cdot\|_\infty$-bounds, which could be easier to

verify. For example, in estimating $\mu_{\mathcal{X}} = \mu_{\theta_0}$, if m is bounded and μ_θ is a generalised linear model with continuous second derivative, then the $\|.\|_\infty$-bounds hold provided $\mathfrak{V}_1 \times \mathfrak{X}$ is compact.

Proposition 3.3 implies that the root-n rate for $\mu_{\mathcal{X}}$ is also attainable in the private setting.

Corollary 3.4 (Regression — Rate of Private Estimator $\mu_{\check{\theta}}$)**.** *Assume that P_{VZ} belongs to the model $\mathcal{P}_{VZ}(Q, \mathrm{d}\mathcal{P}_{VX})$ generated by a model $\mathrm{d}\mathcal{P}_{VX}$ for p_{VX} subject to the parametric assumption $\mu_{\mathcal{X}} = \mu_{\theta_0}$ of (3.71), and by a fixed and known privacy mechanism $Q \in \mathcal{Q}_\psi$. Let $\bar{A}_n$, $\Xi_\theta \coloneqq \Delta^2_{\mu_\theta}$,*

$$\phi_{\tilde{\theta}}(v,x) \coloneqq \mathrm{D}_\theta \Delta^2_{\mu_{\tilde{\theta}}}(v,x)^\intercal = 2(m(v,x) - \mu_{\tilde{\theta}}(v_1,x))\mathrm{D}_\theta \mu_{\tilde{\theta}}(v_1,x)^\intercal, \quad \xi_\theta \coloneqq \mu_\theta$$

satisfy the conditions of Proposition 3.3 pertaining to $\bar{A}_n$, Ξ_θ, $\phi_{\tilde{\theta}}$, ξ_θ therein. Let $\check{\theta}$ be the solution to the estimating equation

$$\tilde{\theta} \mapsto \bar{\Lambda}_n(\tilde{\theta}) \coloneqq \big(\bar{\mathbb{P}}'_n \bar{\phi}^\intercal_{\tilde{\theta}}\big) \bar{A}_n \big(\bar{\mathbb{P}}'_n \bar{\phi}_{\tilde{\theta}}\big) \equiv 0,$$

satisfying $\bar{\Lambda}_n(\check{\theta}) = o_{P_{VZ}}\left(n^{-1/2}\right)$, where $\bar{\phi}_{\tilde{\theta}} \coloneqq \mathrm{D}_\theta \bar{\Xi}_{\tilde{\theta}}$ with $\bar{\Xi}_\theta \coloneqq Q_{\mathcal{X}}^{-1} \Xi_\theta$. Then $\sqrt{n}(\check{\theta} - \theta_0) \overset{P_{VZ}}{\rightsquigarrow} \mathcal{N}(0, \bar{\Sigma})$ as $n \to \infty$ for some $\bar{\Sigma}$, and $\check{\mu}_{\mathcal{X}} \coloneqq \mu_{\check{\theta}}$ satisfies

$$\|\check{\mu}_{\mathcal{X}} - \mu_{\mathcal{X}}\|_{L_2(P_{VX})} = O_{P_{VZ}}\left(n^{-1/2}\right).$$

3.7.3. Estimation of the Riesz Representer

In this section, we consider the estimation of the Riesz representer r. While, arguably, the concept of a Riesz representer is not particularly intuitive, it can coincide with well-understood parameters. For instance, let us recall from Example 3.1 that the Riesz representer of the average treatment effect is

$$r(d,x) = \frac{d}{\pi_{\mathcal{X}}(1\,|\,x)} - \frac{1-d}{1-\pi_{\mathcal{X}}(1\,|\,x)}.$$

Clearly, here, the estimation of r is essentially equivalent to that of the propensity score $\pi_{\mathcal{X}}$ — an endeavour well-known in causal inference.

Before we analyse the estimation of r from the privatised sample $\bar{\mathcal{S}}'$ in Section 3.7.3, it is insightful to analyse the estimation problem in the nonprivate case from the sample $\mathcal{S}'$.

Nonprivate Estimation of the Riesz Representer

Suppose that we could use the nonprivate sample $\mathcal{S}'$ for inference. To construct estimators, we derive an analogue to (3.69), motived by Rotnitzky et al. (2021, Theorem 2 (v)). For brevity, we continue omitting the dependence of the representer on P_{VX} in our notation.

Lemma 3.7 (Identification of r)**.** *For all $\gamma \in \Gamma$, the Riesz representer r_γ of*

$$L_2(P_{V_1}X) \ni \mu \mapsto P_{VX} f(V, X, \mu, \gamma)$$

satisfies

$$r_\gamma = \arg \min_{h \in L_2(P_{V_1}X)} P_{VX} \Upsilon_{\gamma,h}, \quad \Upsilon_{\gamma,h}(v, x) \coloneqq h(v_1, x)^2 - 2f(v, x, h, \gamma). \tag{3.76}$$

Whence,

$$r = \arg \min_{h \in L_2(P_{V_1}X)} P_{VX} \Upsilon_{\gamma_{\mathcal{V}}(c),h}.$$

Compared to (3.69) for the regression $\mu_{\mathcal{X}}$, however, there is a difference: the dependence of the objective function on the unknown parameter $\gamma_{\mathcal{V}}(c)$. Hence, we could first estimate $\gamma_{\mathcal{V}}(c)$ with $\hat{\gamma}_{\mathcal{V}}(c)$ in (3.30) and then r as a regularised minimiser of the $\mathcal{S}'$-based empirical loss

$$\mathbb{P}'_n \Upsilon_{\hat{\gamma}_{\mathcal{V}}(c),h}(V, X) = \frac{1}{n} \sum_{i \in [n]} \Upsilon_{\hat{\gamma}_{\mathcal{V}}(c),h}(V'_i, X'_i)$$

over $h \in L_2(P_{V_1}X)$.

Similarly to $\mu_{\mathcal{X}}$, it is beyond our current scope to study optimisation over the whole set $L_2(P_{V_1}X)$. Nonetheless, we investigate a parametric model for r. Suppose again that $\mathfrak{V} \times \mathfrak{X} = \mathbb{R}^{K_{\mathcal{V}}} \times \mathbb{R}^{K_{\mathcal{X}}}$ for finite positive integers $K_{\mathcal{V}}, K_{\mathcal{X}}$, and that the Riesz representer r_γ of $\mu \mapsto P_{VX} f(V, X, \mu, \gamma)$ is $r_\gamma = r_{\gamma,\theta_0}$ for each $\gamma \in \Gamma$ *uniformly*, where r_{γ,θ_0} belongs to the model

$$\mathcal{R}_\gamma(\Theta) \coloneqq \left\{ r_{\gamma,\theta} : \theta \in \Theta \subset \mathbb{R}^K \right\}, \tag{3.77}$$

with the map $(\gamma, \theta) \mapsto r_{\gamma,\theta}$ known, i.e. the functional form of r_γ is known. This is equivalent to $\mathcal{R}_\gamma(\Theta) = \{\vartheta(\gamma, \rho_\theta(\cdot)) : \theta \in \Theta\}$ for a known and fixed sequence

functions $\rho_\theta : \mathfrak{V}_1 \times \mathfrak{X} \to \mathfrak{R}$, for some codomain $\mathfrak{R}$, and for a known and fixed function $\vartheta : \Gamma \times \mathfrak{R} \to \mathbb{R}$. A prototypical example is a parametric model for the propensity score in inferring the average treatment effect on the treated.

Example 3.2 (Average Treatment Effect on the Treated, continued)**.** *Recall that for $\gamma_{\mathcal{V}}(c) \coloneqq \bar{\gamma} \coloneqq \mathbb{E}D = p_1$, the Riesz representer for $\mathbb{E}\left[Y^0 \,\middle|\, D = 1\right]$ is $r(d,x) = \frac{1-d}{\gamma_{\mathcal{V}}(c)} \frac{\pi_{\mathcal{X}}(1|\,x)}{1-\pi_{\mathcal{X}}(1|\,x)}$, and for $\mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D = 1\right]$, it is $r(d,x) = \frac{d}{\gamma_{\mathcal{V}}(c)} - \frac{1-d}{\gamma_{\mathcal{V}}(c)} \frac{\pi_{\mathcal{X}}(1|\,x)}{1-\pi_{\mathcal{X}}(1|\,x)}$. Suppose that the propensity score $\pi_{\mathcal{X}}(1|\,x) = \pi_{\theta_0}(x)$ for some π_{θ_0} in the presumed model $\left\{\pi_\theta : \theta \in \Theta \subset \mathbb{R}^K\right\}$, for instance, the logistic model $\pi_\theta(x) = (1 + \exp(-\theta^\intercal x))^{-1}$. Then*

$$r_{\gamma,\theta_0}(d,x) = \frac{1-d}{\gamma} \frac{\pi_{\theta_0}(x)}{1-\pi_{\theta_0}(x)}, \quad r_{\gamma,\theta_0}(d,x) = \frac{d}{\gamma} - \frac{1-d}{\gamma} \frac{\pi_{\theta_0}(x)}{1-\pi_{\theta_0}(x)}.$$

are the Riesz representers for $\mathbb{E}\left[Y^0 \,\middle|\, D = 1\right]$ and $\mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D = 1\right]$, respectively, for any $\gamma = \gamma_{\mathcal{V}}(c) = \bar{\gamma} = \mathbb{E}D$. Conversely, $r_{\gamma_{\mathcal{V}}(c),\theta}$ is not a Riesz representer unless $\theta = \theta_0$. Let $\lambda_{0,\theta}(d,x) \coloneqq (1-d)\frac{\pi_\theta(x)}{1-\pi_\theta(x)}$ and $\lambda_{1,\theta}(d,x) \coloneqq d - (1-d)\frac{\pi_\theta(x)}{1-\pi_\theta(x)}$. Then the models for the Riesz representers are

$$\mathcal{R}_\gamma(\Theta) = \left\{\gamma^{-1}\lambda_{0,\theta} : \theta \in \Theta\right\}, \quad \mathcal{R}_\gamma(\Theta) = \left\{\gamma^{-1}\lambda_{1,\theta} : \theta \in \Theta\right\}$$

for $\mathbb{E}\left[Y^0 \,\middle|\, D = 1\right]$ and $\mathbb{E}\left[Y^1 - Y^0 \,\middle|\, D = 1\right]$, respectively.

In the parametric model $\mathcal{R}_\gamma(\Theta)$, the Riesz representer r_{γ,θ_0} is identified by Lemma 3.7 as

$$\begin{aligned} r_{\gamma,\theta_0} &= \arg \min_{\rho \in \mathcal{R}_\gamma(\Theta)} P_{VX} \Upsilon_{\gamma,\rho}(V,X) \\ &= \arg \min_{\rho \in \mathcal{R}_\gamma(\Theta)} P_{VX} \Big[\rho(v_1,x)^2 - 2f(V,X,\rho,\gamma)\Big]. \end{aligned} \tag{3.78}$$

Importantly, notice that for all $\gamma \in \Gamma$ uniformly, $\theta = \theta_0$, and only $\theta = \theta_0$, gives the Riesz representer in the model $\mathcal{R}_\gamma(\Theta)$. An implication is that the previous display is then equivalent to

$$\theta_0 = \arg \min_{\theta \in \Theta} P_{VX} \Upsilon_{\gamma,r_{\gamma,\theta}}(V,X) = \arg \min_{\theta \in \Theta} P_{VX} \Big[r_{\gamma,\theta}(v_1,x)^2 - 2f(V,X,r_{\gamma,\theta},\gamma)\Big], \quad \text{for all } \gamma \in \Gamma. \tag{3.79}$$

Thus, θ_0 does *not* depend on γ. This permits us to infer θ_0 without suffering any bias from the unknown $\gamma_{\mathcal{V}}(c)$. Specifically, we can take an arbitrary, known $\gamma_0 \in \Gamma$ and use (3.79) to infer θ_0 with some estimator $\hat{\theta}_{\gamma_0}$, and then set $\hat{r} := r_{\hat{\gamma}_{\mathcal{V}}(c),\hat{\theta}_{\gamma_0}}$ for $\hat{\gamma}_{\mathcal{V}}(c)$ in (3.30). If $(\gamma,\theta) \mapsto r_{\gamma,\theta}$ is smooth enough, then for a method of moment estimator $\hat{\theta}_{\gamma_0}$, as for $\hat{\mu}_{\mathcal{X}}$, an analogue to Corollary 3.3 holds.

Corollary 3.5 (Riesz Representer — Rate of Nonprivate Estimator $r_{\hat{\gamma}_{\mathcal{V}}(c),\hat{\theta}_{\gamma_0}}$)**.** *Suppose that $r_\gamma = r_{\gamma,\theta_0}$ for the model $\mathcal{R}_\gamma(\Theta)$ of (3.77). Assume that*

(i) The map $(\gamma,\theta) \mapsto r_{\gamma,\theta}(v_1,x)$ is differentiable for all $(v_1,x) \in \mathfrak{V}_1 \times \mathfrak{X}$ at all $(\gamma,\theta) \in \Gamma \times \Theta$ with partial derivatives $\partial_\gamma r_{\tilde{\gamma},\tilde{\theta}}(v_1,x)$ with respect to γ, and $\mathrm{D}_\theta r_{\tilde{\gamma},\tilde{\theta}}(v_1,x)$ with respect to θ, at $(\tilde{\gamma},\tilde{\theta}) \in \Gamma \times \Theta$.

(ii) The map $\theta \mapsto \Upsilon_{\gamma,r_{\gamma,\theta}}(v,x) = r_{\gamma,\theta}(v_1,x)^2 - 2f(v,x,r_{\gamma,\theta},\gamma)$ is differentiable for all $(v,x,\gamma) \in \mathfrak{V} \times \mathfrak{X} \times \Gamma$ at all $\theta \in \Theta$ with $\mathbb{R}^{K\times 1}$-valued derivative $\mathrm{D}_\theta \Upsilon_{\gamma,r_{\gamma,\tilde{\theta}}}(v,x)^\intercal =: \phi_{\gamma,\tilde{\theta}}(v,x)$ at $\tilde{\theta} \in \Theta$.

Fix an arbitrary, known $\gamma_0 \in \Gamma$, and let $\hat{\theta}_{\gamma_0}$ be the solution to the estimating equation

$$\tilde{\theta} \mapsto \Lambda_n(\tilde{\theta}) := \left(\mathbb{P}_n' \phi_{\gamma_0,\tilde{\theta}}^\intercal\right) A_n \left(\mathbb{P}_n' \phi_{\gamma_0,\tilde{\theta}}\right) \equiv 0$$

satisfying $\Lambda_n(\hat{\theta}_{\gamma_0}) = o_{P_{VX}}\left(n^{-1/2}\right)$. If $\Xi_\theta := \Upsilon_{\gamma_0,r_{\gamma_0,\theta}}$, $\phi_\theta := \phi_{\gamma_0,\theta}$, and A_n satisfy the conditions of Lemma 3.6 pertaining to $(\Xi_\theta,\phi_\theta,A_n)$ therein, then $\sqrt{n}(\hat{\theta}_{\gamma_0} - \theta_0) \overset{P_{VX}}{\rightsquigarrow} \mathcal{N}(0,\Sigma_{\gamma_0})$ as $n \to \infty$ for some Σ_{γ_0}. If, in addition,

$$\left\| \partial_\gamma r_{\tilde{\gamma}_{\mathcal{V}}(c),\tilde{\theta}_{\gamma_0}} \right\|_{L_2(P_{VX})} = O_{P_{VX}}(1), \tag{3.80}$$

$$\left\| \left\| \mathrm{D}_\theta r_{\tilde{\gamma}_{\mathcal{V}}(c),\tilde{\theta}_{\gamma_0}} \right\|_2 \right\|_{L_2(P_{VX})} = O_{P_{VX}}(1) \tag{3.81}$$

for some $(\tilde{\gamma}_{\mathcal{V}}(c),\tilde{\theta}_{\gamma_0})$ between $(\gamma_{\mathcal{V}}(c),\theta_0)$ and $(\hat{\gamma}_{\mathcal{V}}(c),\hat{\theta}_{\gamma_0})$, then

$$\|\hat{r} - r\|_{L_2(P_{VX})} = O_{P_{VX}}\left(n^{-1/2}\right),$$

where $\hat{r} := r_{\hat{\gamma}_{\mathcal{V}}(c),\hat{\theta}_{\gamma_0}}$, $r = r_{\gamma_{\mathcal{V}}(c),\theta_0}$ for the model in (3.77) and $\hat{\gamma}_{\mathcal{V}}(c)$ in (3.30).

Private Estimation of the Riesz Representer

In this section, we study the estimation of the Riesz representer r, when we only have access to the privatised sample $\bar{\mathcal{S}}'$. Similarly to the regression $\mu_{\mathcal{X}}$,

Lemma 3.3 (iii) again allows us to study the estimation similarly to the non-private case.

In particular, by Lemma 3.3 (iii), we can identify r from P_{VZ}, rewriting (3.76) as

$$r = \arg \min_{h \in L_2(P_{V_1}X)} P_{VZ} \bar{\Upsilon}_{\gamma_{\mathcal{V}}(c),h}(V,Z), \quad \bar{\Upsilon}_{\gamma_{\mathcal{V}}(c),h}(v,z) \coloneqq (Q_{\mathcal{X}}^{-1} \Upsilon_{\gamma_{\mathcal{V}}(c),h})(v,z).$$

With $\gamma_{\mathcal{V}}(c)$ unknown, we estimate it with $\check{\gamma}_{\mathcal{V}}(c)$ in (3.68), and then we could estimate r as a regularised minimiser of the $\bar{\mathcal{S}}'$-based empirical loss

$$\bar{\mathbb{P}}'_n \bar{\Upsilon}_{\check{\gamma}_{\mathcal{V}}(c),h}(V,Z) = \frac{1}{n} \sum_{i \in [n]} \bar{\Upsilon}_{\check{\gamma}_{\mathcal{V}}(c),h}(V'_i, Z'_i)$$

over $h \in L_2(P_{V_1}X)$.

With further analysis of such generosity beyond our scope, we turn our attention to the parametric model for r as in the nonprivate case. Assuming the model $\mathcal{R}_\gamma(\Theta)$ in (3.77) for the Riesz representer r_γ in Lemma 3.7, we obtain a result in the same vein as Corollaries 3.4 and 3.5. Corollary 3.6 shows that for a smooth enough model, a rate $\|\check{r} - r\|_{L_2(P_{VX})} = O_{P_{VZ}}\left(n^{-1/2}\right)$ is attainable.

Corollary 3.6 (Riesz Representer — Rate of Private Estimator $r_{\check{\gamma}_{\mathcal{V}}(c),\tilde{\theta}}$)**.** *Assume that P_{VZ} belongs to the model $\mathcal{P}_{VZ}(Q, \mathrm{d}P_{VX})$ generated by a model $\mathrm{d}P_{VX}$ for p_{VX} subject to the parametric assumption $r_\gamma = r_{\gamma,\theta_0}$ of (3.77), and by a fixed and known privacy mechanism $Q \in \mathcal{Q}_\psi$. Fix an arbitrary, known $\gamma_0 \in \Gamma$, and let $\check{\theta}_{\gamma_0}$ be the solution to the estimating equation*

$$\tilde{\theta} \mapsto \bar{\Lambda}_n(\tilde{\theta}) \coloneqq \left(\bar{\mathbb{P}}'_n \bar{\phi}_{\gamma_0,\tilde{\theta}}\right) \bar{A}_n \left(\bar{\mathbb{P}}'_n \bar{\phi}_{\gamma_0,\tilde{\theta}}\right) \equiv 0$$

satisfying $\bar{\Lambda}_n(\check{\theta}_{\gamma_0}) = o_{P_{VZ}}\left(n^{-1/2}\right)$, with $\bar{\phi}_{\gamma,\tilde{\theta}} \coloneqq \mathrm{D}_\theta \bar{\Upsilon}_{\gamma,r_{\gamma,\tilde{\theta}}}$, $\bar{\Upsilon}_{\gamma,r_{\gamma,\tilde{\theta}}} \coloneqq Q_{\mathcal{X}}^{-1} \Upsilon_{\gamma,r_{\gamma,\tilde{\theta}}}$ and $\bar{A}_n$ an arbitrary matrix satisfying the conditions of Proposition 3.3. Let $\phi_{\gamma,\tilde{\theta}} \coloneqq \mathrm{D}_\theta \Upsilon^{\intercal}_{\gamma,r_{\gamma,\tilde{\theta}}}$, and suppose that $\phi_\theta \coloneqq \phi_{\gamma_0,\theta}$ and $\Xi_\theta \coloneqq \Upsilon_{\gamma_0,r_{\gamma_0,\theta}}$ satisfy the conditions of Proposition 3.3 pertaining to ϕ_θ and Ξ_θ therein. Then $\sqrt{n}(\check{\theta}_{\gamma_0} - \theta_0) \overset{P_{VZ}}{\rightsquigarrow} \mathcal{N}(0, \bar{\Sigma}_{\gamma_0})$ as $n \to \infty$ for some $\bar{\Sigma}_{\gamma_0}$. If, in addition,

$$\left\| \partial_\gamma r_{\check{\gamma}_{\mathcal{V}}(c),\tilde{\theta}_{\gamma_0}} \right\|_{L_2(P_{VX})} = O_{P_{VZ}}(1), \tag{3.82}$$

$$\left\| \left\| \mathrm{D}_\theta r_{\check{\gamma}_{\mathcal{V}}(c),\tilde{\theta}_{\gamma_0}} \right\|_2 \right\|_{L_2(P_{VX})} = O_{P_{VZ}}(1) \tag{3.83}$$

for some $(\tilde{\gamma}_{\mathcal{V}}(c), \tilde{\theta}_{\gamma_0})$ *between* $(\gamma_{\mathcal{V}}(c), \theta_0)$ *and* $(\check{\gamma}_{\mathcal{V}}(c), \check{\theta}_{\gamma_0})$*, then*

$$\|\check{r} - r\|_{L_2(P_{VX})} = O_{P_{VZ}}\left(n^{-1/2}\right),$$

where $\check{r} := r_{\check{\gamma}_{\mathcal{V}}(c), \check{\theta}_{\gamma_0}}$*,* $r = r_{\gamma_{\mathcal{V}}(c), \theta_0}$ *for the model in* (3.77) *and* $\check{\gamma}_{\mathcal{V}}(c)$ *in* (3.68)*.*

In summary of this section, we studied the estimation of the nuisance parameters $\eta_{\setminus e} = (\mu_{\mathcal{X}}, r, \gamma_{\mathcal{V}}(c), p_{V_2}(c))$. We found that the estimation of the low-dimensional parameters $\gamma_{\mathcal{V}}(c), p_{V_2}(c)$ is trivial, both in the nonprivate as well as in the private setting. Estimation of the infinite-dimensional parameters $\mu_{\mathcal{X}}, r$ remains challenging. However, identifying $\mu_{\mathcal{X}}, r$ as solutions to certain optimisation problems, we saw that their estimation is not particularly more difficult in the private than in the nonprivate setting. Specifically, we showed that under parametric assumptions on $\mu_{\mathcal{X}}, r$, they can be estimated at root-n rates. Thus demonstrating that the rate double robustness property of trading off parametric assumptions on $\mu_{\mathcal{X}}$ for a flexible but slower estimation of r, and the other way around, which holds in the nonprivate setting, prevails in the private setting.

3.8. Conclusion

In this chapter, we reconciled local privacy protection — where the covariates of *each* unit in the sample are privacy protected — and rate double robust inference. First focusing on the nonprivate-data model, we proposed a novel class of estimands which enjoys a rate double robustness property in the nonparametric model, encompassing various estimands of interest. For example, the average treatment effect on the treated is included in our proposed class, facilitating its asymptotically efficient *and* double robust estimation. Our class extends already existent ones by including a potentially nonlinear but smooth dependence on a low-dimensional regression in a way that double robustness is preserved. For example, in the case of the average treatment effect on the treated, this regression is simply the marginal probability of treatment. The low-dimensionality of the regression entering nonlinearly our estimand is key to our construction: owing to its root-n estimability, we can retain double robustness.

Next, we turned towards the private setting, where the covariates are protected by a local privacy mechanism. We presented conditions for the privacy mechanism that guarantee the identifiability of estimands, and derived the semiparametric properties of the resulting private-data model. When the covariates are distributed on a finite set, the class of admissible mechanisms is richer; when the covariates are generic (but absolutely continuous), we restricted our attention to a particular total-variationally private mechanism. This latter mechanism leaves the covariates unchanged with some small probability α, and sets them to pure noise with probability $1 - \alpha$. Although one may raise objections grounded in privacy concerns against this α-probability identity, it is rather powerful with respect to statistical inference. First, it allows for generic image space of the covariates as opposed to restricting them, for instance, to some subset of $\mathbb{R}^K$. Second, it allows for the identification of *any* parameter from the private-data model if and only if the parameter is identified in the nonprivate-data model. Third, we showed that under this mechanism (or under any admissible mechanism for a finitely distributed covariate), the semiparameric properties of the nonprivate-data model carry over to the private-data model. Indeed, one of our main results is to show that the rate double robustness property of our proposed class of parameters continue to hold in the private setting.

Finally, we investigated the estimation of nuisance parameters which govern the double robustness property, with special emphasis on the two infinite-dimensional parameters which arise from our parameter class. It is left for future work to fully address this infinite-dimensional estimation. However, methods already present in the literature, for instance in Chernozhukov et al. (2022), may as well be adaptable to our setting with due care, as we show that estimation of these infinite-dimensional parameters are not particularly harder in the private than in the nonprivate setting. That is, we showed how empirical risk minimiser procedures in the nonprivate setting can be directly translated to the private setting. Moreover, we showed that under functional-form assumptions about either one of the infinite-dimensional parameters, one can afford a more flexible estimation procedure for the other one. For example, we witnessed the well-known trade-off in inferring average treatment effects.

Namely, one may entertain a parametric model for the propensity score, but deploy a nonparametric estimator for the outcome regression, or the other way around. We achieved this by developing a privatised version of the method of moments estimator — a development interesting in its own right.

3.A. Double Robust Inference

This section contains the proof of results in Section 3.4 — Proposition 3.1, Lemma 3.1, and Theorem 3.1 —, and, in Section 3.A.1, some additional results (Proposition 3.4) for finitely distributed covariates.

Proof of Proposition 3.1. First, note that by the definition of $\mu_{\mathcal{X}}, \gamma_{\mathcal{V}}$,

$$\begin{aligned}
&\mathbb{E}r(V_1, X)(m(V, X) - \mu_{\mathcal{X}}(V_1, X)) \\
&= \mathbb{E}\mathbb{E}\left[r(V_1, X)(m(V, X) - \mu_{\mathcal{X}}(V_1, X)) \mid V_1, X\right] \\
&= \mathbb{E}r(V_1, X)\mu_{\mathcal{X}}(V_1, X) - \mathbb{E}r(V_1, X)\mu_{\mathcal{X}}(V_1, X) = 0, \\
\mathbb{E}\mathbb{1}_{V_2=c}(g(V, X) - \gamma_{\mathcal{V}}(c)) &= \mathbb{E}\left[\mathbb{1}_{V_2=c}g(V, X)\right] - p_{V_2}(c)\gamma_{\mathcal{V}}(c) \\
&= p_{V_2}(c)\mathbb{E}\left[g(V, X) \mid V_2 = c\right] - p_{V_2}(c)\gamma_{\mathcal{V}}(c) = 0,
\end{aligned}$$

because V_2 is distributed on a finite set. Hence, by the definition of $\chi(P_{VX})$ in (3.4), $\mathbb{E}\tilde{\chi} = 0$. Because $r \in L_2(P_{V_1X})$, the assumptions (3.5), (3.6), (3.7) imply that $\tilde{\chi}$ is in the tangent set $L_2^0(P_{VX})$ of the nonparametric model.

Consider now a regular submodel $t \mapsto p_{VX,t}$ for $t \in \mathbb{R}$ in the neighbourhood of zero with $p_{VX,0} = p_{VX}$ and the property that $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} p_{VX,t} = sp_{VX}$ for the score function s running through the tangent set $L_2^0(P_{VX})$ (see e.g. Bolthausen et al. (2002, Part III, Example 1.12) for such a construction). For the assertion (3.9), it suffices to show that $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\chi(P_{VX,t}) = \mathbb{E}\tilde{\chi}s$. Assuming that differentiation and expectation commutes, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\chi(P_{VX,t}) &= \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\int_{\mathfrak{V}\times\mathfrak{X}} f(v, x, \mu_{\mathcal{X},t}, \gamma_{\mathcal{V},t}(c))p_{VX,t}(v, x)\,\mathrm{d}\nu_X(x)\,\mathrm{d}\nu_V(v) \\
&= \int_{\mathfrak{V}\times\mathfrak{X}} \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} f(v, x, \mu_{\mathcal{X},t}, \gamma_{\mathcal{V},t}(c))p_{VX}(v, x)\,\mathrm{d}\nu_X(x)\,\mathrm{d}\nu_V(v) \\
&\quad + \mathbb{E}f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))s(V, X).
\end{aligned}$$

Here,

$$\begin{aligned}
&\int_{\mathfrak{V}\times\mathfrak{X}} \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} f(v, x, \mu_{\mathcal{X},t}, \gamma_{\mathcal{V},t}(c))p_{VX}(v, x)\,\mathrm{d}\nu_X(x)\,\mathrm{d}\nu_V(v) \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\mathbb{E}f(V, X, \mu_{\mathcal{X},t}, \gamma_{\mathcal{V}}(c)) + \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\mathbb{E}f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V},t}(c)).
\end{aligned}$$

By (3.8), $\mathbb{E}f(V,X,\mu_{\mathcal{X},t},\gamma_{\mathcal{V}}(c)) = \mathbb{E}r(V_1,X)\mu_{\mathcal{X},t}(V_1,X)$. Taking derivatives, the properties of (conditional) expectation give

$$\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\mu_{\mathcal{X},t}(v_1,x) = \mathbb{E}\left[(m(V,X)-\mu_{\mathcal{X}}(v_1,x))s(V,X) \mid V_1=v_1, X=x\right], \tag{3.84}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\gamma_{\mathcal{V},t}(c) &= \mathbb{E}\left[(g(V,X)-\gamma_{\mathcal{V}}(c))s(V,X) \mid V_2=c\right] \\ &= \mathbb{E}\frac{\mathbb{1}_{V_2=c}}{p_{V_2}(c)}(g(V,X)-\gamma_{\mathcal{V}}(c))s(V,X),\end{aligned} \tag{3.85}$$

implying

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\mathbb{E}r(V_1,X)\mu_{\mathcal{X},t}(V_1,X) \\ &= \mathbb{E}\left[r(V_1,X)\mathbb{E}\left[(m(V,X)-\mu_{\mathcal{X}}(V_1,X))\,s(V,X) \mid V_1,X\right]\right] \\ &= \mathbb{E}r(V_1,X)\left(m(V,X)-\mu_{\mathcal{X}}(V_1,X)\right)s(V,X), \\ \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\mathbb{E}f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V},t}(c)) &= \mathbb{E}\left[\partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\gamma_{\mathcal{V},t}(c)\right] \\ &= \mathbb{E}\left[\partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\mathbb{E}\left[\frac{\mathbb{1}_{V_2=c}}{p_{V_2}(c)}(g(V,X)-\gamma_{\mathcal{V}}(c))s(V,X)\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\right]\frac{\mathbb{1}_{V_2=c}}{p_{V_2}(c)}(g(V,X)-\gamma_{\mathcal{V}}(c))s(V,X)\right].\end{aligned}$$

Hence,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\chi(P_{VX,t}) &= \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\mathbb{E}f(V,X,\mu_{\mathcal{X},t},\gamma_{\mathcal{V}}(c)) + \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\mathbb{E}f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V},t}(c)) \\ &\quad + \mathbb{E}\left[f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))s(V,X)\right] \\ &= \mathbb{E}r(V_1,X)\left(m(V,X)-\mu_{\mathcal{X}}(V_1,X)\right)s(V,X) \\ &\quad + \mathbb{E}\left[\mathbb{E}\left[\partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\right]\frac{\mathbb{1}_{V_2=c}}{p_{V_2}(c)}(g(V,X)-\gamma_{\mathcal{V}}(c))s(V,X)\right] \\ &\quad + \mathbb{E}\left[f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))s(V,X)\right] \\ &= \mathbb{E}\tilde{\chi}s,\end{aligned}$$

as was to be shown, where the last equality follows from $\mathbb{E}s = 0$. ■

Proof of Lemma 3.1. Throughout, we apply that if a P_{VX}-random function $\hat{q} \in L_2(P_{VX})$ is independent of the random sample generating the process $\mathbb{P}_n$, then

$$\int (\hat{q}(v,x) - q(v,x))^2 \, \mathrm{d}P_{VX}(v,x) = o_{P_{VX}}(1) \text{ implies } \sqrt{n}(\mathbb{P}_n - P_{VX})(\hat{q} - q) = o_{P_{VX}}(1); \tag{3.86}$$

see e.g. Kennedy (2023, Lemma 1).

By the definitions (3.9) and (3.11),

$$\begin{aligned}
\hat{\tilde{\chi}}(v,x) - \tilde{\chi}(v,x) &= T_1(v,x) + T_2(v,x) + T_3(v,x) + T_4, \qquad (3.87)\\
T_1(v,x) &:= \hat{r}(v_1,x)(m(v,x) - \hat{\mu}_{\mathcal{X}}(v_1,x)) \\
&\quad - r(v_1,x)(m(v,x) - \mu_{\mathcal{X}}(v_1,x)), \\
T_2(v,x) &:= \frac{\mathbb{1}_{v_2=c}}{\hat{p}_{V_2}(c)}(g(v,x) - \hat{\gamma}_{\mathcal{V}}(c))\hat{e} - \frac{\mathbb{1}_{v_2=c}}{p_{V_2}(c)}(g(v,x) - \gamma_{\mathcal{V}}(c))e, \\
T_3(v,x) &:= f(v,x,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - f(v,x,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)), \\
T_4 &:= -\chi(\hat{P}_{VX}) + \chi(P_{VX}).
\end{aligned}$$

As T_4 is constant, not depending on (v,x), $(\mathbb{P}_n - P_{VX})T_4 = 0$. It remains to show $(\mathbb{P}_n - P_{VX})T_j = o_{P_{VX}}\left(n^{-1/2}\right)$ for $j = 1,2,3$ by the linearity of the process $\mathbb{P}_n - P_{VX}$.

Term T_1. Suppressing the arguments, write

$$\begin{aligned}
T_1 = \hat{r}(m - \hat{\mu}_{\mathcal{X}}) - r(m - \mu_{\mathcal{X}}) &= (\hat{r} - r + r)(m - \hat{\mu}_{\mathcal{X}}) - r(m - \mu_{\mathcal{X}}) \qquad (3.88)\\
&= (\hat{r} - r)(m - \hat{\mu}_{\mathcal{X}}) + r(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}). \qquad (3.89)
\end{aligned}$$

By Assumption 3.1, $\|m - \hat{\mu}_{\mathcal{X}}\|_\infty = O_{P_{VX}}(1)$; either directly by (3.23), or by (3.21) and (3.22), noting that $\|m - \hat{\mu}_{\mathcal{X}}\|_\infty \leq \|m - \mu_{\mathcal{X}}\|_\infty + \|\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}\|_\infty = O_{P_{VX}}(1) + o_{P_{VX}}(1) = O_{P_{VX}}(1)$. Then the $L_2(P_{V_1X})$-convergence (3.18) of r implies that $(\mathbb{P}_n - P_{VX})((\hat{r} - r)(m - \hat{\mu}_{\mathcal{X}})) = o_{P_{VX}}\left(n^{-1/2}\right)$ by (3.86) as

$$\begin{aligned}
&P_{VX}((\hat{r} - r)^2(m - \hat{\mu}_{\mathcal{X}})^2) \\
&\leq \|m - \hat{\mu}_{\mathcal{X}}\|_\infty^2 P_{VX}(\hat{r} - r)^2 = O_{P_{VX}}(1)\, o_{P_{VX}}(1) = o_{P_{VX}}(1)
\end{aligned}$$

since $\left\||q|^2\right\|_\infty = \|q\|_\infty^2$.

By Assumption 3.1, either (3.22), or (3.24) and (3.25) hold. In the former case,

$$P_{VX}(r^2(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}})^2) \leq \|\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}\|_\infty^2 P_{VX} r^2 = o_{P_{VX}}(1),$$

by (3.86) because $r \in L_2(P_{V_1X})$. In the latter case, letting

$$B \coloneqq \left\{(V_1, X) \in \mathfrak{V}_1 \times \mathfrak{X} : |r(V_1, X)| > \bar{R}\right\}$$

with complement B^C, we have

$$\begin{aligned} P_{VX}(r^2(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}})^2) &= \int_{B^C} r^2(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}})^2 \, \mathrm{d}P_{VX} + \int_B r^2(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}})^2 \, \mathrm{d}P_{VX} \\ &\leq \bar{R}^2 \left\|\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}}\right\|^2_{L_2(P_{V_1X})} + 0 = o_{P_{VX}}(1), \end{aligned}$$

since by (3.25), $P_{VX}(B) = 0$ and $\hat{\mu}_{\mathcal{X}}$ is $L_2(P_{V_1X})$-convergent by (3.24). Thus, $(\mathbb{P}_n - P_{VX})(r(\mu_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}})) = o_{P_{VX}}\left(n^{-1/2}\right)$ by (3.86). Conclude that $(\mathbb{P}_n - P_{VX})T_1 = o_{P_{VX}}\left(n^{-1/2}\right)$.

Term T_2. By the mean-value theorem there exists $(\tilde{\gamma}_{\mathcal{V}}(c), \tilde{p}_{V_2}(c), \tilde{e})$ between

$$(\gamma_{\mathcal{V}}(c), p_{V_2}(c), e) \text{ and } (\hat{\gamma}_{\mathcal{V}}(c), \hat{p}_{V_2}(c), \hat{e})$$

such that

$$\begin{aligned} T_2(v, x) = &\frac{\mathbb{1}_{v_2=c}}{\hat{p}_{V_2}(c)}(g(v, x) - \hat{\gamma}_{\mathcal{V}}(c))\hat{e} - \frac{\mathbb{1}_{v_2=c}}{p_{V_2}(c)}(g(v, x) - \gamma_{\mathcal{V}}(c))e \\ = &-\frac{\mathbb{1}_{v_2=c}}{\tilde{p}_{V_2}(c)}\tilde{e}(\hat{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c)) \\ &-\frac{\mathbb{1}_{v_2=c}}{\tilde{p}_{V_2}(c)^2}(g(v, x) - \tilde{\gamma}_{\mathcal{V}}(c))\tilde{e}(\hat{p}_{V_2}(c) - p_{V_2}(c)) \\ &+\frac{\mathbb{1}_{v_2=c}}{\tilde{p}_{V_2}(c)}(g(v, x) - \tilde{\gamma}_{\mathcal{V}}(c))(\hat{e} - e). \end{aligned}$$

By the standard central limit theorem, $\sqrt{n}(\mathbb{P}_n - P_{VX})\mathbb{1}_{V_2=c} = O_{P_{VX}}(1)$. By the linearity of the process $\mathbb{P}_n - P_{VX}$,

$$\begin{aligned} &\sqrt{n}(\mathbb{P}_n - P_{VX})\left[\mathbb{1}_{V_2=c}g(V, X) - \mathbb{1}_{V_2=c}\tilde{\gamma}_{\mathcal{V}}(c)\right] \\ = &\sqrt{n}(\mathbb{P}_n - P_{VX})\mathbb{1}_{V_2=c}g(V, X) + \tilde{\gamma}_{\mathcal{V}}(c)\sqrt{n}(\mathbb{P}_n - P_{VX})\mathbb{1}_{V_2=c} \\ = &(1 + \tilde{\gamma}_{\mathcal{V}}(c))O_{P_{VX}}(1) = O_{P_{VX}}(1) \end{aligned}$$

again by the standard central limit theorem and (3.19). Suppose that $\hat{e} - e = o_{P_{VX}}(1)$, which we show later. Then by (3.19) and (3.20), $(\mathbb{P}_n - P_{VX})T_2 = o_{P_{VX}}\left(n^{-1/2}\right)$.

Term T_3. Recall that $T_3(v, x) = f(v, x, \hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c)) - f(v, x, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))$. The continuity (C.S), together with the consistency of $\hat{\gamma}_{\mathcal{V}}$ and $\hat{\mu}_{\mathcal{X}}$ ((3.19) and (3.22)

or (3.24)) imply that $\int T_3(v,x)^2 \,\mathrm{d}P_{VX}(v,x) = o_{P_{VX}}(1)$ by the continuous mapping theorem. Conclude by (3.86) that $(\mathbb{P}_n - P_{VX})T_3 = o_{P_{VX}}\left(n^{-1/2}\right)$.

Consistency of $\hat{e}$. By the definition of e and $\hat{e}$,

$$\begin{aligned}
\hat{e} - e =& \mathbb{P}_n'' \partial_\gamma f(V, X, \hat{\mu}_\mathcal{X}, \hat{\gamma}_\mathcal{V}(c)) - P_{VX} \partial_\gamma f(V, X, \mu_\mathcal{X}, \gamma_\mathcal{V}(c)) \\
=& \mathbb{P}_n'' \partial_\gamma f(V, X, \hat{\mu}_\mathcal{X}, \hat{\gamma}_\mathcal{V}(c)) - P_{VX} \partial_\gamma f(V, X, \hat{\mu}_\mathcal{X}, \hat{\gamma}_\mathcal{V}(c)) \\
&+ P_{VX} \partial_\gamma f(V, X, \hat{\mu}_\mathcal{X}, \hat{\gamma}_\mathcal{V}(c)) - P_{VX} \partial_\gamma f(V, X, \mu_\mathcal{X}, \gamma_\mathcal{V}(c)) \\
=& (\mathbb{P}_n'' - P_{VX}) \partial_\gamma f(V, X, \hat{\mu}_\mathcal{X}, \hat{\gamma}_\mathcal{V}(c)) \\
&+ P_{VX} \left[\partial_\gamma f(V, X, \hat{\mu}_\mathcal{X}, \hat{\gamma}_\mathcal{V}(c)) - \partial_\gamma f(V, X, \mu_\mathcal{X}, \gamma_\mathcal{V}(c))\right] \\
=& (\mathbb{P}_n'' - P_{VX}) \partial_\gamma f(V, X, \mu_\mathcal{X}, \gamma_\mathcal{V}(c)) \\
&+ (\mathbb{P}_n'' - P_{VX}) \left[\partial_\gamma f(V, X, \hat{\mu}_\mathcal{X}, \hat{\gamma}_\mathcal{V}(c)) - \partial_\gamma f(V, X, \mu_\mathcal{X}, \gamma_\mathcal{V}(c))\right] \\
&+ P_{VX} \left[\partial_\gamma f(V, X, \hat{\mu}_\mathcal{X}, \hat{\gamma}_\mathcal{V}(c)) - \partial_\gamma f(V, X, \mu_\mathcal{X}, \gamma_\mathcal{V}(c))\right].
\end{aligned}$$

Here, the first term is $O_{P_{VX}}\left(n^{-1/2}\right) = o_{P_{VX}}(1)$ by the standard central limit theorem, and the second and third term are $o_{P_{VX}}(1)$ by the continuity (C.DS) along the same arguments concerning T_3 above. ■

Proof of Theorem 3.1. First,

$$P_{VX}\left[-r(V_1, X)h(V_1, X) + f(V, X, h, \gamma_\mathcal{V}(c))\right] = 0, \tag{3.90}$$

$$P_{VX}\left[(m(V, X) - \mu_\mathcal{X}(V_1, X))h(V_1, X)\right] = 0 \tag{3.91}$$

for all $h \in L_2(P_{V_1X})$, where the first equality is by (3.8) and the second is by the definition of $\mu_\mathcal{X}$ and the tower property of expectation. Then, because $\tilde{\chi}$ is an influence function satisfying $P_{VX}\tilde{\chi} = 0$, and $\tilde{\chi}, \chi_0$ — not depending on

(v, x) — are constants with respect to P_{VX}-integration,

$$\begin{aligned}
\chi' - \chi_0 + P_{VX}\tilde{\chi}' &= P_{VX}\left[\chi' + \tilde{\chi}'\right] - P_{VX}\left[\chi_0 + \tilde{\chi}\right] \\
&= P_{VX}\left\{r'(V_1, X)(m(V, X) - \mu'_{\mathcal{X}}(V_1, X)) + \frac{\mathbb{1}_{V_2=c}}{p'_{V_2}(c)}(g(V, X) - \gamma'_{\mathcal{V}}(c))e'' \right. \\
&\qquad \left. + f(V, X, \mu'_{\mathcal{X}}, \gamma'_{\mathcal{V}}(c))\right\} \\
&\quad - P_{VX}\left\{r(V_1, X)(m(V, X) - \mu_{\mathcal{X}}(V_1, X)) \right. \\
&\quad \left. + \frac{\mathbb{1}_{V_2=c}}{p_{V_2}(c)}(g(V, X) - \gamma_{\mathcal{V}}(c))P_{VX}\partial_\gamma f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) + f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))\right\} \\
&\quad - P_{VX}\left\{-r(V_1, X)(\mu'_{\mathcal{X}}(V_1, X) - \mu_{\mathcal{X}}(V_1, X)) \right. \\
&\quad \left. + f(V, X, \mu'_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) - f(V, X, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))\right\} \\
&\quad - P_{VX}\left\{[m(V, X) - \mu_{\mathcal{X}}(V_1, X)]\left[r'(V_1, X) - r(V_1, X)\right]\right\},
\end{aligned}$$

where the last two expectations are zero: we applied (3.90) and (3.91) choosing h to be $\mu'_{\mathcal{X}} - \mu_{\mathcal{X}}$ and $r' - r$, respectively, and using the linearity of f in (C.L). Collecting terms, noting that $P_{VX}\frac{\mathbb{1}_{V_2=c}}{p_{V_2}(c)}(g(V, X) - \gamma_{\mathcal{V}}(c)) = 0$ by the tower property and the definition of $\gamma_{\mathcal{V}}$,

$$\begin{aligned}
\chi' - \chi_0 + P_{VX}\tilde{\chi}' &= -P_{VX}(r - r')(\mu_{\mathcal{X}} - \mu'_{\mathcal{X}}) \\
&\quad + P_{VX}\left\{\frac{\mathbb{1}_{V_2=c}}{p'_{V_2}(c)}(g(V, X) - \gamma'_{\mathcal{V}}(c))e'' \right. \\
&\quad \left. + f(V, X, \mu'_{\mathcal{X}}, \gamma'_{\mathcal{V}}(c)) - f(V, X, \mu'_{\mathcal{X}}, \gamma_{\mathcal{V}}(c))\right\}.
\end{aligned}$$

Next, a Taylor-approximation of $\gamma \mapsto f(V, X, \mu'_{\mathcal{X}}, \gamma)$ by (C.D) with a mean-value representation of the the remainder term gives

$$\begin{aligned}
f(V, X, \mu'_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) = f(V, X, \mu'_{\mathcal{X}}, \gamma'_{\mathcal{V}}(c)) &+ \partial_\gamma f(V, X, \mu'_{\mathcal{X}}, \gamma'_{\mathcal{V}}(c))\left[\gamma_{\mathcal{V}}(c) - \gamma'_{\mathcal{V}}(c)\right] \\
&+ \frac{1}{2}\partial^2_\gamma f(V, X, \mu'_{\mathcal{X}}, \widetilde{\gamma_{\mathcal{V}}(c)})\left[\gamma_{\mathcal{V}}(c) - \gamma'_{\mathcal{V}}(c)\right]^2.
\end{aligned}$$

As $P_{VX}\mathbb{1}_{V_2=c}(g(V,X)-\gamma'_{\mathcal{V}}(c))=p_{V_2}(c)(\gamma_{\mathcal{V}}(c)-\gamma'_{\mathcal{V}}(c))$,

$$\begin{aligned}P_{VX}\left\{\frac{\mathbb{1}_{V_2=c}}{p'_{V_2}(c)}(g(V,X)-\gamma'_{\mathcal{V}}(c))e''+f(V,X,\mu'_{\mathcal{X}},\gamma'_{\mathcal{V}}(c))-f(V,X,\mu'_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\right\}\\ =\frac{p_{V_2}(c)}{p'_{V_2}(c)}(\gamma_{\mathcal{V}}(c)-\gamma'_{\mathcal{V}}(c))e''-P_{VX}\partial_\gamma f(V,X,\mu'_{\mathcal{X}},\gamma'_{\mathcal{V}}(c))\left[\gamma_{\mathcal{V}}(c)-\gamma'_{\mathcal{V}}(c)\right]\\ -\frac{1}{2}P_{VX}\partial^2_\gamma f(V,X,\mu'_{\mathcal{X}},\widetilde{\gamma_{\mathcal{V}}(c)})\left[\gamma_{\mathcal{V}}(c)-\gamma'_{\mathcal{V}}(c)\right]^2\end{aligned}$$

which yields the assertion by the definition of e'. ■

Proof of Corollary 3.1. Follows from Lemma 3.1 and Theorem 3.1, noting that, for e' in (3.29),

$$\hat{e}-e'=(\mathbb{P}''_n-P_{VX})\partial_\gamma f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c))$$

is $o_{P_{VX}}(1)$ by the consistency proof of $\hat{e}$ in Lemma 3.1. ■

3.A.1. Additional Results

Proposition 3.4 (Asymptotic Efficiency of the Plug-in Estimator for Discrete Covariates)**.** *Suppose that $\mathfrak{X}$ and $\mathfrak{V}_1$ are finite, and define*

$$\begin{aligned}\chi(\hat{P}_{VX})&:=\frac{1}{n}\sum_{i\in[n]}f(V_i,X_i,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)),\\ \hat{\mu}_{\mathcal{X}}(v_1,x)&:=\frac{1}{N_{v_1x}}\sum_{i\in[n]}\mathbb{1}_{V_{1i}=v_1,X_i=x}m(V_i,X_i),\quad N_{v_1x}:=\sum_{i\in[n]}\mathbb{1}_{V_{1i}=v_1,X_i=x},\\ \hat{\gamma}_{\mathcal{V}}(c)&:=\frac{1}{N_c}\sum_{i\in[n]}\mathbb{1}_{V_{2i}=c}g(V_i,X_i),\quad N_c:=\sum_{i\in[n]}\mathbb{1}_{V_{2i}=c}.\end{aligned}$$

Assume that $\|\mu_{\mathcal{X}}\|_\infty<\infty$ and that there exists an $\hat{r}:\mathfrak{V}_1\times\mathfrak{X}\to\mathbb{R}$ such that $\|\hat{r}-r\|_\infty=o_{P_{VX}}(1)$, where r is the Riesz representer of $\mu\mapsto\mathbb{E}f(V,X,\mu,\gamma_{\mathcal{V}}(c))=\chi(P_{VX})$ as before. Further assume that for every fixed $\mu\in L_2(P_{V_1X})$ and for every $\tilde{\gamma}_{\mathcal{V}}(c)$ between $\gamma_{\mathcal{V}}(c)$ and $\hat{\gamma}_{\mathcal{V}}(c)$, the stochastic bound conditions

$$(\mathbb{P}_n-P_{VX})\partial_\gamma f(V,X,\mu,\tilde{\gamma}_{\mathcal{V}}(c))=O_{P_{VX}}(1),\tag{3.92}$$

$$(\mathbb{P}_n-P_{VX})\partial^2_\gamma f(V,X,\mu,\tilde{\gamma}_{\mathcal{V}}(c))=O_{P_{VX}}(1),\tag{3.93}$$

$$P_{VX}\partial_\gamma f(V,X,\mu_{\mathcal{X}},\tilde{\gamma}_{\mathcal{V}}(c))=O_{P_{VX}}(1),\tag{3.94}$$

$$P_{VX}\partial^2_\gamma f(V,X,\mu_{\mathcal{X}},\tilde{\gamma}_{\mathcal{V}}(c))=O_{P_{VX}}(1),\tag{3.95}$$

$$P_{VX}\partial^2_\gamma f(V,X,\hat{\mu}_{\mathcal{X}},\tilde{\gamma}_{\mathcal{V}}(c))=O_{P_{VX}}(1)\tag{3.96}$$

hold. Then $\sqrt{n}(\chi(\hat{P}_{VX}) - \chi(P_{VX})) \overset{P_{VX}}{\rightsquigarrow} \mathcal{N}(0, P_{VX}\tilde{\chi}^2)$ *as* $n \to \infty$.

Proof of Proposition 3.4. We start with an auxiliary result. Define

$$\begin{aligned}\hat{\tilde{\chi}}(v,x) &\coloneqq \hat{r}(v_1,x)(m(v,x) - \hat{\mu}_{\mathcal{X}}(v_1,x)) + \frac{\mathbb{1}_{v_2=c}}{\hat{p}_{V_2}(c)}(g(v,x) - \hat{\gamma}_{\mathcal{V}}(c))\hat{e} \\ &\quad + f(v,x,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - \chi(\hat{P}_{VX}) \\ \hat{p}_{V_2}(c) &\coloneqq N_c/n.\end{aligned}$$

First we show that

$$\begin{aligned}\mathbb{P}_n\hat{\tilde{\chi}} = \mathbb{P}_n\left[\hat{r}(V_1,X)(m(V,X) - \hat{\mu}_{\mathcal{X}}(V_1,X)\right] \\ + \frac{\hat{e}}{\hat{p}_{V_2}(c)}\mathbb{P}_n\left[\mathbb{1}_{V_2=c}g(V,X) - \hat{\gamma}_{\mathcal{V}}(c)\mathbb{1}_{V_2=c}\right] + \mathbb{P}_n\left[f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - \chi(\hat{P}_{VX}))\right]\end{aligned}$$

is zero. We apply an 'empirical tower property' to the first term to get

$$\begin{aligned}\mathbb{P}_n\hat{r}(V_1,X)(m(V,X) - \hat{\mu}_{\mathcal{X}}(V_1,X)) &= \frac{1}{n}\sum_{i\in[n]}\hat{r}(V_{1i},X_i)(m(V_i,X_i) - \hat{\mu}_{\mathcal{X}}(V_{1i},X_i)) \\ &= \frac{1}{n}\sum_{(v_1,x)\in\mathfrak{V}_1\times\mathfrak{X}}\sum_{i:(V_{1i},X_i)=(v_1,x)}\hat{r}(V_{1i},X_i)(m(V_i,X_i) - \hat{\mu}_{\mathcal{X}}(V_{1i},X_i)) \\ &= \frac{1}{n}\sum_{(v_1,x)\in\mathfrak{V}_1\times\mathfrak{X}}\left\{\hat{r}(v_1,x)\right. \\ &\quad \left.\times\left[\left(\sum_{i:(V_{1i},X_i)=(v_1,x)} m(V_i,X_i)\right) - \left(\sum_{i:(V_{1i},X_i)=(v_1,x)}\hat{\mu}_{\mathcal{X}}(v_1,x)\right)\right]\right\} \\ &= \frac{1}{n}\sum_{(v_1,x)\in\mathfrak{V}_1\times\mathfrak{X}}\hat{r}(v_1,x)\left[N_{v_1x}\hat{\mu}_{\mathcal{X}}((v_1,x)) - N_{v_1x}\hat{\mu}_{\mathcal{X}}((v_1,x))\right] = 0.\end{aligned}$$

The second term satisfies

$$\begin{aligned}\mathbb{P}_n\left[\mathbb{1}_{V_2=c}g(V,X) - \hat{\gamma}_{\mathcal{V}}(c)\mathbb{1}_{V_2=c}\right] &= \mathbb{P}_n\mathbb{1}_{V_2=c}g(V,X) - \hat{\gamma}_{\mathcal{V}}(c)\mathbb{P}_n\mathbb{1}_{V_2=c} \\ &= N_c\hat{\gamma}_{\mathcal{V}}(c) - \hat{\gamma}_{\mathcal{V}}(c)N_c = 0\end{aligned}$$

by the definition of $\hat{\gamma}_{\mathcal{V}}(c), N_c$. The last term satisfies

$$\mathbb{P}_n\left[f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - \chi(\hat{P}_{VX}))\right] = \mathbb{P}_n f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - \chi(\hat{P}_{VX}) = 0$$

by the definition of $\chi(\hat{P}_{VX})$ as $\chi(\hat{P}_{VX})$ is constant with respect to the empirical measure $\mathbb{P}_n$.

Now we show asymptotic efficiency. Using that $\mathbb{P}_n\hat{\tilde{\chi}} = 0$, we write

$$\sqrt{n}(\chi(\hat{P}_{VX}) - \chi(P_{VX})) = \sqrt{n}\mathbb{P}_n\tilde{\chi} + \sqrt{n}(\mathbb{P}_n - P_{VX})(\hat{\tilde{\chi}} - \tilde{\chi}) + \sqrt{n}R_n,$$
$$R_n = \chi(\hat{P}_{VX}) - \chi(P_{VX}) + P_{VX}\hat{\tilde{\chi}},$$

for $\tilde{\chi}$ in (3.9). The first term satisfies $\sqrt{n}\mathbb{P}_n\tilde{\chi} \overset{P_{VX}}{\rightsquigarrow} \mathcal{N}(0, P_{VX}\tilde{\chi}^2)$. Because the arguments in Theorem 3.1 also apply when $(\hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}})$ and $\chi(\hat{P}_{VX})$ are computed from the same sample $\mathcal{S}$, we have $\sqrt{n}R_n = o_{P_{VX}}(1)$. This follows from (3.27) of Theorem 3.1, because $\hat{\mu}_{\mathcal{X}}$, $\hat{\gamma}_{\mathcal{V}}(c)$ and $\hat{p}_{V_2}(c)$ are asymptotically normal, $\hat{r}$ is consistent , $P_{VX}\partial^2_\gamma f(V, X, \hat{\mu}_{\mathcal{X}}, \tilde{\gamma}_{\mathcal{V}}(c) = O_{P_{VX}}(1)$ by (3.96), and, as we show at the end of this proof, $\hat{e} - e' = o_{P_{VX}}(1)$ for $e' = P_{VX}\partial_\gamma f(V, X, \hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c))$.

It remains to show $\sqrt{n}(\mathbb{P}_n - P_{VX})(\hat{\tilde{\chi}} - \tilde{\chi}) = o_{P_{VX}}(1)$. Because Lemma 3.1 assumes that $(\hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}})$ are computed from a sample independent from what generates $\mathbb{P}_n$, we need to adapt the arguments therein. As in (3.87), write

$$\begin{aligned}
\hat{\tilde{\chi}}(v,x) - \tilde{\chi}(v,x) &= T_1(v,x) + T_2(v,x) + T_3(v,x) + T_4, \qquad (3.97)\\
T_1(v,x) &:= \hat{r}(v_1,x)(m(v,x) - \hat{\mu}_{\mathcal{X}}(v_1,x))\\
&\quad - r(v_1,x)(m(v,x) - \mu_{\mathcal{X}}(v_1,x)),\\
T_2(v,x) &:= \frac{\mathbb{1}_{v_2=c}}{\hat{p}_{V_2}(c)}(g(v,x) - \hat{\gamma}_{\mathcal{V}}(c))\hat{e} - \frac{\mathbb{1}_{v_2=c}}{p_{V_2}(c)}(g(v,x) - \gamma_{\mathcal{V}}(c))e,\\
T_3(v,x) &:= f(v,x,\hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c)) - f(v,x,\mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)),\\
T_4 &:= -\chi(\hat{P}_{VX}) + \chi(P_{VX}).
\end{aligned}$$

Again, T_4 being a constant under $\mathbb{P}_n - P_{VX}$, $\sqrt{n}(\mathbb{P}_n - P_{VX})T_4 = 0$, so we need show $(\mathbb{P}_n - P_{VX})T_j = o_{P_{VX}}\left(n^{-1/2}\right)$ for $j = 1, 2, 3$.

Term T_1. Write

$$\begin{aligned}
\sqrt{n}(\mathbb{P}_n - P_{VX})T_1(V,X) &= \sqrt{n}(\mathbb{P}_n - P_{VX})\left[m(V,X)(\hat{r}(V_1,X) - r(V_1,X))\right]\\
&\quad + \sqrt{n}(\mathbb{P}_n - P_{VX})\left[\hat{\mu}_{\mathcal{X}}(V_1,X)(\hat{r}(V_1,X) - r(V_1,X))\right] \qquad (3.98)\\
&\quad + \sqrt{n}(\mathbb{P}_n - P_{VX})\left[r(V_1,X)(\mu_{\mathcal{X}}(V_1,X) - \hat{\mu}_{\mathcal{X}}(V_1,X))\right]. \qquad (3.99)
\end{aligned}$$

Because $\mathfrak{V}_1 \times \mathfrak{X}$ is finite, we can write

$$r(v_1, x) = \sum_{(\bar{v}_1, \bar{x}) \in \mathfrak{V}_1 \times \mathfrak{X}} \varrho_{\bar{v}_1, \bar{x}} \mathbb{1}_{(\bar{v}_1, \bar{x})}(v_1, x), \quad \varrho_{\bar{v}_1, \bar{x}} \coloneqq r(\bar{v}_1, \bar{x}),$$
$$\hat{r}(v_1, x) = \sum_{(\bar{v}_1, \bar{x}) \in \mathfrak{V}_1 \times \mathfrak{X}} \hat{\varrho}_{\bar{v}_1, \bar{x}} \mathbb{1}_{(\bar{v}_1, \bar{x})}(v_1, x), \quad \hat{\varrho}_{\bar{v}_1, \bar{x}} \coloneqq \hat{r}(\bar{v}_1, \bar{x}).$$

But then

$$\begin{aligned}
&\sqrt{n}(\mathbb{P}_n - P_{VX}) \left[m(V, X)(\hat{r}(V_1, X) - r(V_1, X)) \right] \\
&= \sqrt{n}(\mathbb{P}_n - P_{VX}) \left[m(V, X) \left(\sum_{(\bar{v}_1, \bar{x}) \in \mathfrak{V}_1 \times \mathfrak{X}} (\hat{\varrho}_{\bar{v}_1, \bar{x}} - \varrho_{\bar{v}_1, \bar{x}}) \mathbb{1}_{(\bar{v}_1, \bar{x})}(V_1, X) \right) \right] \\
&= \sum_{(\bar{v}_1, \bar{x}) \in \mathfrak{V}_1 \times \mathfrak{X}} (\hat{\varrho}_{\bar{v}_1, \bar{x}} - \varrho_{\bar{v}_1, \bar{x}}) \sqrt{n}(\mathbb{P}_n - P_{VX}) \left[m(V, X) \mathbb{1}_{(\bar{v}_1, \bar{x})}(V_1, X) \right],
\end{aligned}$$

which is $o_{P_{VX}}(1)$, because $\hat{r}$ is consistent,

$$\sqrt{n}(\mathbb{P}_n - P_{VX}) \left[m(V, X) \mathbb{1}_{(\bar{v}_1, \bar{x})}(V_1, X) \right] = O_{P_{VX}}(1)$$

by the standard central limit theorem, and $|\mathfrak{V}_1 \times \mathfrak{X}|$ is finite. The terms (3.98), (3.99) can be handled similarly because $\|\hat{\mu}_{\mathcal{X}}\|_\infty \leq \|\hat{\mu}_{\mathcal{X}} - \mu_{\mathcal{X}}\|_\infty + \|\mu_{\mathcal{X}}\|_\infty = o_{P_{VX}}(1) + O(1) = O_{P_{VX}}(1)$ by assumption. Thus $\sqrt{n}(\mathbb{P}_n - P_{VX})T_1 = o_{P_{VX}}(1)$.

Term T_2. Same arguments as in Lemma 3.1 apply, yielding $\sqrt{n}(\mathbb{P}_n - P_{VX})T_2 = o_{P_{VX}}(1)$, because $\hat{e} - e = o_{P_{VX}}(1)$ — which we show at the end of this proof —, and $\hat{\gamma}_{\mathcal{V}}(c), \hat{p}_{V_2}(c)$ are consistent.

Term T_3. Write

$$\begin{aligned}
T_3(v, x) &= f(v, x, \hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c)) - f(v, x, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)) \\
&= f(v, x, \hat{\mu}_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c)) - f(v, x, \mu_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c)) \\
&\quad + f(v, x, \mu_{\mathcal{X}}, \hat{\gamma}_{\mathcal{V}}(c)) - f(v, x, \mu_{\mathcal{X}}, \gamma_{\mathcal{V}}(c)).
\end{aligned} \tag{3.100}$$

As with r before, represent $\mu_{\mathcal{X}}(v_1, x) = \sum_{(\bar{v}_1, \bar{x}) \in \mathfrak{V}_1 \times \mathfrak{X}} \lambda_{\bar{v}_1, \bar{x}} \mathbb{1}_{(\bar{v}_1, \bar{x})}(v_1, x)$ for some parameter $\lambda \in \mathbb{R}^{|\mathfrak{V}_1 \times \mathfrak{X}|}$ with $\lambda_{\bar{v}_1, \bar{x}} \coloneqq \mu_{\mathcal{X}}(\bar{v}_1, \bar{x})$; similarly, write

$$\hat{\mu}_{\mathcal{X}}(v_1, x) = \sum_{(\bar{v}_1, \bar{x}) \in \mathfrak{V}_1 \times \mathfrak{X}} \hat{\lambda}_{\bar{v}_1, \bar{x}} \mathbb{1}_{(\bar{v}_1, \bar{x})}(v_1, x),$$
$$\hat{\lambda}_{\bar{v}_1, \bar{x}} \coloneqq \hat{\mu}_{\mathcal{X}}(\bar{v}_1, \bar{x}).$$

By the linearity (C.L) of f and the mean value theorem,

$$\begin{aligned}
&\sqrt{n}(\mathbb{P}_n - P_{VX})\left[f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\right] \\
&\qquad = \sqrt{n}(\mathbb{P}_n - P_{VX})\left[f(V,X,\hat{\mu}_{\mathcal{X}} - \mu_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c))\right] \\
&= \sqrt{n}(\mathbb{P}_n - P_{VX})\left[\sum_{(\bar{v}_1,\bar{x})\in\mathfrak{V}_1\times\mathfrak{X}} (\hat{\lambda}_{\bar{v}_1,\bar{x}} - \lambda_{\bar{v}_1,\bar{x}}) f(V,X,\mathbb{1}_{\bar{v}_1,\bar{x}},\hat{\gamma}_{\mathcal{V}}(c))\right] \\
&= \sum_{(\bar{v}_1,\bar{x})\in\mathfrak{V}_1\times\mathfrak{X}} (\hat{\lambda}_{\bar{v}_1,\bar{x}} - \lambda_{\bar{v}_1,\bar{x}})\sqrt{n}(\mathbb{P}_n - P_{VX})\left[f(V,X,\mathbb{1}_{\bar{v}_1,\bar{x}},\hat{\gamma}_{\mathcal{V}}(c))\right] \\
&= \sum_{(\bar{v}_1,\bar{x})\in\mathfrak{V}_1\times\mathfrak{X}} (\hat{\lambda}_{\bar{v}_1,\bar{x}} - \lambda_{\bar{v}_1,\bar{x}})\sqrt{n}(\mathbb{P}_n - P_{VX})\left[f(V,X,\mathbb{1}_{\bar{v}_1,\bar{x}},\gamma_{\mathcal{V}}(c))\right] \\
&\quad + \sum_{(\bar{v}_1,\bar{x})\in\mathfrak{V}_1\times\mathfrak{X}} (\hat{\lambda}_{\bar{v}_1,\bar{x}} - \lambda_{\bar{v}_1,\bar{x}})(\mathbb{P}_n - P_{VX})\left[\partial_\gamma f(V,X,\mathbb{1}_{\bar{v}_1,\bar{x}},\tilde{\gamma}_{\mathcal{V}}(c))\right] \\
&\qquad\qquad \times\sqrt{n}(\hat{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c))
\end{aligned}$$

for some $\tilde{\gamma}_{\mathcal{V}}(c)$ between $\gamma_{\mathcal{V}}(c)$ and $\hat{\gamma}_{\mathcal{V}}(c)$. But this is $o_{P_{VX}}(1)$ by the standard central limit theorem, and because $\sqrt{n}(\hat{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c)) = O_{P_{VX}}(1)$, $\hat{\mu}_{\mathcal{X}}$ being consistent and $(\mathbb{P}_n - P_{VX})\left[\partial_\gamma f(V,X,\mathbb{1}_{\bar{v}_1,\bar{x}},\tilde{\gamma}_{\mathcal{V}}(c))\right] = O_{P_{VX}}(1)$ by (3.92). For the term (3.100), apply the mean-value theorem twice to get

$$\begin{aligned}
&T_{3,1}(v,x) \coloneqq f(v,x,\mu_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - f(v,x,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) \\
&= (\hat{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c))\left\{\partial_\gamma f(v,x,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) + (\tilde{\gamma}_{\mathcal{V}}(c) - \gamma_{\mathcal{V}}(c))\partial_\gamma^2 f(v,x,\mu_{\mathcal{X}},\tilde{\gamma}'_{\mathcal{V}}(c))\right\}
\end{aligned}$$

for some $\tilde{\gamma}'_{\mathcal{V}}(c)$ between $\gamma_{\mathcal{V}}(c)$ and $\hat{\gamma}_{\mathcal{V}}(c)$. Thus $\sqrt{n}(\mathbb{P}_n - P_{VX})T_{3,1} = o_{P_{VX}}(1)$ by the standard central limit theorem, $\sqrt{n}$-consistency of $\hat{\gamma}_{\mathcal{V}}(c)$, and stochastic boundedness (3.93) of the second derivative. Hence, $\sqrt{n}(\mathbb{P}_n - P_{VX})T_3 = o_{P_{VX}}(1)$.

Consistency of $\hat{e}$. Write

$$\begin{aligned}
\hat{e} - e &= \mathbb{P}_n\partial_\gamma f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - P_{VX}\partial_\gamma f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) \\
&= (\mathbb{P}_n - P_{VX})\partial_\gamma f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) \qquad (3.101) \\
&\quad + P_{VX}\left[\partial_\gamma f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - \partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\right]. \qquad (3.102)
\end{aligned}$$

As $\mu \mapsto \partial_\gamma f(v,x,\mu,\bar{\gamma})$ is linear, (3.101) is $o_{P_{VX}}(1)$ along similar arguments as T_3 above; because we only need consistency, we only need the existence and

stochastic boundedness (3.93) of the second derivative; no need for higher order derivatives. For the term (3.101), write it as

$$P_{VX}\left[\partial_\gamma f(V,X,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - \partial_\gamma f(V,X,\mu_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c))\right]$$
$$+P_{VX}\left[\partial_\gamma f(V,X,\mu_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c)) - \partial_\gamma f(V,X,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\right].$$

By the linearity of $\mu \mapsto \partial_\gamma f(v,x,\mu,\hat{\gamma}_{\mathcal{V}}(c))$, the first term here is $o_{P_{VX}}(1)$, following a mean-value expansion, by the consistency of $\hat{\mu}_{\mathcal{X}}$ and the assumed boundedness (3.95) of $P_{VX}\partial_\gamma f(V,X,\mu_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c))$. The second term is $o_{P_{VX}}(1)$ by the same arguments under assumption (3.95) for $P_{VX}\partial^2_\gamma f(V,X,\mu_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c))$. Hence, $\hat{e}-e=o_{P_{VX}}(1)$. Finally, note that $\hat{e}-e'=o_{P_{VX}}(1)$ too as we claimed, because it is equal to (3.101). ■

Note that for particular forms of f, the conditions of Proposition 3.4 are easier to verify. Namely, if $f(v,x,\mu,\gamma)$ factors as

$$f(v,x,\mu,\gamma)=f_1(v,x,\mu)f_2(\gamma)$$

where $\frac{\partial^2 f_2}{\partial\gamma^2}$ exists and is continuous and $(v,x)\mapsto f_1(v,x,\mu)$ belongs to $L_2(P_{VX})$ for any $\mu\in L_2(P_{V_1X})$, then the boundedness conditions (3.92),(3.93),(3.94), (3.95) are easily met by the standard central limit theorem and the continuous mapping theorem. For instance, this includes the average treatment effect on the treated in Example 3.2.

3.B. Privacy

This section contains the proofs of results in Section 3.5 — Lemmas 3.2 to 3.4 and Proposition 3.2 — and in Section 3.6 — Lemma 3.5 and Corollary 3.2. In Section 3.B.1, auxiliary results Lemmas 3.8 and 3.9 are presented.

Proof of Lemma 3.2. Let Q be α-LTVP. As $|Q|=Q$, we have for all $B\in\mathscr{F}_{\mathfrak{Z}}$,

$$Q(B|\,x)=|Q(B|\,x)|=|Q(B|\,x)-Q(B|\,x')+Q(B|\,x')|$$
$$\leq |Q(B|\,x')|+|Q(B|\,x)-Q(B|\,x')|\leq Q(B|\,x')+\alpha\leq e^{\tilde{\alpha}}Q(B|\,x')+\alpha$$

for all $x,x'\in\mathfrak{X}$ and for any $\tilde{\alpha}\geq 0$. Hence, Q is $(\tilde{\alpha},\alpha)$-LDP.

Now let Q be an (α, δ)-LDP mechanism. Then for all $x, x' \in \mathfrak{X}$ and $B \in \mathscr{F}_3$,

$$\begin{aligned}
|Q(B|\,x) - Q(B|\,x')| &= \begin{cases} Q(B|\,x) - Q(B|\,x') & \text{if } Q(B|\,x) - Q(B|\,x') \geq 0 \\ Q(B|\,x') - Q(B|\,x) & \text{if } Q(B|\,x) - Q(B|\,x') < 0 \end{cases} \\
&\leq \begin{cases} e^\alpha Q(B|\,x') + \delta - Q(B|\,x') & \text{if } Q(B|\,x) - Q(B|\,x') \geq 0 \\ e^\alpha Q(B|\,x) + \delta - Q(B|\,x) & \text{if } Q(B|\,x) - Q(B|\,x') < 0 \end{cases} \\
&= \begin{cases} (e^\alpha - 1)Q(B|\,x') + \delta & \text{if } Q(B|\,x) - Q(B|\,x') \geq 0 \\ (e^\alpha - 1)Q(B|\,x) + \delta & \text{if } Q(B|\,x) - Q(B|\,x') < 0 \end{cases} \\
&\leq e^\alpha - 1 + \delta,
\end{aligned}$$

because Q maps to $[0, 1]$ and $e^\alpha - 1 \geq 0$ for all $\alpha \geq 0$. Hence, if $e^\alpha - 1 + \delta \leq 1$, Q is $(e^\alpha - 1 + \delta)$-LTVP. ■

Proof of Lemma 3.3. Assertion (i). Follows directly from $Z \mid (V, X) \sim Q(\cdot|\,X)$.

Assertion (ii). By definition, the adjoint $Q^*_{\mathcal{X}}$ of $Q_{\mathcal{X}}$ satisfies $P_{VX}[(Q_{\mathcal{X}}k)h] = P_{VZ}[kQ^*_{\mathcal{X}}h]$. By the tower property of expectation we can verify that, for $Q^*_{\mathcal{X}}$ given in (ii),

$$\begin{aligned}
&P_{VX}[(Q_{\mathcal{X}}k)h] = \mathbb{E}\left[\mathbb{E}\left[k(V,Z) \mid V, X\right] h(V,X)\right] = \mathbb{E}\left[\mathbb{E}\left[k(V,Z)h(V,X) \mid V, X\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[k(V,Z)h(V,X) \mid V, Z\right]\right] = \mathbb{E}\left[k(V,Z)\mathbb{E}\left[h(V,X) \mid V, Z\right]\right] = P_{VZ}[kQ^*_{\mathcal{X}}h].
\end{aligned}$$

Assertion (iii). By the tower property of expectation,

$$P_{VZ}k = \mathbb{E}k(V,Z) = \mathbb{E}\mathbb{E}\left[k(V,Z) \mid V, X\right] = P_{VX}Q_{\mathcal{X}}k.$$

Assertion (iv). Under the discrete model for X,

$$\int_{\mathfrak{X}} k(v,z)Q(\mathrm{d}z|\,x) = \sum_{z \in \mathfrak{X}} k(v,z)Q(\{z\}|\,x),$$

from which the assertion directly follows.

Assertion (v). Under the mechanism Q in (3.41), we have

$$\begin{aligned}
(Q_{\mathcal{X}}k)(v,x) = \int_{\mathfrak{X}} k(v,z)Q(\mathrm{d}z|\,x) &= \int_{\mathfrak{X}} k(v,z)\left[\alpha\delta_x(\mathrm{d}z) + (1-\alpha)\bar{Q}(\mathrm{d}z)\right] \\
&= \alpha k(v,x) + (1-\alpha)\int_{\mathfrak{X}} k(v,z)\bar{Q}(\mathrm{d}z).
\end{aligned}$$

Next we show that $Q_{\mathcal{X}} : L_2(P_{VZ}) \to L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})$ is bounded. Take some $k \in L_2(P_{VZ})$, and let $(Q_{\mathcal{X}}k)(v,x) = \alpha k(v,x) + (1-\alpha)\int_{\mathfrak{X}} k(v,z)\bar{Q}(\mathrm{d}z) =: \phi_1(v,x) + \phi_2(v)$. By Minkowski's inequality,

$$\|Q_{\mathcal{X}}k\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})} \le \|\phi_1\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})} + \|\phi_2\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})},$$

where, by definition (3.44),

$$\|\phi_1\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})} = \|\phi_1\|_{L_2(P_{VX})} + \|\phi_1\|_{L_2(P_V\otimes\bar{Q})}$$

with $\|\phi_1\|_{L_2(P_{VX})} = \alpha\|k\|_{L_2(P_{VX})} \le \sqrt{\alpha}\|k\|_{L_2(P_{VZ})}$ by Lemma 3.8(iv), and

$$\|\phi_2\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})} = 2\|\phi_2\|_{L_2(P_V)}.$$

By Jensen's inequality,

$$\left(\int_{\mathfrak{X}} k(v,z)\bar{Q}(\mathrm{d}z)\right)^2 \le \int_{\mathfrak{X}} k(v,z)^2\bar{Q}(\mathrm{d}z),$$

thus

$$\begin{aligned}\|\phi_2\|_{L_2(P_V)} &\le (1-\alpha)\sqrt{\int_{\mathfrak{V}}\int_{\mathfrak{X}} k(v,z)^2\bar{Q}(\mathrm{d}z)P_V(\mathrm{d}v)} \\ &= (1-\alpha)\sqrt{\int_{\mathfrak{V}}\int_{\mathfrak{X}} k(v,z)^2\bar{q}(z)\nu_X(\mathrm{d}z)p_V(v)\nu_V(\mathrm{d}v)} \\ &\le \sqrt{(1-\alpha)}\sqrt{\int_{\mathfrak{V}}\int_{\mathfrak{X}} k(v,z)^2 p_{VZ}(v,z)\nu_X(\mathrm{d}z)\nu_V(\mathrm{d}v)} = \sqrt{(1-\alpha)}\|k\|_{L_2(P_{VZ})},\end{aligned}$$

where we used that by Lemma 3.8(iii),

$$\bar{q}(z)p_V(v) = \frac{1}{1-\alpha}p_{VZ}(v,z) - \frac{\alpha}{1-\alpha}p_{VX}(v,z) \le \frac{1}{1-\alpha}p_{VZ}(v,z)$$

for $\alpha \in (0,1)$. Finally, by the previous display, $\|\phi_1\|_{L_2(P_V\otimes\bar{Q})} \le \frac{1}{1-\alpha}\|\phi_1\|_{L_2(P_{VZ})}$, yielding

$$\begin{aligned}&\|Q_{\mathcal{X}}k\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})} \\ &\le \sqrt{\alpha}\|k\|_{L_2(P_{VZ})} + \frac{\alpha}{1-\alpha}\|k\|_{L_2(P_{VZ})} + 2\sqrt{(1-\alpha)}\|k\|_{L_2(P_{VZ})}.\end{aligned}$$

Assertion (vi). It suffices to show $Q_{\mathcal{X}}(Q_{\mathcal{X}}^{-1}h) = h$ for all $h \in L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})$. Under (3.41),

$$(Q_{\mathcal{X}}(Q_{\mathcal{X}}^{-1}h))(v,x) = \alpha(Q_{\mathcal{X}}^{-1}h)(v,x) + (1-\alpha)\int_{\mathfrak{X}}(Q_{\mathcal{X}}^{-1}h)(v,z)\bar{Q}(\mathrm{d}z).$$

Here, the first term is

$$\begin{aligned}\alpha(Q_{\mathcal{X}}^{-1}h)(v,x) &= \alpha\left\{\frac{1}{\alpha}h(v,x) - \frac{1-\alpha}{\alpha}\int_{\mathfrak{X}} h(v,t)\bar{Q}(\mathrm{d}t)\right\}\\ &= h(v,x) - (1-\alpha)\int_{\mathfrak{X}} h(v,t)\bar{Q}(\mathrm{d}t),\end{aligned}$$

while the second term is

$$\begin{aligned}&(1-\alpha)\int_{\mathfrak{X}}(Q_{\mathcal{X}}^{-1}h)(v,z)\bar{Q}(\mathrm{d}z)\\ &= (1-\alpha)\int_{\mathfrak{X}}\left\{\frac{1}{\alpha}h(v,z) - \frac{1-\alpha}{\alpha}\int_{\mathfrak{X}} h(v,t)\bar{Q}(\mathrm{d}t)\right\}\bar{Q}(\mathrm{d}z)\\ &= (1-\alpha)\left\{\frac{1}{\alpha}\int_{\mathfrak{X}} h(v,z)\bar{Q}(\mathrm{d}z) - \frac{1-\alpha}{\alpha}\int_{\mathfrak{X}} h(v,t)\bar{Q}(\mathrm{d}t)\right\}\\ &= (1-\alpha)\int_{\mathfrak{X}} h(v,t)\bar{Q}(\mathrm{d}t),\end{aligned}$$

where we used that $\bar{Q}$ is a probability measure with $\int_{\mathfrak{X}}\bar{Q}(\mathrm{d}z) = 1$. Collecting terms, conclude that $Q_{\mathcal{X}}(Q_{\mathcal{X}}^{-1}h) = h$. We showed that $Q_{\mathcal{X}}^{-1}$ is the inverse of $Q_{\mathcal{X}}$, which is by (v) a bounded operator.

To see that $Q_{\mathcal{X}}^{-1} : L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q}) \to L_2(P_{VZ})$ is bounded, take some $h \in L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})$. As $(Q_{\mathcal{X}}^{-1}h)(v,z) = \frac{1}{\alpha}h(v,z) - \frac{1-\alpha}{\alpha}\int_{\mathfrak{X}} h(v,x)\bar{Q}(\mathrm{d}x)$, we have

$$\left\|Q_{\mathcal{X}}^{-1}h\right\|_{L_2(P_{VZ})} \le \frac{1}{\alpha}\|h\|_{L_2(P_{VZ})} + \frac{1-\alpha}{\alpha}\left\|\int_{\mathfrak{X}} h(\cdot,x)\bar{Q}(\mathrm{d}x)\right\|_{L_2(P_V)},$$

where, by Lemma 3.8 (ii),

$$\begin{aligned}\|h\|_{L_2(P_{VZ})} &= \sqrt{\int h^2\,\mathrm{d}[\alpha P_{VX} + (1-\alpha)P_V\otimes\bar{Q}]}\\ \le \sqrt{\|h\|^2_{L_2(P_{VX})} + \|h\|^2_{L_2(P_V\otimes\bar{Q})}} &\le \sqrt{(\|h\|_{L_2(P_{VX})} + \|h\|_{L_2(P_V\otimes\bar{Q})})^2}\\ &\le \|h\|_{L_2(P_{VX})} + \|h\|_{L_2(P_V\otimes\bar{Q})} = \|h\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})}.\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}&\left\|\int_{\mathfrak{X}} h(\cdot,x)\bar{Q}(\mathrm{d}x)\right\|_{L_2(P_V)}\\ &\le \sqrt{\int h^2\,\mathrm{d}P_V\otimes\bar{Q}} = \|h\|_{L_2(P_V\otimes\bar{Q})} \le \|h\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})},\end{aligned}$$

whereby

$$\left\|Q_{\mathcal{X}}^{-1}h\right\|_{L_2(P_{VZ})} \leq \frac{1}{\alpha}\left\|h\right\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})} + \frac{1-\alpha}{\alpha}\left\|h\right\|_{L_2(P_{VX})\cap L_2(P_V\otimes\bar{Q})}.$$

■

Proof of Lemma 3.4. Assertion (i). Take a regular submodel $t \mapsto p_{VX,t}$ for $t \in \mathbb{R}$ in the neighbourhood of zero with $p_{VX,0} = p_{VX}$ and the property that $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} p_{VX,t} = s p_{VX}$ for the score function $s \in \mathcal{T}_{VX} \subset L_2^0(P_{VX})$ (Bolthausen et al., 2002).

First, consider the finitely distributed covariate. Propagating the submodel $t \mapsto p_{VX,t}$ through the density (3.37) in Example 3.3, we obtain

$$p_{VZ,t}(v,z) = \sum_{x\in\mathfrak{X}} Q(\{z\}\,|\,x)p_{VX,t}(v,x), \quad (v,z) \in \mathfrak{V}\times\mathfrak{Z}.$$

because Q is known. Hence, as $Q(\{z\}\,|\,x)p_{VX}(v,x)$ is the density $p_{VXZ}(v,x,z)$, the score

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\log p_{VZ,t}(v,z) &= \frac{1}{p_{VZ}(v,z)}\sum_{x\in\mathfrak{X}} Q(\{z\}\,|\,x)s(v,x)p_{VX}(v,x)\\ &= \sum_{x\in\mathfrak{X}} Q(\{z\}\,|\,x)s(v,x)p_{X|VZ}(x|\,v,z)\\ &= \mathbb{E}\left[s(V,X)\,|\,V=v,Z=z\right],\end{aligned}$$

which is equal to $(Q_{\mathcal{X}}^* s)(v,z)$, $(v,z) \in \mathfrak{V}\times\mathfrak{Z}$, by Lemma 3.3 (ii).

Second, consider the generic covariate under the model induced by Q in (3.41). From Lemma 3.8 (iii), we have the induced submodel

$$p_{VZ,t}(v,z) = \alpha p_{VX,t}(v,z) + (1-\alpha)\bar{q}(z)\int_{\mathfrak{X}} p_{VX,t}(v)\,\mathrm{d}\nu_X(x), \quad (v,z)\in\mathfrak{V}\times\mathfrak{X},$$

implying the score

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0}\log p_{VZ,t}(v,z) &= \frac{1}{p_{VZ}(v,z)}\\ \times\left\{\alpha s(v,z)p_{VX}(v,z) + (1-\alpha)\bar{q}(z)\int_{\mathfrak{X}} s(v,x)p_{VX,t}(v)\,\mathrm{d}\nu_X(x)\right\},\end{aligned}$$

for $(v,z) \in \mathfrak{V} \times \mathfrak{X}$. But Lemma 3.8 (vi) implies

$$\begin{aligned}\mathbb{E}\left[s(V,X) \mid V=v, Z=z\right] &= \int_{\mathfrak{X}} s(v,x) P_{X|VZ}(\mathrm{d}x|\, v,z) \\ &= \frac{1}{p_{VZ}(v,z)} \left\{ \int_{\mathfrak{X}} \alpha s(v,x) p_{VX}(v,z) \delta_z(\mathrm{d}x) \right. \\ &\quad \left. + (1-\alpha)\bar{q}(z) \int_{\mathfrak{X}} s(v,x) p_V(v) P_{X|V}(\mathrm{d}x|\, v) \right\} \\ = \frac{1}{p_{VZ}(v,z)} &\left\{ \alpha s(v,z) p_{VX}(v,z) + (1-\alpha)\bar{q}(z) \int_{\mathfrak{X}} s(v,u) p_{VX}(v,u)\, \mathrm{d}\nu_X(u) \right\},\end{aligned}$$

which is $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \log p_{VZ,t}(v,z)$.

Assertions (ii) and (iii). We build on Van der Vaart (1998, Chapter 25). By (i), $\mathcal{T}_{VZ} = \left\{Q_{\mathcal{X}}^* s : s \in L_2^0(P_{VX})\right\}$, which is the range $R(Q_{\mathcal{X}}^*)$ of $Q_{\mathcal{X}}^*$. It follows from the defining relation between $Q_{\mathcal{X}}$ and $Q_{\mathcal{X}}^*$ that $R(Q_{\mathcal{X}}^*)^\perp = N((Q_{\mathcal{X}}^*)^*)$, where $(Q_{\mathcal{X}}^*)^* = Q_{\mathcal{X}}$ in Hilbert spaces $L_2(P_{VX})$ and $L_2(P_{VZ})$,

$$R(Q_{\mathcal{X}}^*)^\perp \coloneqq \{k \in L_2(P_{VZ}) : P_{VZ} k \kappa = 0 \text{ holds for all } \kappa \in R(Q_{\mathcal{X}}^*)\}$$

is the orthocomplement of $R(Q_{\mathcal{X}}^*)$, and $N(Q_{\mathcal{X}}) \coloneqq \{k \in L_2(P_{VZ}) : Q_{\mathcal{X}} k = 0\}$ is the kernel of $Q_{\mathcal{X}}$. By properties of Hilbert spaces, it follows that $\overline{R(Q_{\mathcal{X}}^*)} = N(Q_{\mathcal{X}})^\perp$, where $\overline{R(Q_{\mathcal{X}}^*)}$ is the closure of $R(Q_{\mathcal{X}}^*)$ in $L_2(P_{VZ})$. Studying the kernel $N(Q_{\mathcal{X}})$, the relation

$$0 = (Q_{\mathcal{X}} k)(v,x) \text{ for all } (v,x) \in \mathfrak{V} \times \mathfrak{X},$$

for (ii) is equivalent to

$$0_J = Q^\intercal \begin{bmatrix} k(v,z_1) \\ k(v,z_2) \\ \vdots \\ k(v,z_J) \end{bmatrix} \text{ for all } v \in \mathfrak{V}$$

$$\iff (Q^\intercal)^{-1} 0_J = 0_J = \begin{bmatrix} k(v,z_1) \\ k(v,z_2) \\ \vdots \\ k(v,z_J) \end{bmatrix} \text{ for all } v \in \mathfrak{V},$$

by invertibility of Q. Hence, $k = 0$, and thus $\overline{R(Q_{\mathcal{X}}^*)} = N(Q_{\mathcal{X}})^\perp = L_2(P_{VZ})$ so that the model remains nonparametric. For (iii), $Q_{\mathcal{X}}^{-1} 0 = 0$ with $Q_{\mathcal{X}}^{-1}$ in Lemma 3.3 (vi), yields the same conclusion. ■

Proof of Proposition 3.2. We follow Van der Vaart (1998, Chapter 25.5). Consider the submodel $t \mapsto p_{VX,t}$ of Lemma 3.4 with scores $s \in \mathcal{T}_{VX} = L_2^0(P_{VX})$, implying submodel $t \mapsto P_{VX,t}$. Because $Q \in \mathcal{Q}_\psi$, these submodels induce submodel $t \mapsto p_{VZ,t}$ — which in turn induce submodel $t \mapsto P_{VZ,t}$ — and the tangent set $\mathcal{T}_{VZ}(Q, \mathrm{d}\bar{P}_{VX}) = \{Q_{\mathcal{X}}^* s : s \in L_2^0(P_{VX})\}$ by Lemma 3.4. Lemma 3.4 also implies that $\psi(P_{VZ,t}) = \chi(P_{VX,t})$ for all $t \in \mathbb{R}$ if $Q \in \mathcal{Q}_\psi$.

The efficient influence function $\tilde{\psi}$ of $\psi(P_{VZ})$ exists if and only if

$$\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \psi(P_{VZ,t}) = P_{VZ}[\tilde{\psi}(Q_{\mathcal{X}}^* s)]$$

for all regular submodels $P_{VZ,t}$ with score $Q_{\mathcal{X}}^* s \in \{Q_{\mathcal{X}}^* s : s \in L_2^0(P_{VX})\}$. But $\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \psi(P_{VZ,t}) = \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \chi(P_{VX,t})$ by the previous paragraph. Since $\tilde{\chi}$ is the efficient influence function of $\chi(P_{VX})$ by assumption, we must also have that

$$\frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \chi(P_{VX,t}) = P_{VX}[s\tilde{\chi}].$$

Thus, for all $s \in L_2^0(P_{VX})$,

$$P_{VZ}[\tilde{\psi}(Q_{\mathcal{X}}^* s)] = \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \psi(P_{VZ,t}) = \frac{\mathrm{d}}{\mathrm{d}t}\big|_{t=0} \chi(P_{VX,t}) = P_{VX}[s\tilde{\chi}].$$

By the definition of the adjoint $(Q_{\mathcal{X}}^*)^* = Q_{\mathcal{X}}$, the inner product $P_{VZ}[\tilde{\psi}(Q_{\mathcal{X}}^* s)]$ is equal to $P_{VZ}[\tilde{\psi}(Q_{\mathcal{X}}^* s)] = P_{VX}[(((Q_{\mathcal{X}}^*)^*)\tilde{\psi})s] = P_{VX}[(Q_{\mathcal{X}}\tilde{\psi})s]$, which by the last display is equal to $P_{VX}[s\tilde{\chi}]$ for all $s \in L_2^0(P_{VX})$. Hence $P_{VX}[(Q_{\mathcal{X}}\tilde{\psi})s] = P_{VX}[s\tilde{\chi}]$ for all $s \in L_2^0(P_{VX})$, or, by rearrangement, $P_{VX}[(Q_{\mathcal{X}}\tilde{\psi} - \tilde{\chi})s] = 0$ for all $s \in L_2^0(P_{VX})$. Equivalently, $Q_{\mathcal{X}}\tilde{\psi} - \tilde{\chi}$ must be in the orthocomplement of $L_2^0(P_{VX}) \subset L_2(P_{VX})$, which, as we show below, is

$$L_2^0(P_{VX})^\perp = \{f \in L_2(P_{VX}) : f - P_{VX} f = 0 \, P_{VX}\text{-a.s.}\}\,.$$

Now, $\tilde{\chi}$ is the efficient influence function for $\chi(P_{VX})$, so $P_{VX}\tilde{\chi} = 0$. For $\tilde{\psi}$ to be an influence function for $\psi(P_{VZ})$, we must have $P_{VZ}\tilde{\psi} = 0$, but by Lemma 3.3(iii), $P_{VZ}\tilde{\psi} = P_{VX}Q_{\mathcal{X}}\tilde{\psi}$. Hence, $\tilde{\chi}, Q_{\mathcal{X}}\tilde{\psi} \in L_2^0(P_{VX})$ and $Q_{\mathcal{X}}\tilde{\psi} - \tilde{\chi} \in L_2^0(P_{VX})$. But $Q_{\mathcal{X}}\tilde{\psi} - \tilde{\chi} \in L_2^0(P_{VX})^\perp$ too as we showed above. Since $L_2^0(P_{VX}) \cap L_2^0(P_{VX})^\perp = \{f : f = 0 \, P_{VX}\text{-a.s.}\}$, we must have $Q_{\mathcal{X}}\tilde{\psi} = \tilde{\chi}$ P_{VX}-a.s.. Because $Q \in \mathcal{Q}_\psi$, $Q_{\mathcal{X}}^{-1}$ exists, giving $\tilde{\psi} = Q_{\mathcal{X}}^{-1}\tilde{\chi}$.

To see that the orthocomplement $L_2^0(P_{VX})^\perp$ of $L_2^0(P_{VX})$ in $L_2(P_{VX})$ is

$$\{f \in L_2(P_{VX}) : f - P_{VX} f = 0 \, P_{VX}\text{-a.s.}\}\,,$$

take some $f \in L_2^0(P_{VX})^\perp$. Because $f \in L_2^0(P_{VX})^\perp$ and $f - P_{VX}f \in L_2^0(P_{VX})$, we must have $P_{VX}[f(f - P_{VX}f)] = 0$. Because $P_{VX}[f(f - P_{VX}f)] = P_{VX}[(f - P_{VX}f)(f - P_{VX}f)]$, we must have $f - P_{VX}f = 0$ P_{VX}-a.s., hence $f = 0$ P_{VX}-a.s.. Because $f \in L_2^0(P_{VX})^\perp$ was chosen arbitrarily, the assertion follows. ■

Proof of Lemma 3.5. As in the proof of Lemma 3.1, we apply that if a P_{VZ}-random function $\hat{q} \in L_2(P_{VZ})$ is independent of the random sample generating the process $\bar{\mathbb{P}}_n$, then

$$\int (\hat{q}(v,z) - q(v,z))^2 \, \mathrm{d}P_{VZ}(v,z) = o_{P_{VZ}}(1) \text{ implies } \sqrt{n}(\bar{\mathbb{P}}_n - P_{VZ})(\hat{q} - q) = o_{P_{VZ}}(1). \tag{3.103}$$

In particular, we shall combine this with Lemma 3.9, establishing the boundedness of $Q_{\mathcal{X}}^{-1}$ for $\|\cdot\|_{\mathrm{L}_2}$, to show the convergence of

$$\left\|Q_{\mathcal{X}}^{-1}T\right\|_{L_2(P_{VZ})} \lesssim \|T\|_{\mathrm{L}_2} \tag{3.104}$$

to zero in P_{VZ}-probability for some $T \in \mathrm{L}_2$.

By the linearity of $Q_{\mathcal{X}}^{-1}$, $\hat{\check{\psi}} - \tilde{\psi} = Q_{\mathcal{X}}^{-1}(\check{\check{\chi}} - \tilde{\chi})$. By the definitions (3.9) and (3.49),

$$\begin{aligned}
\check{\check{\chi}}(v,x) - \tilde{\chi}(v,x) &= \bar{T}_1(v,x) + \bar{T}_2(v,x) + \bar{T}_3(v,x) + \bar{T}_4, \qquad (3.105)\\
\bar{T}_1(v,x) &:= \check{r}(v_1,x)(m(v,x) - \check{\mu}_{\mathcal{X}}(v_1,x)) \\
&\quad - r(v_1,x)(m(v,x) - \mu_{\mathcal{X}}(v_1,x)),\\
\bar{T}_2(v,x) &:= \frac{\mathbb{1}_{v_2=c}}{\check{p}_{V_2}(c)}(g(v,x) - \check{\gamma}_{\mathcal{V}}(c))\check{e} - \frac{\mathbb{1}_{v_2=c}}{p_{V_2}(c)}(g(v,x) - \gamma_{\mathcal{V}}(c))e,\\
\bar{T}_3(v,x) &:= f(v,x,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) - f(v,x,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)),\\
\bar{T}_4 &:= -\psi(\hat{P}_{VZ}) + \psi(P_{VZ}).
\end{aligned}$$

As $\bar{T}_4$ is constant, not depending on (v,x), $Q_{\mathcal{X}}^{-1}\bar{T}_4 = \bar{T}_4$ and $(\bar{\mathbb{P}}_n - P_{VZ})Q_{\mathcal{X}}^{-1}\bar{T}_4 = 0$. It remains to show $(\bar{\mathbb{P}}_n - P_{VZ})Q_{\mathcal{X}}^{-1}\bar{T}_j = o_{P_{VZ}}\left(n^{-1/2}\right)$ for $j = 1,2,3$ by the linearity of the process $\bar{\mathbb{P}}_n - P_{VZ}$.

Term $\bar{T}_1$. In the light of (3.104), $(\bar{\mathbb{P}}_n - P_{VZ})Q_{\mathcal{X}}^{-1}\bar{T}_1 = o_{P_{VZ}}\left(n^{-1/2}\right)$ can be established along the same steps as that of T_1 in the proof of Lemma 3.1. In particular, $\left\|Q_{\mathcal{X}}^{-1}\bar{T}_1\right\|_{L_2(P_{VZ})} \lesssim \left\|\bar{T}_1\right\|_{\mathrm{L}_2}$. For completeness, we present the

derivation. Suppressing the arguments, write

$$\begin{aligned}\bar{T}_1 = \check{r}(m-\check{\mu}_{\mathcal{X}}) - r(m-\mu_{\mathcal{X}}) &= (\check{r}-r+r)(m-\check{\mu}_{\mathcal{X}}) - r(m-\mu_{\mathcal{X}}) \\ &= (\check{r}-r)(m-\check{\mu}_{\mathcal{X}}) + r(\mu_{\mathcal{X}}-\check{\mu}_{\mathcal{X}}).\end{aligned}$$

By Assumption 3.3, $\|m-\check{\mu}_{\mathcal{X}}\|_\infty = O_{P_{VZ}}(1)$; either directly by (3.62), or by (3.60) and (3.61), noting that $\|m-\check{\mu}_{\mathcal{X}}\|_\infty \leq \|m-\mu_{\mathcal{X}}\|_\infty + \|\mu_{\mathcal{X}}-\check{\mu}_{\mathcal{X}}\|_\infty = O(1) + o_{P_{VZ}}(1) = O_{P_{VZ}}(1)$. Then the convergence (3.57) of r implies that $(\bar{\mathbb{P}}_n - P_{VZ})Q_{\mathcal{X}}^{-1}((\check{r}-r)(m-\check{\mu}_{\mathcal{X}})) = o_{P_{VZ}}\left(n^{-1/2}\right)$ by (3.104) as

$$\|(\check{r}-r)(m-\check{\mu}_{\mathcal{X}})\|_{\mathrm{L}_2} \leq \|m-\check{\mu}_{\mathcal{X}}\|_\infty \|\check{r}-r\|_{\mathrm{L}_2} = O_{P_{VZ}}(1)\, o_{P_{VZ}}(1) = o_{P_{VZ}}(1)$$

since $\left\||q|^2\right\|_\infty = \|q\|_\infty^2$.

By Assumption 3.3, either (3.61), or (3.63) and (3.64). In the former case,

$$\|r(\mu_{\mathcal{X}}-\check{\mu}_{\mathcal{X}})\|_{\mathrm{L}_2} \leq \|\mu_{\mathcal{X}}-\check{\mu}_{\mathcal{X}}\|_\infty \|r\|_{\mathrm{L}_2} = o_{P_{VZ}}(1),$$

because $r \in \mathrm{L}_2$. In the latter case,

$$\|r(\mu_{\mathcal{X}}-\check{\mu}_{\mathcal{X}})\|_{\mathrm{L}_2} \leq \bar{R}\,\|\mu_{\mathcal{X}}-\check{\mu}_{\mathcal{X}}\|_{\mathrm{L}_2} = o_{P_{VZ}}(1),$$

since (3.64) bounds r and $\check{\mu}_{\mathcal{X}}$ is convergent by (3.63). Thus, $(\bar{\mathbb{P}}_n - P_{VZ})Q_{\mathcal{X}}^{-1}(r(\mu_{\mathcal{X}}-\check{\mu}_{\mathcal{X}})) = o_{P_{VZ}}\left(n^{-1/2}\right)$ by (3.103). Conclude that $(\bar{\mathbb{P}}_n - P_{VZ})Q_{\mathcal{X}}^{-1}\bar{T}_1 = o_{P_{VZ}}\left(n^{-1/2}\right)$.

Term $\bar{T}_2$. By the mean-value theorem there exists $(\tilde{\gamma}_{\mathcal{V}}(c), \tilde{p}_{V_2}(c), \tilde{e})$ between $(\gamma_{\mathcal{V}}(c), p_{V_2}(c), e)$ and $(\check{\gamma}_{\mathcal{V}}(c), \check{p}_{V_2}(c), \check{e})$ such that

$$\begin{aligned}\bar{T}_2(v,x) &= \frac{\mathbb{1}_{v_2=c}}{\check{p}_{V_2}(c)}(g(v,x)-\check{\gamma}_{\mathcal{V}}(c))\check{e} - \frac{\mathbb{1}_{v_2=c}}{p_{V_2}(c)}(g(v,x)-\gamma_{\mathcal{V}}(c))e \\ &= -\frac{\mathbb{1}_{v_2=c}}{\tilde{p}_{V_2}(c)}\tilde{e}(\check{\gamma}_{\mathcal{V}}(c)-\gamma_{\mathcal{V}}(c)) \\ &\quad - \frac{\mathbb{1}_{v_2=c}}{\tilde{p}_{V_2}(c)^2}(g(v,x)-\tilde{\gamma}_{\mathcal{V}}(c))\tilde{e}(\check{p}_{V_2}(c)-p_{V_2}(c)) \\ &\quad + \frac{\mathbb{1}_{v_2=c}}{\tilde{p}_{V_2}(c)}(g(v,x)-\tilde{\gamma}_{\mathcal{V}}(c))(\check{e}-e).\end{aligned}$$

The standard central limit theorem applies to the i.i.d. sequence

$$\left((Q_{\mathcal{X}}^{-1}(v,x) \mapsto \mathbb{1}_{v_2=c})(V_i,Z_i)\right)_{i\in[n]},$$

hence $\sqrt{n}(\bar{\mathbb{P}}_n - P_{VZ})((Q_{\mathcal{X}}^{-1}(v,x) \mapsto \mathbb{1}_{v_2=c})(V,Z)) = O_{P_{VZ}}(1)$. By the linearity of the process $\sqrt{n}(\bar{\mathbb{P}}_n - P_{VZ})Q_{\mathcal{X}}^{-1}$,

$$\begin{aligned}
\sqrt{n}(\bar{\mathbb{P}}_n - P_{VZ})\big\{\big[Q_{\mathcal{X}}^{-1}(v,x) \mapsto \mathbb{1}_{v_2=c} g(v,x) - \mathbb{1}_{v_2=c}\tilde{\gamma}_{\mathcal{V}}(c)\big](V,Z)\big\} \\
= \sqrt{n}(\bar{\mathbb{P}}_n - P_{VZ})\big\{\big[Q_{\mathcal{X}}^{-1}(v,x) \mapsto \mathbb{1}_{v_2=c} g(v,x)\big](V,Z)\big\} \\
-\tilde{\gamma}_{\mathcal{V}}(c)\sqrt{n}(\bar{\mathbb{P}}_n - P_{VZ})\big\{\big[Q_{\mathcal{X}}^{-1}(v,x) \mapsto \mathbb{1}_{v_2=c}\big](V,Z)\big\} \\
= (1 - \tilde{\gamma}_{\mathcal{V}}(c))O_{P_{VZ}}(1) = O_{P_{VZ}}(1)
\end{aligned}$$

again by the standard central limit theorem and (3.58). Suppose that $\check{e} - e = o_{P_{VZ}}(1)$, which we show later. Then by (3.58) and (3.59), $(\bar{\mathbb{P}}_n - P_{VZ})Q_{\mathcal{X}}^{-1}\bar{T}_2 = o_{P_{VZ}}\left(n^{-1/2}\right)$.

Term $\bar{T}_3$. Recall that $\bar{T}_3(v,x) = f(v,x,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) - f(v,x,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))$. By the consistency of $\check{\gamma}_{\mathcal{V}}$ and $\check{\mu}_{\mathcal{X}}$ ((3.58) and (3.61) or (3.63)), we have $\left\|\bar{T}_3\right\|_{\mathrm{L}_2} = o_{P_{VZ}}(1)$ by the continuous mapping theorem and (3.55). Conclude by (3.103) and (3.104) that $(\bar{\mathbb{P}}_n - P_{VZ})Q_{\mathcal{X}}^{-1}\bar{T}_3 = o_{P_{VZ}}\left(n^{-1/2}\right)$.

Consistency of $\check{e}$. By the definition of $e, \check{e}$,

$$\begin{aligned}
\check{e} - e =& \bar{\mathbb{P}}_n''\partial_\gamma \bar{f}(V,Z,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) - P_{VZ}\partial_\gamma \bar{f}(V,Z,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) \\
=& \bar{\mathbb{P}}_n''\partial_\gamma \bar{f}(V,Z,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) - P_{VZ}\partial_\gamma \bar{f}(V,Z,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) \\
&+ P_{VZ}\partial_\gamma \bar{f}(V,Z,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) - P_{VZ}\partial_\gamma f(V,Z,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) \\
=& (\bar{\mathbb{P}}_n'' - P_{VZ})\partial_\gamma \bar{f}(V,Z,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) \\
&+ P_{VZ}\big[\partial_\gamma \bar{f}(V,Z,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) - \partial_\gamma \bar{f}(V,Z,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\big] \\
=& (\bar{\mathbb{P}}_n'' - P_{VZ})\partial_\gamma \bar{f}(V,Z,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c)) \\
&+ (\bar{\mathbb{P}}_n'' - P_{VZ})\big[\partial_\gamma \bar{f}(V,Z,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) - \partial_\gamma \bar{f}(V,Z,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\big] \\
&+ P_{VZ}\big[\partial_\gamma \bar{f}(V,Z,\check{\mu}_{\mathcal{X}},\check{\gamma}_{\mathcal{V}}(c)) - \partial_\gamma \bar{f}(V,Z,\mu_{\mathcal{X}},\gamma_{\mathcal{V}}(c))\big]
\end{aligned}$$

Here, the first term is $O_{P_{VZ}}\left(n^{-1/2}\right) = o_{P_{VZ}}(1)$ by the standard central limit theorem, and the second and third term are $o_{P_{VZ}}(1)$ by the continuity (3.56) of $\partial_\gamma f$ using that $\partial_\gamma \bar{f} = Q_{\mathcal{X}}^{-1}\partial_\gamma f$ along the same arguments concerning $\bar{T}_3$ above. ■

Proof of Corollary 3.2. Follows from Lemma 3.5 and Theorem 3.1, noting that, for $\bar{e}'$ in (3.66),

$$\check{e} - \bar{e}' = (\bar{\mathbb{P}}_n'' - P_{VZ})\partial_\gamma \bar{f}(V,Z,\hat{\mu}_{\mathcal{X}},\hat{\gamma}_{\mathcal{V}}(c))$$

is $o_{P_{VZ}}(1)$ by the consistency proof of $\check{e}$ in Lemma 3.5. ■

3.B.1. Auxiliary Results

Lemma 3.8 (Distributions under (3.41))**.** *Suppose that the privacy mechanism $Q \in \mathcal{Q}(\mathfrak{X} \to \mathfrak{Z})$ is equal to (3.41). Then the following assertions hold true.*

(i) The joint distribution of (V, X, Z) is $P_{VXZ}(B_{\mathsf{v}}, B_{\mathsf{x}}, B_{\mathsf{z}}) = \alpha P_{VX}(B_{\mathsf{v}}, B_{\mathsf{x}} \cap B_{\mathsf{z}}) + (1-\alpha)\bar{Q}(B_{\mathsf{z}})P_{VX}(B_{\mathsf{v}}, B_{\mathsf{x}})$ for $B_{\mathsf{v}} \in \mathscr{F}_{\mathfrak{V}}, B_{\mathsf{x}} \in \mathscr{F}_{\mathfrak{X}}, B_{\mathsf{z}} \in \mathscr{F}_{\mathfrak{Z}}$.

(ii) The joint distribution of (V, Z) is $P_{VZ}(B_{\mathsf{v}}, B_{\mathsf{z}}) = \alpha P_{VX}(B_{\mathsf{v}}, B_{\mathsf{z}}) + (1-\alpha)\bar{Q}(B_{\mathsf{z}})P_V(B_{\mathsf{v}})$ for $B_{\mathsf{v}} \in \mathscr{F}_{\mathfrak{V}}, B_{\mathsf{z}} \in \mathscr{F}_{\mathfrak{Z}}$.

(iii) The $\nu_V \times \nu_X$-density of P_{VX} is $p_{VZ}(v, z) \coloneqq \alpha p_{VX}(v, z) + (1-\alpha)\bar{q}(z)p_V(v)$ for $(v, z) \in \mathfrak{V} \times \mathfrak{X}$.

(iv) For all $h : \mathfrak{V} \times \mathfrak{X} \to \mathbb{R}$ and all $p \in [1, \infty]$, $\|h\|_{L_p(P_{VX})} \leq \alpha^{-1/p} \|h\|_{L_p(P_{VZ})}$.

(v) The Markov kernel

$$\begin{aligned} P_{X|Z}(B|\, z) &\coloneqq \alpha \frac{p_X(z)}{p_Z(z)} \delta_z(B) + (1-\alpha)\frac{\bar{q}(z)}{p_Z(z)} \int_B p_X(x)\, \mathrm{d}\nu_X(x) \\ &= \alpha \frac{p_X(z)}{p_Z(z)} \delta_z(B) + (1-\alpha)\frac{\bar{q}(z)}{p_Z(z)} P_X(B) \end{aligned}$$

for $B \in \mathscr{F}_{\mathfrak{X}}$ is the conditional distribution of X given $Z = z$, $z \in \mathfrak{X}$.

(vi) The Markov kernel

$$\begin{aligned} P_{X|V_{\mathrm{s}}Z}(B|\, v_{\mathrm{s}}, z) \coloneqq{}& \alpha \frac{p_{V_{\mathrm{s}}X}(v_{\mathrm{s}}, z)}{p_{V_{\mathrm{s}}Z}(v_{\mathrm{s}}, z)} \delta_z(B) \\ &+ (1-\alpha)\frac{\bar{q}(z)}{p_{V_{\mathrm{s}}Z}(v_{\mathrm{s}}, z)} \int_B p_{V_{\mathrm{s}}X}(v_{\mathrm{s}}, x)\, \mathrm{d}\nu_X(x) \\ ={}& \alpha \frac{p_{V_{\mathrm{s}}X}(v_{\mathrm{s}}, z)}{p_{V_{\mathrm{s}}Z}(v_{\mathrm{s}}, z)} \delta_z(B) \\ &+ (1-\alpha)\frac{\bar{q}(z)}{p_{V_{\mathrm{s}}Z}(v_{\mathrm{s}}, z)} p_{V_{\mathrm{s}}}(v_{\mathrm{s}}) P_{X|V_{\mathrm{s}}}(B|\, v_{\mathrm{s}}) \end{aligned}$$

for $B \in \mathscr{F}_{\mathfrak{X}}$ is the conditional distribution of X given $(V_{\mathrm{s}}, Z) = (v_{\mathrm{s}}, z)$, $(v_{\mathrm{s}}, z) \in \mathfrak{V}_{\mathrm{s}} \times \mathfrak{X}$ for any subset V_{s} of V.

(vii) Let $V_{\mathrm{s}}, V_{\bar{\mathrm{s}}}$ be a partition of V, that is $V_{\mathrm{s}} \cup V_{\bar{\mathrm{s}}} = V$ and $V_{\mathrm{s}} \cap V_{\bar{\mathrm{s}}} = \varnothing$. The Markov kernel

$$P_{V_{\bar{\mathrm{s}}}X|V_{\mathrm{s}}Z}(B_{\mathsf{v}_{\mathrm{s}}}, B_{\mathsf{x}}|\, v_{\mathrm{s}}, z) := \alpha \frac{p_{V_{\mathrm{s}}X}(v_{\mathrm{s}}, z)}{p_{V_{\mathrm{s}}Z}(v_{\mathrm{s}}, z)} P_{V_{\bar{\mathrm{s}}}|V_{\mathrm{s}}X}(B_{\mathsf{v}_{\mathrm{s}}}|\, v_{\mathrm{s}}, z)\delta_z(B_{\mathsf{x}}) + (1-\alpha)\frac{\bar{q}(z)}{p_{V_{\mathrm{s}}Z}(v_{\mathrm{s}}, z)} p_{V_{\mathrm{s}}}(v_{\mathrm{s}}) P_{V_{\bar{\mathrm{s}}}X|V_{\mathrm{s}}}(B_{\mathsf{v}_{\mathrm{s}}}, B_{\mathsf{x}}|\, v_{\mathrm{s}})$$

for $B_{\mathsf{v}_{\mathrm{s}}} \in \mathscr{F}_{\mathfrak{V}_{\bar{\mathrm{s}}}}, B_{\mathsf{x}} \in \mathscr{F}_{\mathfrak{X}}$ is the conditional distribution of $(V_{\bar{\mathrm{s}}}, X)$ given $(V_{\mathrm{s}}, Z) = (v_{\mathrm{s}}, z)$, $(v_{\mathrm{s}}, z) \in \mathfrak{V}_{\mathrm{s}} \times \mathfrak{X}$.

Proof of Lemma 3.8. Assertion (i). Plug (3.41) into (3.32), using the definition of the Dirac measure $\delta_x(B) = \mathbb{1}_{x\in B}$, to find

$$\begin{aligned} P_{VXZ}(B_{\mathsf{v}}, B_{\mathsf{x}}, B_{\mathsf{z}}) &= \alpha \int \mathbb{1}_{(v,x)\in B_{\mathsf{v}}\times B_{\mathsf{x}}} \mathbb{1}_{x\in B_{\mathsf{z}}} \,\mathrm{d}P_{VX}(v,x) \\ &\quad + (1-\alpha)\bar{Q}(B_{\mathsf{z}}) \int \mathbb{1}_{(v,x)\in B_{\mathsf{v}}\times B_{\mathsf{x}}} \,\mathrm{d}P_{VX}(v,x) \\ &= \alpha \int \mathbb{1}_{(v,x)\in B_{\mathsf{v}}\times (B_{\mathsf{x}}\cap B_{\mathsf{z}})} \,\mathrm{d}P_{VX}(v,x) \\ &\quad + (1-\alpha)\bar{Q}(B_{\mathsf{z}}) \int \mathbb{1}_{(v,x)\in B_{\mathsf{v}}\times B_{\mathsf{x}}} \,\mathrm{d}P_{VX}(v,x) \\ &= \alpha P_{VX}(B_{\mathsf{v}}, B_{\mathsf{x}}\cap B_{\mathsf{z}}) + (1-\alpha)\bar{Q}(B_{\mathsf{z}})P_{VX}(B_{\mathsf{v}}, B_{\mathsf{x}}). \end{aligned}$$

Assertion (ii). Follows from (i) as the marginal distribution by setting $B_{\mathsf{x}} := \mathfrak{X}$.

Assertion (iii). A convex combination of two densities, p_{VZ} is nonnegative. From (ii), write $P_{VZ}(B_{\mathsf{v}}, B_{\mathsf{z}})$ as

$$\begin{aligned} &\alpha \int_{B_{\mathsf{v}}}\int_{B_{\mathsf{z}}} p_{VX}(v,x)\,\mathrm{d}\nu_X(x)\,\mathrm{d}\nu_V(v) + (1-\alpha)\int_{B_{\mathsf{z}}} \bar{q}(z)\,\mathrm{d}\nu_X(x)\int_{B_{\mathsf{v}}} p_V(v)\,\mathrm{d}\nu_V(v) \\ &\qquad = \alpha \int_{B_{\mathsf{v}}}\int_{B_{\mathsf{z}}} \left\{\alpha p_{VX}(v,x) + (1-\alpha)\bar{q}(x)p_V(v)\right\}\mathrm{d}\nu_X(x)\,\mathrm{d}\nu_V(v). \end{aligned}$$

Assertion (iv). By (iii),

$$p_{VX}(v,x) = \frac{1}{\alpha}p_{VZ}(v,x) - \frac{1-\alpha}{\alpha}\bar{q}(x)p_V(v) \le \frac{1}{\alpha}p_{VZ}(v,x)$$

because $-\frac{1-\alpha}{\alpha}\bar{q}(x)p_V(v) \le 0$ for $\alpha \in (0,1]$. Then

$$\begin{aligned} \|h\|_{L_p(P_{VX})} &= \left(\int_{\mathfrak{X}}\int_{\mathfrak{V}} |h(v,x)|^p p_{VX}(v,x)\,\mathrm{d}\nu_V(v)\,\mathrm{d}\nu_X(x)\right)^{1/p} \\ &\le \alpha^{-1/p}\left(\int_{\mathfrak{X}}\int_{\mathfrak{V}} |h(v,z)|^p p_{VZ}(v,x)\,\mathrm{d}\nu_V(v)\,\mathrm{d}\nu_X(x)\right)^{1/p} = \alpha^{-1/p}\,\|h\|_{L_p(P_{VZ})}. \end{aligned}$$

Assertion (v). It is sufficient and necessary to verify that for the $P_{X|Z}$ given in (v), $P_{XZ}(B_{\mathsf{x}}, B_{\mathsf{z}}) = \int_{B_{\mathsf{z}}} P_{X|Z}(B_{\mathsf{x}}|\, z) \frac{\mathrm{d}P_Z}{\mathrm{d}\nu_X}(z)\, \mathrm{d}\nu_X(z)$, where $\frac{\mathrm{d}P_Z}{\mathrm{d}\nu_X} = p_Z$ is the ν_X-density of P_Z derived from (iii). From the right, using the definition of the Dirac measure,

$$\begin{aligned}
&\int_{B_{\mathsf{z}}} P_{X|Z}(B_{\mathsf{x}}|\, z) p_Z(z)\, \mathrm{d}\nu_X(z) \\
&= \int_{B_{\mathsf{z}}} \left\{ \alpha \frac{p_X(z)}{p_Z(z)} \delta_z(B_{\mathsf{x}}) + (1-\alpha) \frac{\bar{q}(z)}{p_Z(z)} \int_{B_{\mathsf{x}}} p_X(x)\, \mathrm{d}\nu_X(x) \right\} p_Z(z)\, \mathrm{d}\nu_X(z) \\
&= \alpha \int_{B_{\mathsf{z}} \cap B_{\mathsf{x}}} p_X(z)\, \mathrm{d}\nu_X(z) + (1-\alpha) \int_{B_{\mathsf{z}}} \bar{q}(z)\, \mathrm{d}\nu_X(z) \int_{B_{\mathsf{x}}} p_X(x)\, \mathrm{d}\nu_X(x) \\
&= \alpha P_X(B_{\mathsf{x}} \cap B_{\mathsf{z}}) + (1-\alpha) \bar{Q}(B_{\mathsf{z}}) P_X(B_{\mathsf{x}}),
\end{aligned}$$

in which we recognise $P_{XZ}(B_{\mathsf{x}}, B_{\mathsf{z}})$ by (i).

Assertions (vi) and (vii). Follow from the same arguments as (v). ■

Lemma 3.9 (Norms and continuity of $Q_{\mathcal{X}}, Q_{\mathcal{X}}^{-1}$)**.** *Let $c_Q > 0$ denote a constant which depends only on Q and whose value may differ in every display in this lemma. Assume that $Q \in \mathcal{Q}_\psi$.*

If in the case of $|\mathfrak{X}| = J < \infty$, we also have $\inf_{(z,x)\in\mathfrak{X}^2} Q(\{z\}\,|\,x) > 0$, then for all $h : \mathfrak{V} \times \mathfrak{X} \to \mathbb{R}$ and all $p \in [1, \infty]$,

$$\|h\|_{L_p(P_{VX})} \leq c_Q \|h\|_{L_p(P_{VZ})}. \tag{3.106}$$

Let $Q_{\mathcal{X}}$ in (3.43) be induced by $Q \in \mathcal{Q}_\psi$. Then for all $k \in L_2(P_{VZ})$,

$$\|Q_{\mathcal{X}} k\|_{L_2(P_{VX})} \leq c_Q \|k\|_{L_2(P_{VZ})}. \tag{3.107}$$

If (V, X) is distributed on a finite set and $\inf_{(v,x)\in\mathfrak{V}\times\mathfrak{X}} p_{VX}(v, x) > 0$, then for all $h \in L_2(P_{VX})$,

$$\left\|Q_{\mathcal{X}}^{-1} h\right\|_{L_2(P_{VZ})} \leq c_Q \|h\|_{L_2(P_{VX})}. \tag{3.108}$$

If Q is equal to (3.41), we have, for all $h \in L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})$,

$$\left\|Q_{\mathcal{X}}^{-1} h\right\|_{L_2(P_{VZ})} \leq c_Q \|h\|_{L_2(P_{VX}) \cap L_2(P_V \otimes \bar{Q})}. \tag{3.109}$$

Proof of Lemma 3.9. Suppose that Q is (3.41). Then (3.106) is Lemma 3.8(iv); (3.107) and (3.109) are Lemma 3.3(v) and (vi), respectively.

Suppose now that X is distributed on a finite set $\mathfrak{X} = \mathfrak{Z}$. First, we show (3.106). We have, for any $c > 0$,

$$\begin{aligned} &\|h\|^p_{L_p(P_{VZ})} - \frac{1}{c}\|h\|^p_{L_p(P_{VX})} \\ &= \int_{\mathfrak{V}} \sum_{x\in\mathfrak{X}} |h(v,x)|^p \left(p_{VZ}(v,x) - \frac{1}{c} p_{VX}(v,x) \right) \mathrm{d}\nu_V(v). \end{aligned}$$

By assumption,

$$\underline{q} \coloneqq \min_{(x,\bar{x})\in\mathcal{X}^2} Q(\{x\} \,|\, \bar{x}) > 0.$$

Since $p_{VZ}(v,x) = \sum_{\bar{x}\in\mathfrak{X}} Q(\{x\} \,|\, \bar{x}) p_{VX}(v,\bar{x})$, we have

$$\begin{aligned} p_{VZ}(v,x) - \frac{1}{c} p_{VX}(v,x) &\geq \underline{q} \sum_{\bar{x}\in\mathfrak{X}} p_{VX}(v,\bar{x}) - \frac{1}{c} p_{VX}(v,x) \\ &= \underline{q} \sum_{\bar{x}\in\mathfrak{X}:\bar{x}\neq x} p_{VX}(v,\bar{x}) + \left(\underline{q} - \frac{1}{c} \right) p_{VX}(v,x). \end{aligned}$$

Hence, setting $c \coloneqq 1/\underline{q}$ implies $p_{VZ}(v,x) - \frac{1}{c} p_{VX}(v,x) \geq 0$. Thus

$$\|h\|^p_{L_p(P_{VZ})} - \frac{1}{c}\|h\|^p_{L_p(P_{VX})} \geq 0.$$

Second, we show (3.107). By Lemma 3.3(i) and Jensen's inequality,

$$((Q_{\mathcal{X}} k)(v,x))^2 \leq \mathbb{E}\left[k(V,Z)^2 \,\middle|\, V = v, X = x \right].$$

But then

$$\begin{aligned} &\|Q_{\mathcal{X}} k\|^2_{L_2(P_{VX})} = \mathbb{E}\left[((Q_{\mathcal{X}} k)(V,X))^2 \right] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[k(V,Z)^2 \,\middle|\, V, X \right] \right] = \mathbb{E} k(V,Z)^2 = \|k\|^2_{L_2(P_{VZ})}. \end{aligned}$$

Third, we show (3.108). Consider the matrix representation of $Q_{\mathcal{X}}^{-1}$ in Lemma 3.3 (iv). The linear operator (matrix) $Q : \mathbb{R}^{J\times J} \to \mathbb{R}^{J\times 1}$ has inverse Q^{-1} because $Q \in \mathcal{Q}_\psi$. As $(Q^{-1})^\intercal : \mathbb{R}^{J\times J} \to \mathbb{R}^{J\times 1}$ is a linear operator on a finite-dimensional space, it is a continuous and bounded linear operator (e.g. Kress (2014, Theorem 2.4)). Whence, $\left\|(Q^{-1})^\intercal t\right\|_2 \lesssim \|t\|_2$ for all $t \in \mathbb{R}^{J\times 1}$,

where $\|.\|_2$ is the Euclidean norm on $\mathbb{R}^{J\times 1}$. The assertion follows by

$$\begin{aligned}\left\|Q_{\mathcal{X}}^{-1}h\right\|_{L_2(P_{VZ})}^2 &= \sum_{v\in\mathfrak{V}}\sum_{z\in\mathfrak{Z}}\left[(Q_{\mathcal{X}}^{-1}h)(v,z)\right]^2 p_{VZ}(v,z) \leq \sum_{v\in\mathfrak{V}}\sum_{z\in\mathfrak{Z}}\left[(Q_{\mathcal{X}}^{-1}h)(v,z)\right]^2 \\ &= \sum_{v\in\mathfrak{V}}\sum_{z\in\mathfrak{Z}}\left[\sum_{x\in\mathfrak{X}}((Q^{-1})^\intercal)_{z,x}h(v,x)\right]^2 = \sum_{v\in\mathfrak{V}}\left\|(Q^{-1})^\intercal(h(v,x))_{x\in\mathcal{X}}\right\|_2^2 \\ &\lesssim \sum_{v\in\mathfrak{V}}\|(h(v,x))_{x\in\mathcal{X}}\|_2^2 = \sum_{v\in\mathfrak{V}}\sum_{x\in\mathfrak{X}}h(v,x)^2 \\ &\simeq \sum_{v\in\mathfrak{V}}\sum_{x\in\mathfrak{X}}h(v,x)^2 p_{VX}(v,x) = \|h\|_{L_2(P_{VX})},\end{aligned}$$

since $\inf_{(v,x)\in\mathfrak{V}\times\mathfrak{X}} p_{VX}(v,x) > 0$ by assumption. ■

3.C. Estimation of Nuisance Parameters

This section contains the proofs of results in Section 3.7: Lemmas 3.6 and 3.7, Proposition 3.3, and Corollaries 3.5 and 3.6. The proofs of Corollaries 3.3 and 3.4 are omitted.

Proof of Lemma 3.6. We have the identification $P_{VX}\phi_{\theta_0} = 0_K$. The conditions of the lemma are versions of Assumptions 2.1–2.5 and 3.1–3.6 in Hansen (1982) adapted to our setting, whereby Theorems 2.1 and 3.2 *ibid* apply, giving $\sqrt{n}(\hat{\theta}-\theta_0) \overset{P_{VX}}{\rightsquigarrow} \mathcal{N}(0,\Sigma)$ (see also Newey and McFadden (1994, Theorems 2.6, 3.4)). Then by the mean-value theorem and (3.74),

$$\left\|\xi_{\hat{\theta}} - \xi_{\theta_0}\right\|_{L_2(P_{VX})} \leq \left\|\hat{\theta}-\theta_0\right\|_2 \left\|\sup_{\tilde{\theta}\in\mathrm{Nb}(\theta_0)}\left\|\mathrm{D}_\theta\xi_{\tilde{\theta}}\right\|_2\right\|_{L_2(P_{VX})} = O_{P_{VX}}\left(n^{-1/2}\right).$$

■

Proof of Proposition 3.3. Because $Q_{\mathcal{X}}^{-1}$ and differentiation with respect to θ commutes,

$$\bar{\phi}_{\tilde{\theta}}^\intercal = \mathrm{D}_\theta\bar{\Xi}_{\tilde{\theta}} = \mathrm{D}_\theta Q_{\mathcal{X}}^{-1}\Xi_{\tilde{\theta}} = Q_{\mathcal{X}}^{-1}\mathrm{D}_\theta\Xi_{\tilde{\theta}} = Q_{\mathcal{X}}^{-1}\phi_{\tilde{\theta}}^\intercal,$$

and then also

$$\mathrm{D}_\theta\bar{\phi}_{\tilde{\theta}} = \mathrm{D}_\theta Q_{\mathcal{X}}^{-1}\phi_{\tilde{\theta}} = Q_{\mathcal{X}}^{-1}\mathrm{D}_\theta\phi_{\tilde{\theta}} = Q_{\mathcal{X}}^{-1}\dot{\phi}_{\tilde{\theta}}.$$

(Here, we denote with $Q_{\mathcal{X}}^{-1}\phi_{\tilde{\theta}}^{\intercal}$ the vector and with $Q_{\mathcal{X}}^{-1}\dot{\phi}_{\tilde{\theta}}$ the matrix where $Q_{\mathcal{X}}^{-1}$ is applied element-wise to the coordinate functions of $\phi_{\tilde{\theta}}^{\intercal}$ and $\dot{\phi}_{\tilde{\theta}}$, respectively.) Lemma 3.3 (iii) implies the identification $P_{VZ}\bar{\phi}_{\theta_0}^{\intercal} = P_{VX}\phi_{\theta_0}^{\intercal} = 0_K$, as $P_{VX}\phi_{\theta_0}^{\intercal} = 0_K$ by the identifiability conditions of Lemma 3.6; and also $P_{VZ}\mathrm{D}_\theta\bar{\phi}_{\tilde{\theta}} = P_{VZ}Q_{\mathcal{X}}^{-1}\dot{\phi}_{\tilde{\theta}} = P_{VX}\dot{\phi}_{\tilde{\theta}}$. By this, and the strengthening of $\|\cdot\|_{L_p(P_{VX})}$ to $\|\cdot\|_{L_p(P_{VZ})}$ in (3.74),(3.72), and (3.73), all assumptions in Lemma 3.6 that hold under $(\mathbb{P}_n', P_{VX})$ continue to hold under $(\bar{\mathbb{P}}_n', P_{VZ})$. Thus, the assertion follows with $\dot{\Phi}$ being the same as in Lemma 3.6. ■

Proof of Lemma 3.7. By the definition of r_γ,

$$P_{VX}f(V, X, h, \gamma) = P_{VX}r_\gamma(V_1, X)h(V_1, X) \text{ for all } h \in L_2(P_{V_1X}).$$

Then

$$P_{VX}\Upsilon_{\gamma,h}(V, X) = P_{VX}\left[h^2 - 2f\right] = P_{VX}\left[h^2 - 2hr_\gamma\right] = P_{VX}\left[(h - r_\gamma)^2 - r_\gamma^2\right],$$

whereby $\arg\min_{h\in L_2(P_{V_1X})} P_{VX}\Upsilon_{\gamma,h} = \arg\min_{h\in L_2(P_{V_1X})} P_{VX}\left[(h - r_\gamma)^2\right] = r_\gamma$. ■

Proof of Corollary 3.5. The asymptotic normality of $\hat{\theta}_{\gamma_0}$ follows directly from Lemma 3.6 via the identification (3.79), whereby $P_{VX}\phi_{\gamma,\theta_0} = 0_K$ for any fixed $\gamma \in \Gamma$. A mean-value expansion of $r_{\hat{\gamma}_\mathcal{V}(c),\hat{\theta}_{\gamma_0}}$ via (i) in combination with (3.80) and (3.81) yields the second assertion as $\hat{\gamma}_\mathcal{V}(c) - \gamma_\mathcal{V}(c) = O_{P_{VX}}\left(n^{-1/2}\right)$, and $\hat{\theta}_{\gamma_0} - \theta_0 = O_{P_{VX}}\left(n^{-1/2}\right)$ by the first assertion. ■

Proof of Corollary 3.6. Follows from Proposition 3.3 as Corollary 3.5 follows from Lemma 3.6. ■

Conclusion

We contributed to the application and theoretical development of methods in causal inference and privacy. In addition to being of theoretical interest, these results also speak to recent, socially relevant issues, and, as such, they are of practical relevance.

Causal research questions are ubiquitous in various scientific fields studying practical questions. To name but a few, the assessment of medical treatments, vocational training programs, or of the effect of PhD studies on mental well-being are all potential applications of causal inference methods. Our theoretical contributions mainly apply to settings where experiments are unavailable, thus, inference must rely on observations of the uninterfered-with behaviour of individuals (observational data). For instance, it might be considered ethically questionable to randomise individuals within an experiment to undertake PhD studies or not; instead, it is merely observed whether they do so or not. In these cases, one must adjust for systematic differences between individuals to avoid bias stemming from self-selection into treatment. For instance, individuals opting for PhD studies may in the first place be in a more (or less) favourable mental health state than those who do not opt for PhD. In Chapter 1, we studied caliper matching, a method capable of such bias adjustment. The main appeal of caliper matching is its simplicity: the quality of adjustment is explicitly controlled via the caliper, which makes it an intuitive method for practitioners. By deriving its theoretical properties, we made it suitable for applications which also require formal uncertainty guarantees, such as the construction of confidence intervals.

Our contribution to causal inference in Chapter 2 is of applied nature. Therein, we compared the efficacy of two hip-fracture treatments combining data from experimental and observational sources, tackling missing-data issues. Unfortunately, our results are not of immediate practical relevance, because no treatment is found decisively better than the other. Nonetheless, our work could potentially be used as a starting point for other applied studies wishing to combine multiple data sources for causal inference and facing similar issues.

Privacy, similarly to causal inference, is a pressing practical issue. In an era with data being collected about numerous aspects of our life, it is only desirable that inference be performed in a privacy-preserving manner. Our theoretical contributions to privacy-preserving inference in Chapter 3 are in this vein. They add to the literature by proposing a simple and intuitive method to ensure individual-level privacy, which enables empirical applications for various data types. It was also shown how this method can be used to infer a rich class of practically relevant parameters amenable to flexible (double robust) modelling, including but not limited to causal effects in the other two chapters. We also established a direct connection between 'traditional', privacy-agnostic inference and privacy-preserving inference. This connection is advantageous in prectice, enabling the translation of privacy-agnostic methods into privacy-preserving ones.

Bibliography

Alberto Abadie and Guido W. Imbens. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica : Journal of the Econometric Society*, 74(1):235–267, January 2006. ISSN 0012-9682, 1468-0262. doi: 10.1111/j.1468-0262.2006.00655.x. URL http://doi.wiley.com/10.1111/j.1468-0262.2006.00655.x.

Alberto Abadie and Guido W. Imbens. Matching on the Estimated Propensity Score. *Econometrica : Journal of the Econometric Society*, 84(2):781–807, 2016. ISSN 0012-9682. doi: 10.3982/ECTA11293. URL https://www.econometricsociety.org/doi/10.3982/ECTA11293.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard Response: Estimating Distributions Privately, Efficiently, and With Little Communication. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1120–1129. PMLR, 2019. URL https://proceedings.mlr.press/v89/acharya19a.html.

Anish Agarwal and Rahul Singh. Causal Inference With Corrupted Data: Measurement Error, Missing Values, Discretization, and Differential Privacy, February 2024. URL http://arxiv.org/abs/2107.02780.

Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan. Differentially Private Simple Linear Regression, July 2020. URL http://arxiv.org/abs/2007.05157.

Kenneth S. Alexander. Rates of Growth and Sample Moduli for Weighted Empirical Processes Indexed by Sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

Hilal Asi and John C. Duchi. Near Instance-Optimality in Differential Privacy, May 2020. URL http://arxiv.org/abs/2005.10630.

Hilal Asi, John C. Duchi, and Omid Javidbakht. Element Level Differential Privacy: The Right Granularity of Privacy, December 2019. URL http://arxiv.org/abs/1912.04042.

Narayanaswamy Balakrishnan and Shanti S. Gupta. 2 Higher Order Moments of Order Statistics from Exponential and Right-Truncated Exponential Distributions and Applications to Life-Testing Problems. In *Order Statistics: Applications*, volume 17 of *Handbook of Statistics*, pages 25–59. Elsevier, 1998. doi: 10.1016/S0169-7161(98)17004-9. URL https://www.sciencedirect.com/science/article/pii/S0169716198170049.

Heejung Bang and James M. Robins. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–973, December 2005. ISSN 0006-341X, 1541-0420. doi: 10.1111/j.1541-0420.2005.00377.x. URL https://academic.oup.com/biometrics/article/61/4/962-973/7296220.

Richard Kwasi Bannor, Gupta Amarnath Krishna Kumar, Helena Oppong-Kyeremeh, and Camillus Abawiera Wongnaa. Adoption and Impact of Modern Rice Varieties on Poverty in Eastern India. *Rice Science*, 27(1): 56–66, January 2020. ISSN 16726308. doi: 10.1016/j.rsci.2019.12.006. URL https://linkinghub.elsevier.com/retrieve/pii/S1672630819301106.

Rina Foygel Barber and John C. Duchi. Privacy and Statistical Risk: Formalisms and Minimax Bounds, December 2014. URL http://arxiv.org/abs/1412.4451.

Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür. Fisher Information Under Local Differential Privacy. *IEEE Journal on Selected Areas in Information Theory*, 1(3):645–659, 2020. doi: 10.1109/JSAIT.2020.3039461.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Differentially Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds, October 2014. URL http://arxiv.org/abs/1405.7085.

Erich Battistin and Andrew Chesher. Treatment Effect Estimation With Covariate Measurement Error. *Journal of Econometrics*, 178(2):707–715, February 2014. ISSN 03044076. doi: 10.1016/j.jeconom.2013.10.010. URL https://linkinghub.elsevier.com/retrieve/pii/S030440761300225X.

Thomas B. Berrett, László Györfi, and Harro Walk. Strongly Universally Consistent Nonparametric Regression and Classification With Privatised Data. *Electronic Journal of Statistics*, 15(1), January 2021. ISSN 1935-7524. doi: 10.1214/21-EJS1845.

Herman J. Bierens. *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*. Cambridge University Press, 1st edition, July 1994. ISBN 978-0-521-56511-0 978-0-521-41900-0 978-0-511-59927-9. doi: 10.1017/CBO9780511599279. URL https://www.cambridge.org/core/product/identifier/9780511599279/type/book.

Patrick Billingsley. *Probability and Measure*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 3rd edition, 1995. ISBN 978-0-471-00710-4.

Erwin Bolthausen, Aad W. van der Vaart, and Edwin Perkins. Semiparametric Statistics. In Pierre Bernard, Jean-Michel Morel, Floris Takens, and Bernard Teissier, editors, *Lectures on Probability Theory and Statistics*, volume 1781 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002. ISBN 978-3-540-43736-9 978-3-540-47944-4. doi: 10.1007/b93152. URL http://link.springer.com/10.1007/b93152.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Routledge, 1st edition, October 2017. ISBN 978-1-315-13947-0. doi: 10.1201/9781315139470. URL https://www.taylorfrancis.com/books/9781351460491.

Elenka Brenna. Should I Care for My Mum or for My Kid? Sandwich Generation and Depression Burden in Italy. *Health policy (Amsterdam, Netherlands)*, 125(3):415–423, March 2021. ISSN 01688510. doi: 10.1016/j.healthpol.2020.11.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0168851020302979.

Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private Hypothesis Selection, January 2021. URL http://arxiv.org/abs/1905.13229.

Kristen Capogrossi and Wen You. The Influence of School Nutrition Programs on the Weight of Low-Income Children: A Treatment Effect Analysis. *Health Economics*, 26(8):980–1000, 2017.

James R. Carpenter and Melanie Smuk. Missing Data: A Statistical Framework for Practice. *Biometrical Journal*, 63(5):915–947, June 2021. ISSN 0323-3847, 1521-4036. doi: 10.1002/bimj.202000196. URL https://onlinelibrary.wiley.com/doi/10.1002/bimj.202000196.

Kamalika Chaudhuri and Daniel J. Hsu. Convergence Rates for Differentially Private Statistical Estimation. *Proceedings of the International Conference on Machine Learning. International Conference on Machine Learning*, 2012:1327–1334, 2012.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially Private Empirical Risk Minimization, February 2011. URL http://arxiv.org/abs/0912.0071.

Shuxiao Chen, Bo Zhang, and Ting Ye. Minimax Rates and Adaptivity in Combining Experimental and Observational Data, September 2021. URL http://arxiv.org/abs/2109.10522.

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Automatic Debiased Machine Learning of Causal and Structural Effects. *Econometrica*, 90(3):967–1027, 2022. ISSN 0012-9682. doi: 10.3982/ECTA18515. URL https://www.econometricsociety.org/doi/10.3982/ECTA18515.

Youngmin Cho. The Effects of Nonstandard Work Schedules on Workers' Health: A Mediating Role of Work-to-Family Conflict. *International Journal of Social Welfare*, 27(1):74–87, January 2018. ISSN 1369-6866, 1468-2397. doi: 10.1111/ijsw.12269. URL https://onlinelibrary.wiley.com/doi/10.1111/ijsw.12269.

William G. Cochran. Matching in Analytical Studies. *American Journal of Public Health and the Nations Health*, 43(6 Pt 1):684–691, June 1953. ISSN 0002-9572. doi: 10.2105/AJPH.43.6_Pt_1.684. URL https://ajph.aphapublications.org/doi/full/10.2105/AJPH.43.6_Pt_1.684.

William G. Cochran and Donald B. Rubin. Controlling Bias in Observational Studies: A Review. *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 35(4):417–446, 1973. ISSN 0581572X. URL http://www.jstor.org/stable/25049893.

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, January 2023. URL http://arxiv.org/abs/2011.08047.

Francisco Cribari-Neto, Nancy Lopes Garcia, and Klaus L. P. Vasconcellos. A Note on Inverse Moments of Binomial Variates. *Brazilian Review of Econometrics*, 20:269–277, 2000.

Lauren Eyler Dang, Jens Magelund Tarp, Trine Julie Abrahamsen, Kajsa Kvist, John B. Buse, Maya Petersen, and Mark J. van der Laan. A Cross-Validated Targeted Maximum Likelihood Estimator for Data-Adaptive Experiment Selection Applied to the Augmentation of RCT Control Arms With External Data, February 2023. URL http://arxiv.org/abs/2210.05802.

Rajeev H. Dehejia and Sadek Wahba. Propensity Score Matching Methods for Non-Experimental Causal Studies. Technical Report w6829, National Bureau of Economic Research, Cambridge, MA, December 1998. URL http://www.nber.org/papers/w6829.pdf.

Rajeev H. Dehejia and Sadek Wahba. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, 84(1):151–161, 2002. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/3211745.

Paul M. Densen, Paul Padget, Bruce Webster, Claude S. Nicol, and Clayton Rich. Studies in Cardiovascular Syphilis. II. Methodologic Problems in the Evaluation of Therapy. *American Journal of Syphilis, Gonorrhea, and Venereal Diseases*, 36(1):64–76, January 1952. ISSN 0096-6738.

Damien Desfontaines and Balázs Pejó. SoK: Differential Privacies, November 2022. URL http://arxiv.org/abs/1906.01337.

Joerg Drechsler, Ira Globus-Harris, Audra McMillan, Jayshree Sarathy, and Adam Smith. Non-Parametric Differentially Private Confidence Intervals for the Median, July 2021. URL http://arxiv.org/abs/2106.10333.

John C. Duchi and Feng Ruan. The Right Complexity Measure in Locally Private Estimation: It Is Not the Fisher Information. *The Annals of Statistics*, 52(1), February 2024. ISSN 0090-5364. doi: 10.1214/22-AOS2227. URL https://projecteuclid.org/journals/annals-of-statistics/volume-52/issue-1/The-right-complexity-measure-in-locally-private-estimation--It/10.1214/22-AOS2227.full.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax Optimal Procedures for Locally Private Estimation. *Journal of the American Statistical Association*, 113(521):182–201, January 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1389735. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1389735.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Shai Halevi, and Tal Rabin, editors, *Theory of Cryptography*, volume 3876, pages 265–284. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-32731-8 978-3-540-32732-5. doi: 10.1007/11681878_14. URL http://link.springer.com/10.1007/11681878_14.

Eva Christine Erhardt. Microfinance Beyond Self-Employment: Evidence for Firms in Bulgaria. *Labour Economics*, 47:75–95, August 2017. ISSN 09275371. doi: 10.1016/j.labeco.2017.04.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0927537116302561.

Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting Privacy Breaches in Privacy Preserving Data Mining. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 211–222, San Diego California, June 2003. ACM. ISBN 978-1-58113-670-8. doi: 10.1145/773153.773174. URL https://dl.acm.org/doi/10.1145/773153.773174.

Jerome H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1), March 1991. ISSN 0090-5364. doi: 10.1214/aos/1176347963. URL https://projecteuclid.org/journals/annals-of-statistics/volume-19/issue-1/Multivariate-Adaptive-Regression-Splines/10.1214/aos/1176347963.full.

Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5), October 2001. ISSN 0090-5364. doi: 10.1214/aos/1013203451. URL https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full.

Jerome H. Friedman. Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, 38(4): 367–378, February 2002. ISSN 01679473. doi: 10.1016/S0167-9473(01)00065-2. URL https://linkinghub.elsevier.com/retrieve/pii/S0167947301000652.

Markus Frölich. Matching Estimators and Optimal Bandwidth Choice. *Statistics and Computing*, 15(3): 197–215, July 2005. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-005-1309-6. URL http://link.springer.com/10.1007/s11222-005-1309-6.

Kazuto Fukuchi, Quang Khai Tran, and Jun Sakuma. Differentially Private Empirical Risk Minimization With Input Perturbation. In Akihiro Yamamoto, Takuya Kida, Takeaki Uno, and Tetsuji Kuboyama, editors, *Discovery Science*, volume 10558, pages 82–90. Springer International Publishing, Cham, 2017. ISBN 978-3-319-67785-9 978-3-319-67786-6. doi: 10.1007/978-3-319-67786-6_6. URL http://link.springer.com/10.1007/978-3-319-67786-6_6.

Craig Gentry. *A Fully Homomorphic Encryption Scheme*. PhD thesis, Stanford University, 2009. URL https://crypto.stanford.edu/craig.

Noah Golowich. Differentially Private Nonparametric Regression Under a Growth Condition, November 2021. URL http://arxiv.org/abs/2111.12786.

Jinyong Hahn. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, 66(2):315, March 1998. ISSN 00129682. doi: 10.2307/2998560. URL https://www.jstor.org/stable/2998560?origin=crossref.

Lars Peter Hansen. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029, July 1982. ISSN 00129682. doi: 10.2307/1912775. URL https://www.jstor.org/stable/1912775?origin=crossref.

Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3), August 1986. ISSN 0883-4237. doi: 10.1214/ss/1177013604. URL https://projecteuclid.org/journals/statistical-science/volume-1/issue-3/Generalized-Additive-Models/10.1214/ss/1177013604.full.

James J. Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an Econometric Evaluation Estimator. *The Review of Economic Studies*, 65(2):261–294, 1998. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/2566973.

Jonas Heese, Mozaffar Khan, and Karthik Ramanna. Is the SEC Captured? Evidence from Comment-Letter Reviews. *Journal of Accounting and Economics*, 64(1):98–122, August 2017. ISSN 01654101. doi: 10.1016/j.jacceco.2017.06.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0165410117300344.

Martin Huber, Michael Lechner, and Andreas Steinmayr. Radius Matching on the Propensity Score With Bias Adjustment: Tuning Parameters and Finite Sample Behaviour. *Empirical Economics*, 49(1):1–31, August 2015a. ISSN 0377-7332, 1435-8921. doi: 10.1007/s00181-014-0847-1. URL http://link.springer.com/10.1007/s00181-014-0847-1.

Martin Huber, Michael Lechner, and Conny Wunsch. Workplace Health Promotion and Labour Market Performance of Employees. *Journal of Health Economics*, 43:170–189, September 2015b. ISSN 01676296. doi: 10.1016/j.jhealeco.2015.07.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0167629615000776.

U. Hucke. Local Differential Privacy and Estimation in the Binomial Model. Master's thesis, University of Freiburg, 2019.

Jonathan Izudi, Imelda K. Tamwesigire, and Francis Bajunirwe. Does Completion of Sputum Smear Monitoring Have an Effect on Treatment Success and Cure Rate Among Adult Tuberculosis Patients in Rural Eastern Uganda? A Propensity Score-Matched Analysis. *PLOS ONE*, 14(12):e0226919, December 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0226919. URL https://dx.plos.org/10.1371/journal.pone.0226919.

Yangdi Jiang, Yi Liu, Xiaodong Yan, Anne-Sophie Charest, Linglong Kong, and Bei Jiang. Analysis of Differentially Private Synthetic Data: A Measurement Error Approach. In *AAAI Conference on Artificial Intelligence*, 2024. URL https://api.semanticscholar.org/CorpusID:268699082.

Olof Johnell and John A. Kanis. An Estimate of the Worldwide Prevalence, Mortality and Disability Associated With Hip Fracture. *Osteoporosis International*, 15(11):897–902, November 2004. ISSN 0937-941X, 1433-2965. doi: 10.1007/s00198-004-1627-0. URL http://link.springer.com/10.1007/s00198-004-1627-0.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal Mechanisms for Local Differential Privacy, November 2015. URL http://arxiv.org/abs/1407.1338.

Gautam Kamath, Vikrant Singhal, and Jonathan R. Ullman. Private Mean Estimation of Heavy-Tailed Distributions. *CoRR*, abs/2002.09464, 2020. URL https://arxiv.org/abs/2002.09464.

Vishesh Karwa and Salil Vadhan. Finite Sample Differentially Private Confidence Intervals, November 2017. URL http://arxiv.org/abs/1711.03908.

Shao-Hsun Keng and Sheng-Jang Sheu. The Effect of Stimulants and Their Combined Use With Cigarettes on Mortality: The Case of Betel Quid. *The European Journal of Health Economics*, 14(4):677–695, August 2013. ISSN 1618-7598, 1618-7601. doi: 10.1007/s10198-012-0415-6. URL http://link.springer.com/10.1007/s10198-012-0415-6.

Edward H. Kennedy. Semiparametric Doubly Robust Targeted Double Machine Learning: A Review, January 2023. URL http://arxiv.org/abs/2203.06469.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private Convex Empirical Risk Minimization and High-Dimensional Regression. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 25.1–25.40, Edinburgh, Scotland, June 2012. PMLR. URL https://proceedings.mlr.press/v23/kifer12.html.

Máté Kormos, Stéphanie L. van der Pas, and Aad W. van der Vaart. Asymptotics of Caliper Matching Estimators for Average Treatment Effects, April 2023. URL http://arxiv.org/abs/2304.08373.

Rainer Kress. *Linear Integral Equations*. Number volume 82 in Applied Mathematical Sciences. Springer, New York, 3rd edition, 2014. ISBN 978-1-4614-9592-5.

Yuvaraj Krishnamoorthy and Tanveer Rehman. Impact of Antenatal Care Visits on Childhood Immunization: A Propensity Score-Matched Analysis Using Nationally Representative Survey. *Family Practice*, 39(4):603–609, July 2022. ISSN 1460-2229. doi: 10.1093/fampra/cmab124. URL https://academic.oup.com/fampra/article/39/4/603/6375553.

Ying-Ying Lee. Efficient Propensity Score Regression Estimators of Multivalued Treatment Effects for the Treated. *Journal of Econometrics*, 204(2):207–222, June 2018. ISSN 03044076. doi: 10.1016/j.jeconom.2018.02.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0304407618300290.

Jing Lei. Differentially Private M-Estimators. In *NIPS*, 2011.

Jiachun Li, Kaining Shi, and David Simchi-Levi. Privacy Preserving Adaptive Experiment Design. In *International Conference on Machine Learning*, 2024. URL https://api.semanticscholar.org/CorpusID:267028376.

Xi Lin, Jens Magelund Tarp, and Robin J. Evans. Data Fusion for Efficiency Gain in ATE Estimation: A Practical Review With Simulations, July 2024. URL http://arxiv.org/abs/2407.01186.

Po-Ling Loh and Martin J. Wainwright. High-Dimensional Regression With Noisy and Missing Data: Provable Guarantees With Nonconvexity. *The Annals of Statistics*, 40(3), June 2012. ISSN 0090-5364. doi: 10.1214/12-AOS1018. URL http://arxiv.org/abs/1109.3714.

Yi Lu, Daniel O. Scharfstein, Maria M. Brooks, Kevin Quach, and Edward H. Kennedy. Causal Inference for Comprehensive Cohort Studies, October 2019. URL http://arxiv.org/abs/1910.03531.

Paul Mangold, Aurélien Bellet, Joseph Salmon, and Marc Tommasi. High-Dimensional Private Empirical Risk Minimization by Greedy Coordinate Descent. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 4894–4916. PMLR, 2023. URL https://proceedings.mlr.press/v206/mangold23a.html.

John A. Nelder and Robert W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370, 1972. ISSN 00359238. doi: 10.2307/2344614. URL https://www.jstor.org/stable/10.2307/2344614?origin=crossref.

Whitney K. Newey and Daniel McFadden. Chapter 36: Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994. ISBN 978-0-444-88766-5. doi: 10.1016/S1573-4412(05)80005-4. URL https://linkinghub.elsevier.com/retrieve/pii/S1573441205800054.

Jerzy Neyman. *On the Applications of the Theory of Probability to Agricultural Experiments*. PhD thesis, University of Warsaw, 1924.

Fengshi Niu, Harsha Nori, Brian Quistorff, Rich Caruana, Donald Ngwe, and Aadharsh Kannan. Differentially Private Estimation of Heterogeneous Causal Effects. In *CLEaR*, 2022. URL https://api.semanticscholar.org/CorpusID:247025804.

Michael Oberst, Alexander D'Amour, Minmin Chen, Yuyan Wang, David Sontag, and Steve Yadlowsky. Understanding the Risks and Rewards of Combining Unbiased and Possibly Biased Estimators, With Applications to Causal Inference, May 2023. URL http://arxiv.org/abs/2205.10467.

Yuki Ohnishi and Jordan Awan. Locally Private Causal Inference for Randomized Experiments, 2023. URL https://arxiv.org/abs/2301.01616.

Anouk Patel-Campillo and V.B. Salas García. Breaking the Poverty Cycle? Conditional Cash Transfers and Higher Education Attainment. *International Journal of Educational Development*, 92:102612, July 2022. ISSN 07380593. doi: 10.1016/j.ijedudev.2022.102612. URL https://linkinghub.elsevier.com/retrieve/pii/S0738059322000621.

Robert Piziak and Patrick L. Odell. *Matrix Theory*. Chapman and Hall/CRC, 1st edition, February 2007. ISBN 978-1-4200-0993-4. doi: 10.1201/9781420009934. URL https://www.taylorfrancis.com/books/9781420009934.

Andrej Dmitrievič Polânin and Aleksandr Vladimirovich Manzhirov. *Handbook of Integral Equations*. CRC press, Boca Raton, London, New York, Washington, 1st edition, 1998. ISBN 978-0-8493-2876-3.

Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics*. Academic Press, New York, 1972. ISBN 978-0-12-585001-8.

Herbert Robbins. A Remark on Stirling's Formula. *The American Mathematical Monthly*, 62(1):26, January 1955. ISSN 00029890. doi: 10.2307/2308012. URL http://www.jstor.org/stable/2308012?origin=crossref.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, September 1994. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1994.10476818. URL https://www.tandfonline.com/doi/full/10.1080/01621459.1994.10476818.

Paul R. Rosenbaum and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/70.1.41. URL https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/70.1.41.

Paul R. Rosenbaum and Donald B. Rubin. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, 39(1):33–38, February 1985. ISSN 0003-1305, 1537-2731. doi: 10.1080/00031305.1985.10479383. URL http://www.tandfonline.com/doi/abs/10.1080/00031305.1985.10479383.

Andrea Rotnitzky, James M. Robins, and Daniel O. Scharfstein. Semiparametric Regression for Repeated Outcomes With Nonignorable Nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339, December 1998. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1998.10473795. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10473795.

Andrea Rotnitzky, Ezequiel Smucler, and James M. Robins. Characterization of Parameters With a Mixed Bias Property. *Biometrika*, 108(1):231–238, March 2021. ISSN 0006-3444, 1464-3510. doi: 10.1093/biomet/asaa054. URL https://academic.oup.com/biomet/article/108/1/231/5899828.

Donald B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. ISSN 0022-0663. doi: 10.1037/h0037350. URL http://content.apa.org/journals/edu/66/5/688.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics. Wiley, 1st edition, June 1987. ISBN 978-0-471-08705-2 978-0-470-31669-6. doi: 10.1002/9780470316696. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316696.

Donald B. Rubin. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434):473–489, June 1996. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1996.10476908. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476908.

Donald B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511810725.

Luca Salmasi and Luca Pieroni. Immigration Policy and Birth Weight: Positive Externalities in Italian Law. *Journal of Health Economics*, 43:128–139, September 2015. ISSN 01676296. doi: 10.1016/j.jhealeco.2015.06.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0167629615000752.

Eugene F. Schuster and Sidney J. Yakowitz. Contributions to the Theory of Nonparametric Regression, With Application to System Identification. *The Annals of Statistics*, 7(1):139–149, 1979. ISSN 00905364. URL http://www.jstor.org/stable/2958838.

Ramalingam Shanmugam and Barry C. Arnold. Characterizations Based on Conditional Distributions Given the Minimum Value in the Sample. *Sankhy: The Indian Journal of Statistics, Series A (1961-2002)*, 50(3):452–459, 1988. ISSN 0581572X. URL http://www.jstor.org/stable/25050713.

Or Sheffet. Differentially Private Ordinary Least Squares. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3105–3114. PMLR, August 2017. URL https://proceedings.mlr.press/v70/sheffet17a.html.

Chung-Hua Shen and Yuan Chang. Ambition Versus Conscience, Does Corporate Social Responsibility Pay off? The Application of Matching Methods. *Journal of Business Ethics*, 88(S1):133–153, April 2009. ISSN 0167-4544, 1573-0697. doi: 10.1007/s10551-008-9826-9. URL http://link.springer.com/10.1007/s10551-008-9826-9.

Galen R. Shorack and Jon A. Wellner. *Empirical Processes With Applications to Statistics*. Number 59 in Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 2009. ISBN 978-0-89871-684-9.

Aleksandra B. Slavkovic and Roberto Molinari. Perturbed M-Estimation: A Further Investigation of Robust Statistics for Differential Privacy, 2021. URL https://api.semanticscholar.org/CorpusID:237194827.

Adam Smith. Efficient, Differentially Private Point Estimators, September 2008. URL http://arxiv.org/abs/0809.4794.

Ezequiel Smucler, Andrea Rotnitzky, and James M. Robins. A Unifying Approach for Doubly-Robust ℓ_1 Regularized Estimation of Causal Contrasts, June 2019. URL http://arxiv.org/abs/1904.03737.

Lukas Steinberger. Efficiency in Local Differential Privacy, January 2023. URL http://arxiv.org/abs/2301.10600.

Jonathan A. C. Sterne, Ian R. White, John B. Carlin, Michael Spratt, Patrick Royston, Michael G. Kenward, Angela M. Wood, and James R. Carpenter. Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. *BMJ*, 338(jun29 1):b2393–b2393, September 2009. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.b2393. URL https://www.bmj.com/lookup/doi/10.1136/bmj.b2393.

Elizabeth A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1 – 21, 2010. doi: 10.1214/09-STS313. URL https://doi.org/10.1214/09-STS313.

Maria C. J. M. Tol, Nienke W. Willigenburg, Hanna C. Willems, Taco Gosens, Ariena Rasker, Martin J. Heetveld, Martijn G. M. Schotanus, Johanna M. van Dongen, Bart Eggen, Máté Kormos, Stéphanie L. van der Pas, Aad W. van der Vaart, Rudolf W. Poolman, and APOLLO Research Group. Posterolateral or Direct Lateral Approach for Cemented Hemiarthroplasty After Femoral Neck Fracture (APOLLO): Protocol for a Multicenter Randomized Controlled Trial With Economic Evaluation and Natural Experiment Alongside. *Acta Orthopaedica*, 93:732–738, September 2022. ISSN 1745-3682, 1745-3674. doi: 10.2340/17453674.2022.4547. URL https://actaorthop.org/actao/article/view/4547.

Maria C. J. M. Tol, Nienke W. Willigenburg, Ariena J. Rasker, Hanna C. Willems, Taco Gosens, Martin J. Heetveld, Martijn G. M. Schotanus, Bart Eggen, Máté Kormos, Stéphanie L. van der Pas, Aad W. van der Vaart, J. Carel Goslings, Rudolf W. Poolman, APOLLO Research Group, Frank van Roon, Martijn van Dijk, Jort Keizer, Anne J.H. Vochteloo, Pieter Joosse, Bert Boonen, Jetse Jelsma, Dieuwertje Theeuwen, Joris J.W. Ploegmakers, Tim Schepers, Evelien van der Meij, Svenhjalmar H. van Helden, Rutger Zuurmond, Bart A. van Dijkman, Thomas D. Berendes, and Hans G.E. Hendriks. Posterolateral or Direct Lateral Surgical Approach for Hemiarthroplasty After a Hip Fracture: A Randomized Clinical Trial Alongside a Natural Experiment. *JAMA Network Open*, 7(1):e2350765, January 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2023.50765. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2813843.

Stef van Buuren and Karin Groothuis-Oudshoorn. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03. URL http://www.jstatsoft.org/v45/i03/.

Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 2007.

Max P.L. van der Sijp, Danny van Delft, Pieta Krijnen, Arthur H.P. Niggebrugge, and Inger B. Schipper. Surgical Approaches and Hemiarthroplasty Outcomes for Femoral Neck Fractures: A Meta-Analysis. *The Journal of Arthroplasty*, 33(5):1617–1627.e9, May 2018. ISSN 08835403. doi: 10.1016/j.arth.2017.12.029. URL https://linkinghub.elsevier.com/retrieve/pii/S088354031731135X.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 1-107-26372-7.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, New York, NY, 1996. ISBN 978-1-4757-2547-6 978-1-4757-2545-2. doi: 10.1007/978-1-4757-2545-2. URL http://link.springer.com/10.1007/978-1-4757-2545-2.

Nerina Vecchio, Debbie Davies, and Nicholas Rohde. The Effect of Inadequate Access to Healthcare Services on Emergency Room Visits. A Comparison Between Physical and Mental Health Conditions. *PLOS ONE*, 13(8):1–15, August 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0202559. URL https://dx.plos.org/10.1371/journal.pone.0202559.

Wei Wang, Lili Lu, Mohammedhamid Mohammedosman Kelifa, Yan Yu, Anqi He, Na Cao, Si Zheng, Wenjun Yan, and Yinmei Yang. Mental Health Problems in Chinese Healthcare Workers Exposed to Workplace Violence During the COVID-19 Outbreak: A Cross-Sectional Study Using Propensity Score Matching Analysis. *Risk Management and Healthcare Policy*, Volume 13:2827–2833, December 2020. ISSN 1179-1594. doi: 10.2147/RMHP.S279170.

Stanley L. Warner. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309):63–69, March 1965. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1965.10480775. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1965.10480775.

Ian R. White, Patrick Royston, and Angela M. Wood. Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, 30(4):377–399, February 2011. ISSN 0277-6715, 1097-0258. doi: 10.1002/sim.4067. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.4067.

Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic Integrative Analysis of Randomized Trial and Real-World Data for Treatment Heterogeneity Estimation, November 2022. URL http://arxiv.org/abs/2005.10579.

Yang Yang, Xindi Huang, Ximeng Liu, Hongju Cheng, Jian Weng, Xiangyang Luo, and Victor Chang. A Comprehensive Survey on Secure Outsourced Computation and Its Applications. *IEEE Access*, 7:159426–159465, October 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2949782.

Yuchen Zhu, Limor Gultchin, Arthur Gretton, Matt Kusner, and Ricardo Silva. Causal Inference With Treatment Measurement Error: A Nonparametric Instrumental Variable Approach, June 2022. URL http://arxiv.org/abs/2206.09186.

Hui Zou and Trevor Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, April 2005. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. URL https://academic.oup.com/jrsssb/article/67/2/301/7109482.

Curriculum Vitae

Kormos Máté[5], born in Budapest, Hungary, on 3 May 1994, completed his secondary education in 2013 at Toldy Ferenc Gimnázium, Budapest, Hungary, and obtained a title of Master of Arts in Economics in 2018 at Central European University, Budapest, Hungary, with thesis *Paired* 2×2 *Factorial Design for Treatment Effect Identification and Estimation in the Presence of Paired Interference and Noncompliance* supervised by Mátyás László and Lieli Robert. He also obtained a title of Master of Science in Econometrics in 2020 at Vrije Universiteit Amsterdam (Tinbergen Institute) with thesis *Kernel Density Estimation on Homomorphically Encrypted Data* supervised by Charles Bos. He conducted his doctoral research at Universiteit Leiden (2020–2021) and Technische Universiteit Delft (2021–2024) under the supervision of Aad van der Vaart and Stéphanie van der Pas, while holding a guest position at Amsterdam Universitair Medische Centra (2021–2024).

[5]Name written according to Hungarian convention: surname first.

Publications

- Máté Kormos, Stéphanie L. van der Pas, and Aad W. van der Vaart. Asymptotics of Caliper Matching Estimators for Average Treatment Effects, April 2023. URL http://arxiv.org/abs/2304.08373
- Maria C. J. M. Tol, Nienke W. Willigenburg, Hanna C. Willems, Taco Gosens, Ariena Rasker, Martin J. Heetveld, Martijn G. M. Schotanus, Johanna M. van Dongen, Bart Eggen, Máté Kormos, Stéphanie L. van der Pas, Aad W. van der Vaart, Rudolf W. Poolman, and APOLLO Research Group. Posterolateral or Direct Lateral Approach for Cemented Hemiarthroplasty After Femoral Neck Fracture (APOLLO): Protocol for a Multicenter Randomized Controlled Trial With Economic Evaluation and Natural Experiment Alongside. *Acta Orthopaedica*, 93:732–738, September 2022. ISSN 1745-3682, 1745-3674. doi: 10.2340/17453674.2022.4547. URL https://actaorthop.org/actao/article/view/4547
- Maria C. J. M. Tol, Nienke W. Willigenburg, Ariena J. Rasker, Hanna C. Willems, Taco Gosens, Martin J. Heetveld, Martijn G. M. Schotanus, Bart Eggen, Máté Kormos, Stéphanie L. van der Pas, Aad W. van der Vaart, J. Carel Goslings, Rudolf W. Poolman, APOLLO Research Group, Frank van Roon, Martijn van Dijk, Jort Keizer, Anne J.H. Vochteloo, Pieter Joosse, Bert Boonen, Jetse Jelsma, Dieuwertje Theeuwen, Joris J.W. Ploegmakers, Tim Schepers, Evelien van der Meij, Svenhjalmar H. van Helden, Rutger Zuurmond, Bart A. van Dijkman, Thomas D. Berendes, and Hans G.E. Hendriks. Posterolateral or Direct Lateral Surgical Approach for Hemiarthroplasty After a Hip Fracture: A Randomized Clinical Trial Alongside a Natural Experiment. *JAMA Network Open*, 7(1):e2350765, January 2024. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2023.50765. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2813843

Suppose that we wish to infer the effect of a medical treatment on the well-being of individuals in a population. This research aim falls into the realm of causal inference, which is the main underlying topic connecting the three chapters of this thesis. Suppose also that the possibility of conducting an experiment — arguably, the most credible method for causal inference — is limited, e.g. for monetary or ethical reasons. In this case, we have to rely on observational data — data not from an experiment — to infer the effect of the treatment. This is the topic of the first and second chapter: the first chapter studies the case when experiments are completely unavailable, while the second chapter combines experimental and observational data. In addition, suppose that we also wish to protect the privacy of individuals. How to perform (causal) inference in a privacy-preserving manner is the content of the third chapter.