Designing for Stability in Causal Inference



Linying Yang

Jesus College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

June 2026

Acknowledgements

I would like to thank my supervisor, Professor Robin Evans. Words cannot express how much Robin has meant to me. I will never forget the moments when he looked into my tearful eyes and helped me regain confidence, when he patiently rehearsed my first talk, or when he paused his meetings to comfort me through my periods of severe depression. During some of the darkest periods of my life, his kindness felt like dawn breaking over a cold and lonely sea. From Robin, I learned as much from his kindness as from his wisdom and generosity. I am fortunate to be his student, and his office will always remain my favourite place in Oxford. If there is any justice in the way goodness returns to the world, I hope life brings him every happiness it can.

I am also deeply grateful to Professor Frank Windmeijer, who has guided me throughout my DPhil. The curiosity and enjoyment he brings to research remind me why I wanted to do it in the first place. Whenever I feel lost or anxious, Frank helps me find my footing again. There is a steadiness and care in him that helps me find my own. Throughout my DPhil, he has been a deeply rooted presence in my life, someone I can always rely on and look up to.

My thanks also go to Professor Nicolai Meinshausen. I have always admired his work, but what I cherish the most is his kindness. His hugs and encouragement have meant far more to me than he may ever know. Every encounter with him feels like a gentle spring breeze.

I am grateful to Professor François Caron. Beyond everything I learned from him academically, I will always be grateful for his warmth, encouragement, and genuine care. I owe my gratitude to Professor Peter Bühlmann, whom I was fortunate enough to collaborate with. His creativity opened many doors in my thinking. Just as importantly, the support and care he shows to young researchers taught me what academic leadership can be at its best.

I am especially thankful to Christl, Maria, and Christina. As a young woman hoping to work in academia, I often drew strength from your presence. You reminded me that empathy is a strength, and that there is much to be proud of in being a woman researcher. To Beverley, Emma, and Frédérique in Oxford, and to Susanne at ETH: thank you for your kindness and support, which made this journey far smoother than it otherwise would have been.

The greatest gift from Oxford was the people who walked alongside me. Dan, thank you for being such a warm and steady companion through these years. Thank you for believing in

me when I could not believe in myself, and for being there through every tear and every laugh. I am grateful to have found a lifelong friend in you. Robin, we have been through difficult times together, and I am grateful that we always found warmth in each other. Thank you for all the growth we shared. Stefano, thank you for your thoughtful and gentle friendship. For all your care and companionship, I am deeply grateful. The world always feels a little softer in your company. Thank you, Chris, for your friendship. You taught me to walk in the park and think. I will miss your coffee, and the joy of research we have had together. Xing, thank you for being such a wonderful collaborator and friend. Our discussions were some of the most enjoyable moments of my DPhil. Arya, I am grateful for your care, the vulnerability we shared, and the early morning calls from Canada. Thank you, Aavash, for all the walks in the park and conversations. Cathal, you always opened your door to me, especially on the sleepless nights when I needed somewhere to be. Thank you Sahra and Valentin for every hug you gave me. There are many more people I wish I could mention individually, including Johannes, Jihoon, Kevin, Jakob, Vanessa, Karla, Hera, Dan, Yasin, Giulia, Alex, Yanzhao, Faaiz, Jake, Tyler, Shirley, Shahine, Mike, Cyrill, Felix, Linus, Maybritt, Morten, Jens, Zhixiao and Pablo. Knowing you has been an unexpected gift of these years.

A special thank you to Xinwei. We have been friends for more than ten years, and I am grateful that our paths have crossed again. I am lucky to have you in my life.

My heartfelt thanks also go to my friends who have become family, in the UK and the USA. To Wenkai, Jin, Lihao, Joe, Jianjie, Xin, Maggie, Yishuang and Edward: thank you for caring for me like a younger sister and for always being there when I needed support.

Finally, I would like to thank my family. Mom, thank you for your unwavering support. I would not have made it this far without you. To my grandparents, thank you for your endless love. And thank you, Eureka, for being the sweetest cat in the world and for accompanying me through countless days and nights.

My years at Oxford contained both darkness and light. Throughout these years, I hoped to keep a transparent heart and eyes that could still be moved to tears. In the people mentioned in these pages, I found a generosity and sincerity greater than I could have hoped for. Because of them, I have been able to keep parts of myself that I once feared I might lose: the softness to be moved, the courage to hope, and the faith in goodness.

To all of you, thank you. I have been far luckier than I deserve. The light I carry forward from these years has been kindled, again and again, by you.

Abstract

Causal and predictive conclusions are more reliable when they are stable under perturbations of the data generation and analysis procedure, or when their sensitivity to perturbations can be quantified. This thesis studies how to design for stability: it develops procedures that are stable by construction and tools that evaluate how stable a conclusion is.

An augmented marginal outcome density ratio estimator is developed for average treatment effects under weak overlap, where inverse propensity score weighting becomes highly variable. By incorporating outcome information, the proposed estimator can achieve lower variance than state-of-the-art methods in certain settings.

Augmenting a primary study with external data can strengthen causal conclusions, but it can also undermine them when incompatibility is not accounted for. A test-then-pool procedure, built on a maximum mean discrepancy equivalence test, is developed to control the Type-I error of the subsequent causal test across a class of source heterogeneities, while recovering power when the sources are compatible.

Frengression is introduced as a flexible and faithful generative framework for causal estimation and simulation. Parameterising the target estimand separately from the observational structure, it ties estimation and simulation to the same target. It provides point estimates of causal quantities and synthetic data with known causal structure, and can be used to build benchmarks for evaluating the stability of causal methods in complex settings.

Predictive models are increasingly trained on one distribution and deployed on another, where their performance can deteriorate. The thesis further studies predictive robustness under distribution shift through causal invariance. A method is proposed that generates synthetic environments from a single dataset, rather than relying on heterogeneous sources, to learn a prediction function and characterise the class of perturbations under which it remains stable.

This thesis shows that stability is not an absolute property of a statistical method. It is a guarantee relative to a specified class of perturbations. Designing for stability means making that class explicit, constructing procedures that respect it, and quantifying how conclusions change when stability is only partial.

Contents

Acknowledgements ii

Abstract iv

1 Introduction 1

2 Stability and causality 3
2.1 The language of causality . . . 3
2.1.1 Potential outcomes . . . 3
2.1.2 Structural causal models . . . 5
2.2 Stability for causal inference . . . 7
2.3 Causality for predictive robustness . . . 8
2.4 Sources of instability in causal inference . . . 10
2.4.1 Finite-sample fragility . . . 10
2.4.2 Structural complexity . . . 11
2.4.3 Distribution shift . . . 11
2.4.4 Threats to identification . . . 11
2.5 Notation . . . 12

3 Stable estimation via outcome-informed weighting 14
3.1 Introduction . . . 14
3.2 Background . . . 15
3.2.1 Problem set-up and notation . . . 15
3.2.2 Efficient estimators and efficiency bound of ATE . . . 16
3.2.3 Related work . . . 17
3.3 Methodology . . . 19
3.3.1 Marginal ratio and augmented marginal ratio estimators . . . 19
3.3.2 Outcome-informed weighting . . . 22
3.3.3 Connection with covariate adjustment . . . 23
3.3.4 Weights estimation . . . 24
3.4 Asymptotic properties . . . 25
3.4.1 Consistency . . . 26
3.4.2 Asymptotic distribution and inference . . . 27
3.4.3 Efficiency interpretation via the submodel . . . 30
3.5 Experiments . . . 31
3.5.1 Synthetic experiment results . . . 31
3.5.2 Application on NHANES data . . . 34
3.6 Discussion . . . 37

4 Stable data fusion via equivalence testing 38
4.1 Introduction . . . 38
4.2 Background . . . 40
4.2.1 Distributional treatment effect . . . 41
4.2.2 MMD and two-sample testing . . . 41
4.2.3 Existing mean-based TTP . . . 44
4.2.4 Warm-up: distributional TTP . . . 45
4.3 Method . . . 46
4.3.1 Fusion test: MMD equivalence testing . . . 47
4.3.2 Causality test: partial bootstrap and partial permutation . . . 49
4.3.2.1 Partial bootstrap . . . 50
4.3.2.2 Normal approximation . . . 52
4.3.2.3 Partial permutation . . . 53
4.3.2.4 Comparison between partial bootstrap and partial permutation . . . 54
4.3.3 The overall procedure . . . 57
4.3.4 Choosing the kernel and bandwidth . . . 60
4.3.5 Choosing the equivalence radius . . . 61
4.4 Synthetic experiment . . . 61
4.4.1 Equivalence TTP controls Type-I error and improves power . . . 62
4.4.2 Effects of varying ρ . . . 63
4.4.3 Comparison between partial bootstrap and partial permutation . . . 64
4.4.4 Comparison of varying kernel bandwidth . . . 65
4.4.5 Comparison with linear kernel . . . 65
4.5 Prospera program . . . 65
4.6 Discussion . . . 67

5 Stability evaluation via faithful causal simulation 69
5.1 Introduction . . . 69
5.1.1 Contribution . . . 70
5.1.2 Related work . . . 71
5.1.2.1 Deep generative models for the causal margin . . . 71
5.1.2.2 Modelling longitudinal data . . . 72
5.1.2.3 Modelling distributional treatment effects . . . 73
5.1.2.4 Generative models for causal data simulation . . . 73
5.1.3 Structure . . . 74
5.2 Method . . . 74
5.2.1 Notation . . . 74
5.2.2 Review of key concepts . . . 75
5.2.3 Model and parameterisation . . . 76
5.2.4 Estimation . . . 79
5.2.4.1 Finite-sample frengression . . . 81

5.2.5 Sampling . . . 83
5.2.6 Sampling-based inference . . . 83
5.3 Model consistency . . . 84
5.4 Extrapolation . . . 87
5.5 Frengression on sequential data . . . 88
5.5.1 Model . . . 89
5.5.2 Estimation . . . 89
5.5.3 Sampling . . . 91
5.6 Numerical examples on supported settings . . . 92
5.6.1 Binary intervention . . . 92
5.6.2 Continuous treatment . . . 94
5.6.3 Longitudinal setting and survival data . . . 96
5.6.4 Distributional regression . . . 98
5.7 Application on LEADER . . . 100
5.8 Discussion . . . 103

6 Stable prediction via causal invariance 105
6.1 Background . . . 105
6.2 Invariance and the causal prediction function . . . 108
6.3 Synthetic environments via reweighting . . . 109
6.4 Cramér–von Mises penalty . . . 111
6.5 Distributional-robustness interpretation . . . 112
6.6 Identification . . . 113
6.7 Using observed environments . . . 114
6.8 Synthetic experiment . . . 115
6.9 Ongoing work . . . 117

7 Discussion and future work 118
7.1 Securing and evaluating stability . . . 118
7.2 Limitations . . . 119
7.3 Future directions . . . 119
7.4 Concluding remarks . . . 120

A Supplementary material for Chapter 3 121
A.1 Main proofs . . . 121
A.1.1 Cross-fitting lemma . . . 121
A.1.2 Proof of Proposition 3.3.3 . . . 123
A.1.3 Proof of Proposition 3.3.5 . . . 124
A.1.4 Proof of Proposition 3.3.6 . . . 125
A.1.5 Proof of Theorem 3.4.3 . . . 126

A.1.6 Proof of Theorem 3.4.4 . . . 128
A.1.7 Proof of Proposition 3.4.6 . . . 130
A.1.8 Discussion of Proposition 3.4.7 . . . 130
A.1.9 Proof of Proposition 3.4.8 . . . 131
A.1.10 Proofs of Theorem 3.4.9 and Corollary 3.4.10 . . . 131
A.2 Weights estimation . . . 133
A.3 Derivation of weights in Example 3.3.4 . . . 133
A.4 Implementation details . . . 136
A.5 Detailed synthetic experiment results . . . 136
A.6 Additional experiment: Amazon review . . . 136

B Supplementary material for Chapter 4 140
B.1 Auxiliary results . . . 140
B.1.1 Consequences of a fixed pooling ratio . . . 140
B.1.2 Validity with merged samples . . . 141
B.1.2.1 Partial bootstrap . . . 141
B.1.2.2 Partial permutation . . . 142
B.1.3 Consistency with merged samples . . . 143
B.2 Main proofs . . . 144
B.2.1 Proof of Theorem 4.3.1 . . . 144
B.2.2 Proof of Proposition 4.3.4 . . . 144
B.2.3 Proof of Theorem 4.3.5 . . . 147
B.2.4 Proof of Theorem 4.3.6 . . . 149
B.2.5 Proof of Theorem 4.3.8 . . . 151
B.2.6 Proof of Theorem 4.3.9 . . . 152
B.2.7 Proof of Theorem 4.3.12 . . . 154
B.2.8 Proof of Theorem 4.3.16 . . . 156
B.2.9 Proof of Corollary 4.3.17 . . . 157
B.2.10 Proof of Corollary 4.3.18 . . . 158
B.2.11 Proof of Theorem 4.3.9 with converging sample size ratios . . . 158
B.3 Additional experiment results on normal approximation . . . 160
B.4 Extending Equivalence-TTP to U-statistic estimators . . . 160

C Supplementary material for Chapter 5 163
C.1 Auxiliary results . . . 163
C.1.1 Distribution consistency from energy distance convergence . . . 163
C.2 Main proofs . . . 165
C.2.1 Proof of Lemma 5.2.6 . . . 165
C.2.2 Proof of Proposition 5.2.7 . . . 165
C.2.3 Proof of Theorem 5.3.3 . . . 166
C.2.4 Proof of Corollary 5.3.4 . . . 173

C.2.5 Proof of Proposition 5.4.2 . . . 174
C.3 Hyperparameters in frengression . . . 176
C.4 Experiment details . . . 176
C.4.1 Binary intervention . . . 176
C.4.2 Longitudinal analysis . . . 177
C.4.2.1 Synthetic data . . . 177
C.4.2.2 Survival analysis: comparison with `dltmle` . . . 177
C.4.3 LEADER . . . 178

D Supplementary material for Chapter 6 180
D.1 Proof of Proposition 6.3.1 . . . 180
D.2 Proof of Proposition 6.3.2 . . . 181
D.3 Proof of Proposition 6.4.1 . . . 181
D.4 Proof of Lemma 6.5.1 . . . 182
D.5 Proof of Theorem 6.5.2 . . . 183
D.6 Proof of Proposition 6.6.2 . . . 183
D.7 Proof of Lemma 6.7.2 . . . 186
D.8 Proof of Proposition 6.7.4 . . . 186

Bibliography 187

Chapter 1

Introduction

A scientific conclusion is more reliable when small perturbations of the conditions under which it was reached do not alter its main claim. Relevant perturbations include those induced by sampling variability, population shifts, and modelling assumptions. *Stability* is the formal name for this property, and it should be required of any conclusion intended to support action.

The relationship between causality and stability runs in two directions, and this thesis addresses both. Causality has long been linked to stability through notions that formalise invariance under perturbations: "autonomy" in the econometric tradition of Haavelmo (1944) and Aldrich (1989), where a structural relation is autonomous if it retains its form under changes elsewhere in the system; "transportability" in the data-fusion framework of Bareinboim and Pearl (2016), where a causal effect identified in one environment remains valid in another under suitable conditions; and explicitly "invariance" in the distributional robustness and invariant prediction lines of Peters et al. (2016); Meinshausen (2018); Bühlmann (2020), where the conditional distribution of a response given its direct causes remains unchanged across environments, including under interventions on other variables. These notions differ in scope and formalisation but share a common idea: causal relationships are characterised by invariance to changes in their environment. Invariance is the limiting form of stability, corresponding to no change over a specified perturbation class.[1] Conversely, conclusions drawn about causal relationships should themselves be stable. Estimators or tests whose sign or decision is reversed by small perturbations do not provide a reliable basis for action, even if valid under the assumed model or asymptotically.

A canonical illustration is the LaLonde benchmark (LaLonde, 1986; Dehejia and Wahba, 1999; Imbens and Xu, 2024), where observational estimates of a single treatment effect vary substantially with the choice of estimator and adjustment set, even though the experimental

[1] In parts of the literature (e.g. Dawid and Didelez, 2010), "stability" is used for what we call "invariance". Here we distinguish the two: stability denotes a broader notion, defined through statistical properties of estimators and inference procedures; see Chapter 2.

benchmark is fixed. The instability arises from the estimation procedure, not from the target itself. Chapter 2 gives further examples where methods regarded as stable yield causal conclusions that fail under relevant perturbations.

With this in view, the thesis first turns to stability for causal inference. Chapter 3 introduces the augmented marginal outcome density ratio estimator for the average treatment effect. Inverse propensity score weighting becomes unstable under weak overlap; the estimator uses outcome information to reduce its variance and keep estimation stable. Chapter 4 develops a test-then-pool procedure for fusing evidence across data sources. Conventional test-then-pool procedures inflate the Type-I error of the downstream causal effect test; the proposed procedure, building on a maximum mean discrepancy equivalence test and calibrated by partial bootstrap or partial permutation, controls it across a class of source heterogeneities. Chapter 5 introduces frengression, a deep generative model with a parameterisation that centres around the target causal estimand. Each component of the frengression model is variation independent and customisable, the causal target in particular, so the fitted model can simulate data with a known causal effect for evaluating causal inference methods.

Chapter 6 then considers predictive robustness from the perspective of causal invariance. In contrast to the preceding chapters, which concern stability for causal inference, the chapter focuses on the invariance that the underlying causal mechanism provides across environments, and the predictive robustness that follows from it.

The thesis concludes in Chapter 7 with a discussion of limitations and directions for future work.

Chapter 2

Stability and causality

2.1 The language of causality

This thesis is written in two related causal languages: the potential-outcomes framework of Neyman (1923) and Rubin (1974), and the structural causal model of Pearl (2009). The two languages provide closely related representations of causal quantities and of the conditions under which they are identifiable, but locate causal structure in different objects, as the closing paragraphs of Sections 2.1.1 and 2.1.2 make explicit. We use whichever language is more economical for the setting at hand. Standard references are Hernán and Robins (2025) and Imbens and Rubin (2015) for potential outcomes, and Pearl (2009) and Peters et al. (2017) for the structural and graphical sides.

2.1.1 Potential outcomes

The potential-outcomes framework (Neyman, 1923; Rubin, 1974) represents causal effects through counterfactual outcomes: a causal effect is defined as a contrast between outcomes that would be observed on a single unit under different treatment values. Holland (1986) summarised this perspective with the maxim "no causation without manipulation", emphasising that causal effects are defined relative to interventions that can, at least conceptually, be applied to the same unit.

Definition 2.1.1 (Potential outcomes)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let X be a treatment taking values in $\mathcal{X}$, Z covariates taking values in $\mathcal{Z}$, and Y an outcome. For each $x \in \mathcal{X}$, let $Y(x)$ be a random variable on $(\Omega, \mathcal{F}, P)$. The collection $\{Y(x) : x \in \mathcal{X}\}$ is called the potential outcomes.*

For each unit, the observed outcome is the potential outcome at the realised treatment, $Y = Y(X)$; the remaining potential outcomes $\{Y(x) : x \in \mathcal{X},\ x \neq X\}$ are counterfactual. This

is the "fundamental problem of causal inference" (Holland, 1986), a missing-data problem at its core. Representing each unit's outcome by a single $Y(x)$ per treatment level, with $Y = Y(X)$, presumes the stable unit treatment value assumption (SUTVA; Rubin, 1980): there is no interference between units, so a unit's potential outcomes do not depend on the treatments assigned to others, and each treatment level has a single version, the latter underpinning the consistency condition in Assumption 2.1.2.

A causal *estimand* is a functional of the distribution of the potential outcomes. When $\mathcal{X} = \{0, 1\}$, the canonical example is the average treatment effect (ATE) $\theta = \mathbb{E}[Y(1) - Y(0)]$. For continuous $\mathcal{X}$, the dose-response curve $x \mapsto \mathbb{E}[Y(x)]$ plays the same role.

To identify these estimands, it then suffices to recover the interventional distribution of $Y(x)$ at every $x \in \mathcal{X}$ from the observed data (Z, X, Y).[2] The g-formula (Robins, 1986) achieves this under the following assumptions, in which $\perp\!\!\!\perp$ denotes conditional independence:

Assumption 2.1.2 (Identification)**.**

(a) Consistency: $X = x$ implies $Y = Y(x)$, $\quad \forall x \in \mathcal{X}$;

(b) Unconfoundedness: $Y(x) \perp\!\!\!\perp X \mid Z$, $\quad \forall x \in \mathcal{X}$;

(c) Positivity: the conditional density satisfies $p_{X \mid Z}(x \mid z) > 0$ for almost every $(x, z) \in \mathcal{X} \times \mathcal{Z}$ (for discrete treatment, $\Pr(X = x \mid Z = z) > 0$).

Proposition 2.1.3 (g-formula)**.** *Under Assumption 2.1.2, for every $x \in \mathcal{X}$ the interventional distribution of $Y(x)$ is identified as*

$$P_{Y(x)}(\cdot) = \int P(Y \in \cdot \mid X = x, Z = z)\, \mathrm{d}P_Z(z), \tag{2.1}$$

and in particular $\mathbb{E}[Y(x)] = \int \mathbb{E}[Y \mid X = x, Z = z]\, \mathrm{d}P_Z(z)$.

This expression coincides with the back-door adjustment in the structural causal model framework (see Proposition 2.1.8).

The potential-outcomes language locates the inferential question on the side of the *outcome*. Stability is a property of the procedure that estimates a target functional of the potential-outcome distribution. It requires that quantities such as variance, Type-I error, or estimation

[2] Identifying $P_{Y(x_0)}$ at a single level x_0 requires the conditions of Assumption 2.1.2 only at x_0; they are stated for all $x \in \mathcal{X}$ because the dose-response curve ranges over $\mathcal{X}$ and the average treatment effect contrasts two levels. As these estimands are means, it is enough that $\mathbb{E}[Y(x)]$, rather than the full distribution $P_{Y(x)}$, is identifiable.

error remain controlled across a specified class of perturbations of the data-generating distribution. Definition 2.2.1 makes this precise.

2.1.2 Structural causal models

The structural causal model (SCM; Pearl, 2009) represents causation through *structural assignments*: each variable is generated as a function of its direct causes and an exogenous noise term. Interventions act by replacing the structural assignment of a variable while leaving all other assignments unchanged.

The graphical structure underlying an SCM is captured by a directed acyclic graph (DAG).

Definition 2.1.4 (Directed acyclic graph)**.** *A directed graph $\mathcal{G} = (V, E)$ consists of a set of nodes V and a set of directed edges $E \subseteq V \times V$, where $(\tilde{a}, \tilde{b}) \in E$ represents an edge from $\tilde{a}$ to $\tilde{b}$. The graph is acyclic if it contains no directed cycles. For $\tilde{b} \in V$, the parents of $\tilde{b}$ are*

$$\mathrm{pa}(\tilde{b}) = \{\tilde{a} \in V : (\tilde{a}, \tilde{b}) \in E\}.$$

A node $\tilde{b}$ is a descendant of $\tilde{a}$ if there exists a sequence $v_0, v_1, \ldots, v_k$ of nodes with $v_0 = \tilde{a}$ and $v_k = \tilde{b}$, such that $(v_{i-1}, v_i) \in E$ for all i.

An SCM is then specified by attaching a structural assignment to each node of the graph, generating it from its parents and an independent noise term.

Definition 2.1.5 (SCM)**.** *An SCM on a DAG $\mathcal{G} = (V, E)$ with endogenous variables*[3] *$V = \{X_1, \ldots, X_d\}$ is a tuple $\mathcal{M} = (\mathcal{G}, U, \mathcal{F}, P_U)$ where:*

- *$U = \{U_1, \ldots, U_d\}$ is a collection of exogenous variables with joint distribution P_U;*
- *$\mathcal{F} = \{f_1, \ldots, f_d\}$ is a collection of measurable functions such that*

$$X_i = f_i(\mathrm{pa}(X_i), U_i), \quad i = 1, \ldots, d,$$

  where $\mathrm{pa}(X_i)$ denotes the parents of X_i in $\mathcal{G}$;

- *the exogenous variables are jointly independent (the Markovian case), i.e.*

[3] As is standard in structural causal models (Pearl, 2009), the endogenous variables X_i are determined within the model by structural assignments, while the exogenous variables U_i are determined outside it and are assigned by no structural equation.

$$P_U = \prod_{i=1}^{d} P_{U_i}.$$

The graph $\mathcal{G}$ encodes the factorisation of the observational distribution: if $\mathcal{G}$ is acyclic and the distribution admits a probability density p, then

$$p(x_1, \dots, x_d) = \prod_{i=1}^{d} p(x_i \mid \mathrm{pa}(x_i)). \tag{2.2}$$

Definition 2.1.6 (Intervention)**.** *For $i \in \{1, \dots, d\}$ and $x \in \mathcal{X}_i$, the intervention $\mathrm{do}(X_i = x)$ replaces the structural assignment for X_i by the constant function $f_i(\cdot) = x$, leaving all other assignments unchanged. The resulting model is denoted $\mathcal{M}_{\mathrm{do}(X_i=x)}$.*

The joint distribution of V induced by $\mathcal{M}_{\mathrm{do}(X_i=x)}$ is denoted $P_{\mathrm{do}(X_i=x)}$. Identification of interventional quantities is characterised by graphical criteria.

Definition 2.1.7 (Back-door criterion)**.** *Let $\mathcal{G}$ be a directed acyclic graph, and let X and Y be two nodes. A set of variables Z satisfies the back-door criterion relative to (X, Y) if:*

(a) no element of Z is a descendant of X in $\mathcal{G}$;

(b) Z blocks[4] every path between X and Y that contains an arrow into X.

Proposition 2.1.8 (Back-door adjustment)**.** *If Z satisfies the back-door criterion relative to (X, Y) in $\mathcal{G}$, then the interventional distribution is identified as*

$$P(Y \in \cdot \mid \mathrm{do}(X = x)) = \int P(Y \in \cdot \mid X = x, Z = z)\, \mathrm{d}P_Z(z),$$

and in particular $\mathbb{E}[Y \mid \mathrm{do}(X = x)] = \int \mathbb{E}[Y \mid X = x, Z = z]\, \mathrm{d}P_Z(z)$.

Both can be found in Pearl (2009, Definition 3.3.1 and Theorem 3.3.2).

An SCM induces potential outcomes via $Y(x) := Y$ in $\mathcal{M}_{\mathrm{do}(X=x)}$, with $Y = Y(X)$ for the observed outcome, providing a correspondence between the two frameworks.

The structural formulation places the inferential question on the side of the *mechanism*. An intervention replaces the structural assignment of the variable it acts on and leaves every other assignment in place (Haavelmo, 1944; Pearl, 2009; Peters et al., 2017). If such an intervention

[4] We use "blocks" in the standard d-separation sense: a path is blocked by Z if it contains a non-collider in Z, or a collider (a node with two incoming arrows along the path) such that neither the collider nor any of its descendants is in Z; see Pearl (2009, §1.2.3).

does not act on Y, it leaves the mechanism for Y in place, so the conditional distribution $P(Y \mid \mathrm{pa}(Y))$ it induces is unchanged. Writing $\mathcal{I}(Y)$ for the set of distributions arising from interventions that do not alter the structural assignment of Y, the invariance of this conditional reads

$$P(Y \in \cdot \mid \mathrm{pa}(Y) = w) = P'(Y \in \cdot \mid \mathrm{pa}(Y) = w) \quad \text{for all } P, P' \in \mathcal{I}(Y) \text{ and all } w. \tag{2.3}$$

This invariance is the basis of invariant causal prediction (Peters et al., 2016), to which we return in Section 2.3.

2.2 Stability for causal inference

A procedure is stable, in the sense used throughout this thesis, if its performance is preserved under perturbations of the data-generating distribution.

Definition 2.2.1 (Stability)**.** *Let $\theta(P)$ be a target functional, $\hat{\theta}$ a procedure or fitted object targeting it, and $J(\hat{\theta}, Q)$ a measure of its performance under a distribution Q. Let Π be a class of distributions and let $\delta \geq 0$. Then $\hat{\theta}$ is* (Π, J, δ)-stable *if*

$$\sup_{Q \in \Pi} J(\hat{\theta}, Q) \leq \delta. \tag{2.4}$$

The triple (Π, J, δ) specifies the perturbations under which performance is controlled, the notion of performance, and the admissible level. Causal invariance is the special case $\delta = 0$. Here $\Pi = \mathcal{I}(Y)$, $\hat{\theta}$ is the conditional $P(Y \mid \mathrm{pa}(Y))$ at a reference $P \in \Pi$, and $J(\hat{\theta}, Q)$ the discrepancy between $Q(Y \mid \mathrm{pa}(Y))$ and $\hat{\theta}$; $\sup_{Q \in \Pi} J(\hat{\theta}, Q) = 0$ is exactly the invariance (2.3). The robust-prediction problem of Section 2.3 is a further instantiation of the same triple, now with $\hat{\theta}$ (parameters of) a prediction function and $J(\hat{\theta}, Q)$ its risk under Q, and δ bounding the worst-case risk over Π.

The methodology chapters draw on this definition in different ways. Chapter 3 takes $\hat{\theta}$ to be the augmented marginal outcome density ratio estimator of the average treatment effect, Π a class of distributions admitting weak overlap, J the asymptotic variance, and δ a finite bound; under suitable conditions $\hat{\theta}$ reduces the variance relative to inverse propensity score weighting (IPW), especially under weak overlap. Chapter 4 addresses a data fusion problem and takes $\hat{\theta}$ to be the test-then-pool procedure for the treatment effect, Π a class of joint distributions over

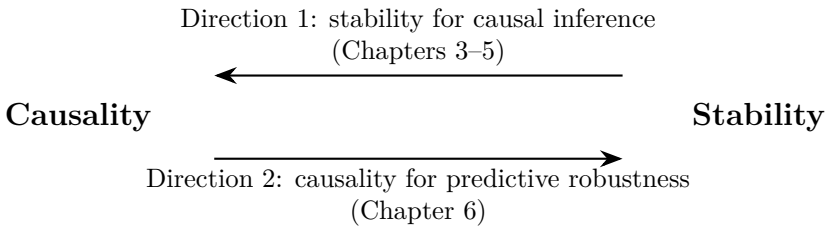


Figure 2.1: Two directions in the relationship between causality and stability developed in this thesis. The first uses stability to support causal conclusions; the second uses causal structure to characterise perturbation classes under which prediction functions remain stable.

heterogeneous data sources, J the Type-I error, and δ the nominal level α. Chapter 5 provides a generative model with known causal structure that supplies the controlled perturbations Π and a ground-truth target, against which the stability of an inference procedure $\hat{\theta}$ can be assessed.

Beyond the discussion in this thesis, several established literatures fit the same template. In classical robustness theory (Hampel, 1974; Huber, 1981), Π is an ε-contamination neighbourhood of a reference distribution, and J measures sensitivity through breakdown points or influence functions. Algorithmic stability (Bousquet and Elisseeff, 2002) considers perturbations of the empirical distribution induced by deleting or reweighting observations, with J measuring sensitivity of the estimator.

The preceding sections study stability of procedures for causal inference: given a causal target, the goal is to construct estimators or tests whose performance remains controlled under perturbations of the data-generating distribution. A complementary perspective asks whether stability can instead be obtained from the structure of the data-generating mechanism (Figure 2.1). The next section develops this perspective for prediction.

2.3 Causality for predictive robustness

Predictive models are increasingly deployed in applications where the data encountered at deployment are not distributed as the data used for training. A model fitted by empirical risk minimisation (ERM) is tuned to the training distribution, and its performance can deteriorate substantially when the test distribution moves away from it. This section studies the other direction between causality and stability, in which the target is predictive robustness under such distribution shifts. In the notation of (2.4), this corresponds to (Π, J, δ)-stability with $\Pi = \mathcal{Q}$ a class of environments, J a prediction risk, and δ a bound on the worst-case risk.

With X the predictors and Y the response, let $f_\theta : \mathcal{X} \to \mathbb{R}$ be a prediction function and ℓ a loss function. For a class $\mathcal{Q}$ of distributions on (X, Y), define

$$J(\theta, Q) = \mathbb{E}_Q[\ell\{Y, f_\theta(X)\}].$$

We formulate robust prediction as the target of inference

$$\arg\min_\theta \sup_{Q \in \mathcal{Q}} J(\theta, Q), \tag{2.5}$$

which selects a prediction function minimising the worst-case risk over $\mathcal{Q}$.

Several approaches to robust prediction share this worst-case form, differing in the set $\mathcal{Q}$ they protect against: adversarial training optimises the prediction against the worst-case bounded perturbation of each input (Sinha et al., 2017), and distributionally robust optimisation takes $\mathcal{Q}$ to be a ball $\{Q : D(Q, P) \leq \rho\}$ around the observed distribution P in a pre-specified distance or divergence D (Duchi and Namkoong, 2021). Such a ball is agnostic to the mechanism generating the shift: a radius large enough to cover the distributions actually encountered also admits many that do not arise, so the prediction function minimising the worst-case risk can be overly conservative, whereas a smaller radius excludes some of the shifts that do arise, so the worst-case guarantee does not extend to them.

Causality-oriented robustness (Meinshausen, 2018; Bühlmann, 2020) instead takes $\mathcal{Q}$ to be the distributions generated by interventions that leave the mechanism of Y unchanged, a class determined by the causal structure rather than by a postulated distance from P. Let $X_S = \mathrm{pa}(Y)$ denote the direct causes of Y, so $P(Y \mid X_S)$ is the same for every $Q \in \mathcal{Q}$. Among predictors built on X_S, the risk-minimiser is determined by $P(Y \mid X_S)$, not by the marginal of X_S; it is therefore optimal at every $Q \in \mathcal{Q}$ at once, thus minimising the worst-case risk over $\mathcal{Q}$.

The optimisation problem (2.5) does not in general recover this invariant conditional distribution. In practice, the objective is often augmented with regularisation terms that penalise variation of the function across environments. The choice of $\mathcal{Q}$ fixes the perturbation class, while regularisation controls how closely the solution approximates invariance. The resulting prediction functions need not be causal, but can exhibit stable performance over $\mathcal{Q}$.

In this direction, causality characterises invariant prediction functions and specifies perturbation classes under which they remain valid. This perspective underlies invariant causal prediction (Peters et al., 2016) and the distributional robustness view of Meinshausen (2018);

Heinze-Deml and Meinshausen (2021); Shen et al. (2025). Chapter 6 develops this perspective further.

2.4 Sources of instability in causal inference

Stability can arise either from the design of a statistical procedure or from invariances implied by the underlying causal structure. In practice, however, achieving it is complicated by the characteristics of the data at hand. Finite samples limit what can be learned reliably; complex structure makes causal quantities difficult to isolate and increases sensitivity to modelling choices; and distribution shift requires conclusions to remain valid beyond the environment in which they are learned. A distinct question is whether the causal target is identifiable from the observable distribution; this issue is discussed separately in Section 2.4.4.

2.4.1 Finite-sample fragility

Finite samples limit the stability of causal estimators even under correct identification. This appears both within a single data source and across multiple sources.

Within a single source. One example is weak overlap. When the propensity score $\Pr(X = x \mid Z)$ approaches 0 or 1, the variance of inverse propensity score weighting diverges. This regime is common in high-dimensional settings and in applications with rare exposures (Stürmer et al., 2010; D'Amour et al., 2021). The instability is therefore driven by the sampling distribution rather than by identification.

Across sources. A typical example arises when the primary study has a limited sample size; a natural solution is to augment it with additional data sources to improve the precision of the estimated causal effect. The difficulty is that the sources need not be fully comparable: pooling may reduce variance but can also introduce bias, so the usual strategy is to test for compatibility and pool only when the test is passed. With limited data the test has little power: incompatibility goes undetected, the uncertainty of this preliminary decision is ignored, and the Type-I error of the resulting causal conclusion is inflated.

These examples reflect the same constraint: the data available do not support stable estimation without additional structure. Chapter 3 addresses weak overlap within a single source, and Chapter 4 the problem of drawing a causal conclusion across heterogeneous sources with controlled Type-I error.

2.4.2 Structural complexity

The preceding subsection focuses on low-dimensional causal targets, typically average treatment effects under simple treatment regimes. Many applications, however, involve substantially greater structural complexity.

Longitudinal treatments, continuous exposures, and high-dimensional outcomes require flexible models for the joint distribution. In these regimes, small perturbations of nuisance components can induce large changes in the estimated causal functional. Instability is therefore structural: it arises from the complexity of the data-generating process rather than from sampling variability alone.

A second consequence is that validation becomes indirect. Since the causal target is unobservable, methods can only be compared through simulation, where synthetic data with a known mechanism provide the ground truth against which estimators are stress-tested (de Vassimon Manela et al., 2025). A generative model that is itself unstable cannot provide a reliable benchmark. The problem is thus twofold: stable estimation of the target and stable simulation of the underlying mechanism. Chapter 5 addresses both aspects.

2.4.3 Distribution shift

In many applications, data are collected in one environment and used in another. The resulting distribution shift can invalidate procedures that are stable under a single distribution but not across a class of environments.

Stability in this setting is necessarily defined relative to a class of perturbations. Without additional structure, performance under one distribution does not extend to others. Causal structure provides one route to such guarantees by identifying conditional relationships that remain invariant across environments, as discussed in Chapter 6.

2.4.4 Threats to identification

The discussion so far assumes that the causal target is identifiable from the observable distribution. Identification and stability address different questions. Identification asks whether the target can be recovered from the observed data under the assumed causal model; stability asks how sensitive the resulting inference is to perturbations of the data or the analysis.

Identification relies on assumptions connecting the observed and causal distributions. These assumptions may fail for a variety of reasons, including unobserved confounding, selection bias, measurement error, adjustment for post-treatment variables, or violations of SUTVA (Pearl, 2009; Hernán and Robins, 2025). When they do, the causal target is no longer determined by the observable distribution, and no statistical procedure can recover it without additional assumptions.

The methods developed in this thesis assume that the causal target, or the causal structure on which it rests, has been justified by a separate identification argument. They address the stability of estimation, testing, and prediction, rather than the identification of the target itself.

Taken together, these sources of instability motivate the design principles developed in this thesis. The methodology chapters address finite-sample fragility and structural complexity by controlling performance under perturbations, while the perspective of Chapter 6 exploits causal structure to achieve prediction stability across environments.

2.5 Notation

Beyond the causal notation introduced earlier in this chapter (potential outcomes $Y(x)$ and conditional independence $\perp\!\!\!\perp$), we use the following conventions throughout. P and Q denote probability distributions and P^n, Q^n their empirical measures based on n observations, with $P_n[f] = \int f \, \mathrm{d}P^n = n^{-1} \sum_{i=1}^n f(X_i)$ the corresponding empirical average operator; $\Pr(A)$ is the probability of an event A; and $\mathbb{E}$ denotes expectation, with $\mathbb{E}f = \int f \, \mathrm{d}P$. Further subscripts identify the distribution in question: P_V and $P_{V \mid W}$ are the distribution of a random variable V and its conditional distribution given W, while P_θ is the distribution indexed by a parameter θ, such as the distribution generated by a model with parameters θ in Chapter 5. We write $X \sim P$ when X has distribution P, $X \overset{d}{=} Y$ for equality in distribution, and "i.i.d." for independent and identically distributed. For a sequence of random variables, $\xrightarrow{p}$, $\rightsquigarrow$, and $\xrightarrow{a.s.}$ denote, respectively, convergence in probability, weak convergence, and almost sure convergence, and $O_P(\cdot)$, $o_P(\cdot)$ denote the usual stochastic orders. For $q \geq 1$, $\|f\|_q = \left(\int |f|^q \, \mathrm{d}P \right)^{1/q}$ is the L_q norm, $L_q(P) = \{f : \|f\|_q < \infty\}$ the corresponding space, and $L_2^0(P) = \{f \in L_2(P) : \mathbb{E}f = 0\}$ its mean-zero subspace; the same notation applies under any marginal, for instance $L_2(P_X)$. Chapter-specific objects are introduced where they first appear.

Symbol	Meaning by context
X	treatment (Chapters 3 and 5); predictors (Sections 2.1.2 and 2.3 and chapter 6); generic integration variable (Chapter 4)
Z	covariates (Chapters 3 and 5); generic integration variable (Chapter 4)
Y	outcome, in causal settings (Chapters 3 and 5); response, in predictive settings (Sections 2.1.2 and 2.3 and chapter 6); generic integration variable (Chapter 4)
W	an observation (Z, X, Y) (Chapter 3); multinomial bootstrap weights (Chapter 4); conditioning variables (Chapter 5)
θ	average treatment effect (Chapter 3); generative model parameters (Chapter 5)
ρ	equivalence radius (Chapter 4); correlation (Chapter 5)
ξ	nuisance parameter (Chapter 3); kernel mean embedding (Chapters 4 and 6); exogenous noise (Chapter 5)
μ	outcome mean (Chapters 3–5)
π	propensity score (Chapters 3 and 5); permutation (Chapter 4)
η	learned representation (Chapter 3); exogenous noise (Chapter 5)
ζ	kernel bandwidth (Chapter 4); exogenous noise (Chapter 5)
γ	pooling weight (Chapter 4); regularisation parameter (Chapter 6)
λ	sample ratio (Chapter 4); regularisation parameter (Chapter 6)
β	regression coefficient (Chapters 3, 5 and 6); angle between kernel mean embeddings (Chapter 4)

Table 2.1: Symbols used with more than one meaning across the thesis, with the meaning intended in each context.

To match the conventions of each area, several symbols are used with more than one meaning across the thesis; we collect them in Table 2.1.

Chapter 3

Stable estimation via outcome-informed weighting

In Chapter 2, an inference procedure is stable when its performance remains controlled under a class of perturbations of the data-generating distribution. This chapter develops one such procedure for estimating the average treatment effect under weak overlap, where inverse propensity score weighting is unstable: its weights carry variation irrelevant to the causal effect. Existing remedies act on the covariates, by selecting an adjustment set, imposing balance, or learning a representation. The proposed augmented marginal outcome density ratio estimator instead removes that variation by outcome-informed weighting, reducing the instability without a priori assumptions on the covariate structure. This chapter is a revised and extended version of joint work with Professor Robin J. Evans (Yang and Evans, 2025).

3.1 Introduction

Doubly robust estimators such as augmented inverse propensity score weighting (AIPW; Robins et al., 1994) remain consistent for the average treatment effect (ATE) if either the propensity score or the outcome regression is correctly specified. Inverse propensity score weighting (IPW), however, exposes them to practical positivity violation (Petersen et al., 2012): when some covariates predict treatment too well, the estimated weights become extreme and the finite-sample variance is large; at the population level, including treatment-predictive variables in the adjustment can also raise the efficiency bound (Robins et al., 2007).

Not all covariates need to enter the propensity score adjustment. Adjusting for instrumental variables, which affect only the treatment, does not reduce bias but can substantially inflate variance (Pearl, 2010). Henckel et al. (2022) and Rotnitzky and Smucler (2020) make this precise: efficient adjustment sets need not include variables that predict treatment only, and

outcome-predictive variables are the relevant ones for efficiency.

This motivates the marginal outcome density ratio estimator (MR) of Taufiq et al. (2024), which *shifts the focus from the covariate space to the outcome space*: the observed outcomes determine which information enters the weights. Introduced for policy learning, MR is revisited in this chapter for treatment-effect estimation, in particular its connection to covariate adjustment. Building on it, we introduce the augmented marginal outcome density ratio estimator (AMR), which inherits the double robustness of AIPW through a bridge identity; the additional estimation problem reduces to a one-dimensional regression on the outcome residual.

Existing methods for handling high-dimensional covariates and limited overlap typically rely on *a priori* structural assumptions, such as covariate-selection rules, balancing constraints, or learned representations, which gain efficiency when correct but propagate misspecification when not. MR and AMR instead reduce variation in the inverse propensity component by a projection applied after nuisance parameter estimation, rather than by selecting covariates or learning a covariate representation; both estimators replace *a priori* structural assumptions with a post-hoc, outcome-based adjustment. We prove an oracle variance reduction relative to IPW and AIPW, and find empirically that MR and AMR are substantially more stable under weak overlap.

The remainder of this chapter is organised as follows. Section 3.2 fixes notation and reviews related work. Section 3.3 defines MR and AMR, describes the role of outcome-informed weighting, relates the estimators to covariate-adjustment methods, and discusses weight estimation. Section 3.4 develops the asymptotic properties of $\hat{\theta}^0_{AMR}$ and $\hat{\theta}_{AMR}$. Empirical results are presented in Section 3.5.

3.2 Background

3.2.1 Problem set-up and notation

Throughout this chapter W denotes the full data vector $W = (Z, X, Y) \sim P \in \mathcal{P}$ for a class $\mathcal{P}$ of probability measures, with $Z \in \mathcal{Z} \subseteq \mathbb{R}^d$ (d pre-treatment covariates) and $X \in \{0, 1\}$ binary. The empirical average operator $P_n[\cdot]$ of Section 2.5 acts on the sample $\{W_i\}_{i=1}^n$. We use the L_q norm $\|\cdot\|_q$ and the spaces $L_q(P)$, $L_2^0(P)$ of Section 2.5, with $\|f\|_2$ the default; its empirical counterpart is $\|f\|_{n,2} = (P_n[f^2])^{1/2}$. When f depends only on a sub-vector of W, $\|f\|_2$ denotes

its L_2 norm under the marginal distribution of that sub-vector: under P_Z when f is a function of the covariates, and under the distribution of the outcome residual when f is a function of that residual.

In the potential outcomes framework (Neyman, 1923; Rubin, 1974), $Y(x)$ denotes the outcome under treatment $x \in \{0,1\}$. Define the conditional mean $\mu^x(z) \coloneqq \mathbb{E}[Y(x) \mid Z = z]$ and the propensity score $\pi(z) \coloneqq \Pr(X = 1 \mid Z = z)$, with estimates $\hat{\pi}(z)$ and $\hat{\mu}^x(z)$ for $x \in \{0,1\}$. The argument of these nuisances, true or estimated, is suppressed when it is clear from context. The target parameter is the ATE $\theta = \mathbb{E}[Y(1) - Y(0)]$, occasionally written $\theta(P)$; $\hat{\theta}$ denotes an estimator of θ.

Throughout this chapter, we assume the identification conditions of Assumption 2.1.2 are satisfied, here with a binary treatment $X \in \{0,1\}$. Thus, $\mu^x(z) = \mathbb{E}[Y \mid X = x, Z = z]$, and the observed outcome can be written as $Y = XY(1) + (1 - X)Y(0)$. The mean of each potential outcome $\mathbb{E}Y(x)$ is identified as $\mathbb{E}\mu^x(Z) = \mathbb{E}[\mathbb{E}[Y \mid Z, X = x]]$, so $\theta = \mathbb{E}[\mu^1(Z) - \mu^0(Z)]$.

Define $\mu^*(Z) \coloneqq \pi(Z)\mu^0(Z) + \{1 - \pi(Z)\}\mu^1(Z)$ and the *outcome residual* $Y^* \coloneqq Y - \mu^*(Z)$, with estimated counterparts $\hat{\mu}^*(Z)$ and $\hat{Y}^* = Y - \hat{\mu}^*(Z)$. Here μ^* is not the conditional mean $\mathbb{E}[Y \mid Z] = \pi(Z)\mu^1(Z) + (1 - \pi(Z))\mu^0(Z)$, so Y^* differs from the ordinary regression residual $Y - \mathbb{E}[Y \mid Z]$. The "clever covariate" from the targeted maximum likelihood estimation (TMLE; Van der Laan et al., 2011) literature is $h(X,Z) = (X - \pi(Z))/(\pi(Z)\{1 - \pi(Z)\})$, with estimate $\hat{h}(X,Z) = (X - \hat{\pi}(Z))/(\hat{\pi}(Z)\{1 - \hat{\pi}(Z)\})$. The IPW and AIPW estimators are then $P_n[\hat{h}(X,Z)Y]$ and $P_n[\hat{h}(X,Z)\hat{Y}^*]$, respectively.

3.2.2 Efficient estimators and efficiency bound of ATE

With θ the target functional and $\xi = (\mu^1, \mu^0, \pi)$ the nuisance parameters,[5] consider the full model

$$\mathcal{M} = \{P : Y(x) \perp\!\!\!\perp X \mid Z,\ 0 < \pi(Z) < 1\}. \tag{3.1}$$

Under Assumption 2.1.2, the efficient influence function (EIF) for θ in $\mathcal{M}$ is

$$\varphi(W;\xi) = \frac{X}{\pi(Z)}(Y - \mu^1(Z)) - \frac{1 - X}{1 - \pi(Z)}(Y - \mu^0(Z)) + (\mu^1(Z) - \mu^0(Z)) - \theta, \tag{3.2}$$

[5] Within this chapter ξ denotes the nuisance parameters, distinct from the kernel mean embedding of Chapter 4 and the exogenous noise of Chapter 5.

which can also be written as $\varphi(W;\xi) = h(X,Z)Y^* - \theta$.

The function φ is the EIF of θ at P. We briefly recall the underlying semiparametric notions, following van der Vaart (2000); Tsiatis (2006). The model $\mathcal{M}$ is a set of distributions; the local analysis instead takes place in the Hilbert space $L_2^0(P)$ of mean-zero, square-integrable functions of W, where the scores and influence functions below reside. A submodel is a subset $\mathcal{M}' \subseteq \mathcal{M}$ of distributions containing P; its local structure is probed through regular parametric submodels, smooth one-parameter families $\{P_\varepsilon\} \subseteq \mathcal{M}'$ indexed by a perturbation ε in a neighbourhood of 0 with $P_0 = P$. The score of such a path, $s = \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \log p_\varepsilon\big|_{\varepsilon=0} \in L_2^0(P)$, is the direction in which the distribution is perturbed away from P, and the tangent space $\mathcal{T}(P,\mathcal{M}')$ is the closed linear span of all such scores. An estimator is *regular* if its limiting distribution is unchanged by such local perturbations of the model; among all regular estimators of θ, the smallest attainable asymptotic variance is the variance of the EIF, the nonparametric efficiency bound.

AIPW and TMLE are constructed from the estimating equation $P_n\varphi(W;\hat{\xi},\hat{\theta}) = 0$ with $\hat{\xi} = (\hat{\mu}^0, \hat{\mu}^1, \hat{\pi})$. Under regularity nuisance-rate conditions, they are asymptotically normal (Chernozhukov et al., 2018; Van der Laan et al., 2011):

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow \mathcal{N}\Big(0, \mathrm{Var}\left(\varphi(W;\xi)\right)\Big),$$

where $\mathrm{Var}\left(\varphi(W;\xi)\right)$ is the nonparametric efficiency bound for θ. We refer to van der Vaart (2000); Tsiatis (2006); Kennedy (2024) for detailed reviews.

3.2.3 Related work

Identification of the ATE rests on no unmeasured confounding and positivity, two conditions that pull against each other: adjusting for more confounders reduces bias[6] but risks violating positivity, inflating the variance of inverse propensity score weighting. This trade-off is especially pronounced in high-dimensional settings (Petersen et al., 2012; D'Amour et al., 2021).

Selection rules for the adjustment set vary across the literature. Rosenbaum (2002) and Rubin (2009) advocate conditioning on all pre-treatment covariates; Vander Weele and Shpitser (2011) restricts adjustment to variables affecting either treatment or outcome; Brookhart et al. (2006) restricts further to variables affecting the outcome; and Pearl (2010) cautions that including instrumental variables can amplify confounding bias.

[6] Though not always, as the *M-bias* example shows (Greenland et al., 1999).

A separate line of work does not select variables by their role, but instead controls the variance of inverse propensity score weighting through the estimation procedure itself. The covariate balancing propensity score (CBPS; Imai and Ratkovic, 2014) and its high-dimensional extension hd-CBPS (Ning et al., 2020) estimate the propensity score under covariate-balancing constraints. Collaborative targeted maximum likelihood estimation (CTMLE; van der Laan and Gruber, 2010; Díaz, 2018; Ju et al., 2019; Benkeser et al., 2020) jointly learns the propensity score and the outcome regression to optimise the bias-variance trade-off, selecting a propensity model from a candidate pool by checking whether outcome-regression bias is reduced; the procedure has many decision points and is computationally demanding. van der Laan and Gruber (2010) observed that when the outcome regression is well estimated, little inverse propensity score weighting is needed. Building on this, Díaz (2018) construct the adapted propensity score (e-score) by regressing the outcome bias on $\hat{\pi}(Z)$ and then regressing $\hat{\pi}(Z)$ on the result, projecting all information back onto Z. In a related variant, Benkeser et al. (2020) propose an estimator conditioning on the consistently estimated outcome mean $\hat{\mu}^x(Z)$.

Representation-learning methods integrate the choice of adjustment into the estimation procedure. Shi et al. (2019) propose DragonNet, which prioritises treatment-predictive information even at the expense of outcome prediction, in contrast with approaches that sacrifice propensity score accuracy to mitigate lack of overlap.

Christgau and Hansen (2024) extend Rotnitzky and Smucler (2020) by adjusting through a learned representation $V = \eta(Z)$. Define $\theta(V) = \mathbb{E}\big[\mathbb{E}[Y \mid V, X = 1] - \mathbb{E}[Y \mid V, X = 0]\big]$. The EIF for $\theta(V)$ takes the form of (3.2) with Z replaced by V, and its variance defines the semiparametric efficiency bound for regular estimators of $\theta(V)$. Define analogously $\theta(Z) = \mathbb{E}\big[\mathbb{E}[Y \mid Z, X = 1] - \mathbb{E}[Y \mid Z, X = 0]\big]$. When $\mathbb{E}_V\big[\mathbb{E}[Y \mid V, X = x]\big] = \mathbb{E}_Z\big[\mathbb{E}[Y \mid Z, X = x]\big]$, $x \in \{0, 1\}$, we have $\theta(Z) = \theta(V)$, and the choice between adjustment by Z and by V reduces to whichever yields the smaller efficiency bound.

Suppose the outcome regression factors as $\mu^x(z) = g(x, \eta(\beta, z))$. Christgau and Hansen (2024) propose the debiased outcome-adapted propensity estimator (DOPE) using $V_\beta = \eta(\beta, Z)$ and show that, under suitable conditions (their Corollary 3.6), $\theta(V_\beta) = \theta(Z)$ and

$$\operatorname{Var}\left(\varphi(Z)\right) \geq \operatorname{Var}\left(\varphi(V_\beta)\right), \tag{3.3}$$

within some submodels. Letting $\hat{\theta}(Z)$ and $\hat{\theta}_\beta$ denote estimators conditioning on Z and V_β with β known, (3.3) implies $\hat{\theta}_\beta$ is at least as efficient as $\hat{\theta}(Z)$. In practice, β is unknown and must

be estimated, introducing additional uncertainty. In particular, Christgau and Hansen (2024) adopt the single-index form

$$Y = g(\alpha X + \beta^\top Z) + \epsilon_Y, \tag{3.4}$$

and estimate β via a neural network with a linear first layer enforcing this structure; the outcome and treatment are then regressed on $\hat{\beta}^\top Z$.

Several earlier methods fit into this framework. Benkeser et al. (2020) can be viewed as a special case of DOPE that conditions on the estimated outcome mean $\hat{\mu}^x(Z)$ rather than on $V_{\hat{\beta}} = \hat{\beta}^\top Z$; Gui and Veitch (2022)'s transformer-based TI estimator for text data shares the same logic; and Díaz's e-score corresponds to $V = \mathbb{E}[Y - \hat{\mu}^x(Z) \,|\, \hat{\pi}(Z)]$.

These methods rely on a priori structural knowledge that is not always available. The back-door criterion (Definition 2.1.7) requires the true causal structure; covariate-balancing methods typically assume linear relationships (Imai and Ratkovic, 2014; Athey et al., 2018; Ning et al., 2020); and representation-learning approaches (Benkeser et al., 2020; Gui and Veitch, 2022; Christgau and Hansen, 2024) depend on a pre-processing step that must recover the relevant representation. Errors in that step propagate in ways that are difficult to control, especially in high dimensions. The semiparametric theory of Christgau and Hansen (2024) applies to specific representation structures and does not account for the uncertainty introduced by estimating $\hat{\beta}$ or for misspecification. As shown in Section 3.5.1, such assumptions can lead to substantial bias.

Another line of work reduces variance by altering the target estimand, such as smoothed propensity score trimming (Rosenbaum, 2002; Crump et al., 2009; Yang and Ding, 2018) and overlap weights (Li et al., 2018). Unlike these, MR and AMR still target the ATE.

3.3 Methodology

3.3.1 Marginal ratio and augmented marginal ratio estimators

Recall that the IPW and AIPW estimators are defined as $\mathbb{E}[h(X,Z)Y]$ and $\mathbb{E}[h(X,Z)Y^*]$ in the population version. We denote the population weights, $w(Y) = \mathbb{E}[h(X,Z) \,|\, Y]$ and $w^*(Y^*) = \mathbb{E}[h(X,Z) \,|\, Y^*]$. These weights define MR and AMR, respectively. Equivalently, MR replaces the IPW weight $h(X,Z)$ by its $L_2(P)$-projection onto $\sigma(Y)$, and AMR applies the same operation on the outcome residual, projecting $h(X,Z)$ onto $\sigma(Y^*)$. The corresponding estimated weight

	Population	Oracle	Estimator
IPW	$\theta_{IPW} = \mathbb{E}[h(X,Z)Y]$	$\hat{\theta}^0_{IPW} = P_n[h(X,Z)Y]$	$\hat{\theta}_{IPW} = P_n[\hat{h}(X,Z)Y]$
AIPW	$\theta_{AIPW} = \mathbb{E}[h(X,Z)Y^*]$	$\hat{\theta}^0_{AIPW} = P_n[h(X,Z)Y^*]$	$\hat{\theta}_{AIPW} = P_n[\hat{h}(X,Z)\hat{Y}^*]$
MR	$\theta_{MR} = \mathbb{E}[w(Y)Y]$	$\hat{\theta}^0_{MR} = P_n[w^0(Y)Y]$	$\hat{\theta}_{MR} = P_n[\hat{w}(Y)Y]$
AMR	$\theta_{AMR} = \mathbb{E}[w^*(Y^*)Y^*]$	$\hat{\theta}^0_{AMR} = P_n[w^{*0}(\hat{Y}^*)\hat{Y}^*]$	$\hat{\theta}_{AMR} = P_n[\hat{w}^*(\hat{Y}^*)\hat{Y}^*]$

Table 3.1: Summary of estimators.

functions are denoted $\hat{w}$ and $\hat{w}^*$, when we use estimated nuisance parameters $\hat{\pi}$ and $\hat{\mu}^x$. With these notations, we define:

Definition 3.3.1 (MR and AMR estimators)**.** *Our proposed empirical MR and AMR estimators are*

$$\hat{\theta}_{MR} = P_n[\hat{w}(Y)Y] = P_n\left\{\hat{\mathbb{E}}\left[\frac{X-\hat{\pi}(Z)}{\hat{\pi}(Z)\{1-\hat{\pi}(Z)\}}\,\middle|\,Y\right]Y\right\}, \tag{3.5}$$

and

$$\hat{\theta}_{AMR} = P_n\left[\hat{w}^*(\hat{Y}^*)\hat{Y}^*\right] = P_n\left\{\hat{\mathbb{E}}\left[\frac{X-\hat{\pi}(Z)}{\hat{\pi}(Z)\{1-\hat{\pi}(Z)\}}\,\middle|\,\hat{Y}^*\right]\hat{Y}^*\right\}. \tag{3.6}$$

For later discussion, we also introduce the oracle estimators

$$w^0 = \mathbb{E}[\hat{h}(X,Z)\,|\,Y] = \mathbb{E}\left[\frac{X-\hat{\pi}(Z)}{\hat{\pi}(Z)\{1-\hat{\pi}(Z)\}}\,\middle|\,Y\right]$$

and

$$w^{*0} = \mathbb{E}\left[\hat{h}(X,Z)\,\middle|\,\hat{Y}^*\right] = \mathbb{E}\left[\frac{X-\hat{\pi}(Z)}{\hat{\pi}(Z)\{1-\hat{\pi}(Z)\}}\,\middle|\,\hat{Y}^*\right]$$

as intermediate quantities. Table 3.1 summarises these definitions. For IPW and AIPW, the superscript 0 denotes that the nuisance parameters π and μ^x are treated as known; for MR and AMR, it denotes that the induced conditional weight function (w^0 or w^{*0}) is treated as known, while the nuisance estimates $\hat{\pi}$ and $\hat{\mu}^x$ may still enter the construction. The "Estimator" column shows the usual case in which the nuisances and weight functions are all estimated.

The weight function w^{*0} is defined as the conditional expectation given the estimated $\hat{Y}^*$, which follows a different distribution than the true Y^* that defines w^*. When the estimated nuisance parameters $\hat{\mu}^x$ and $\hat{\pi}$ exactly match their true values μ^x and π, the two functions w^*

and w^{*0} are identical. In contrast, w^0 is the conditional expectation given Y, aligning it with the definition of w. The two need not be identical in general: w^0 involves $\hat{h}(X, Z)$, whereas w involves $h(X, Z)$. They coincide when $\hat{\pi} = \pi$. A simple example illustrating the differences among w^{*0}, w^*, w^0, and w is provided in Example 3.3.4.

In practice, the weights are estimated by solving[7]

$$\hat{w} = \arg\min_{f \in \mathcal{F}} P_n \left[\frac{X - \hat{\pi}(Z)}{\hat{\pi}(Z)\{1 - \hat{\pi}(Z)\}} - f(Y) \right]^2, \tag{3.7}$$

$$\hat{w}^* = \arg\min_{f \in \mathcal{F}} P_n \left[\frac{X - \hat{\pi}(Z)}{\hat{\pi}(Z)\{1 - \hat{\pi}(Z)\}} - f(\hat{Y}^*) \right]^2 \tag{3.8}$$

within a function class $\mathcal{F}$.

Algorithm 3.1 summarises AMR estimation. The MR estimator follows the same steps without computing $\hat{Y}^*_{-k}$, and instead estimates $\hat{w}_{-k}$ by regressing $\hat{h}_{-k}(X, Z)$ on Y_{-k}.

Algorithm 3.1 Estimating ATE with Augmented Marginal Outcome Density Ratio

Input: Observations $\{Z_i, Y_i, X_i\}_{i=1}^n$.
Output: AMR estimation for ATE, $\hat{\theta}_{AMR}$

Split the data into K folds, denoted $\{D_k\}_{k=1}^K$.
for $k = 1$ to K **do**

Train nuisance parameter estimation models on the training set D_{-k} (all data except fold D_k) to obtain: $\hat{\mu}^x_{-k}$ and $\hat{\pi}_{-k}$ that take covariates $z \in \mathcal{Z}$ as input;

Using the trained models on D_{-k}, compute for all observations in D_{-k}: $\hat{Y}^*_{-k} \coloneqq \{\hat{Y}^*_i = Y_i - \hat{\pi}_{-k}(Z_i)\hat{\mu}^0_{-k}(Z_i) - \{1 - \hat{\pi}_{-k}(Z_i)\}\hat{\mu}^1_{-k}(Z_i), i \in D_{-k}\}$;

Still on the training set D_{-k}, regress $\hat{h}_{-k}(X, Z) \coloneqq \left\{ \frac{X_i - \hat{\pi}_{-k}(Z_i)}{\hat{\pi}_{-k}(Z_i)(1 - \hat{\pi}_{-k}(Z_i))}, i \in D_{-k} \right\}$ on $\hat{Y}^*_{-k}$ to estimate the weight function $\hat{w}^*_{-k}$ using (3.8);
end for
Compute the average across all n observations:

$$\hat{\theta}_{AMR} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}\{i \in D_k\} \hat{w}^*_{-k}(\hat{Y}^*_{i,k}) \cdot \hat{Y}^*_{i,k},$$

where $\hat{Y}^*_{i,k} = Y_i - \hat{\pi}_{-k}(Z_i)\hat{\mu}^0_{-k}(Z_i) - \{1 - \hat{\pi}_{-k}(Z_i)\}\hat{\mu}^1_{-k}(Z_i)$.

Remark 3.3.2 (Cross-fitting)**.** *In Algorithm 3.1 we fit $\hat{w}^*_{-k}$ on the same fold along with other nuisance parameters $\hat{\pi}_{-k}$ and $\hat{\mu}^x_{-k}, x \in \{0, 1\}$. One can also split the fold D_{-k} into D^0_{-k} and D^1_{-k}, fit the nuisance parameters $\hat{\pi}_{-k}$ and $\hat{\mu}^x_{-k}, x \in \{0, 1\}$ on D^0_{-k}, and $\hat{w}^*_{-k}$ on D^1_{-k}. In the finite sample context, we recommend following Algorithm 3.1 that uses all the data in D_{-k} to estimate the weight function, since this makes the best use of potentially small amounts of data.*

[7] Other loss functions, such as the Huber loss can also be used.

When the nuisance parameters and the projection weight functions are known, MR and AMR have variance no larger than IPW and AIPW:

Proposition 3.3.3 (Variance comparison with IPW and AIPW)**.** *The marginal weights* $w(Y) = \mathbb{E}[h(X,Z)\,|\,Y]$ *and* $w^*(Y^*) = \mathbb{E}[h(X,Z)\,|\,Y^*]$ *are the* $L_2(P)$ *projections of* $h(X,Z)$ *onto* $\sigma(Y)$ *and* $\sigma(Y^*)$*, and by the law of total variance*

$$\begin{aligned}\operatorname{Var}\big(h(X,Z)Y\big) - \operatorname{Var}\big(w(Y)Y\big) &= \mathbb{E}\big[\operatorname{Var}\big(h(X,Z)\,|\,Y\big)Y^2\big] \geq 0,\\ \operatorname{Var}\big(h(X,Z)Y^*\big) - \operatorname{Var}\big(w^*(Y^*)Y^*\big) &= \mathbb{E}\big[\operatorname{Var}\big(h(X,Z)\,|\,Y^*\big)Y^{*2}\big] \geq 0.\end{aligned}$$

Each gap is the variance of the component of $h(X,Z)$ removed by the projection, strictly positive whenever $\Pr\{Y^2 \operatorname{Var}(h(X,Z)\,|\,Y) > 0\} > 0$ (for MR) or $\Pr\{(Y^*)^2 \operatorname{Var}(h(X,Z)\,|\,Y^*) > 0\} > 0$ (for AMR); the proof is in Appendix A.1.2. When the nuisances are known, the oracle estimators are sample means of these quantities and, dividing by n, inherit the ordering: $\operatorname{Var}(\hat{\theta}^0_{MR}) \leq \operatorname{Var}(\hat{\theta}^0_{IPW})$ and $\operatorname{Var}(\hat{\theta}^0_{AMR}) \leq \operatorname{Var}(\hat{\theta}^0_{AIPW})$. The AMR gap is small when Y^* is small in magnitude and large when $\operatorname{Var}(h(X,Z)\,|\,Y^*)$ is large.

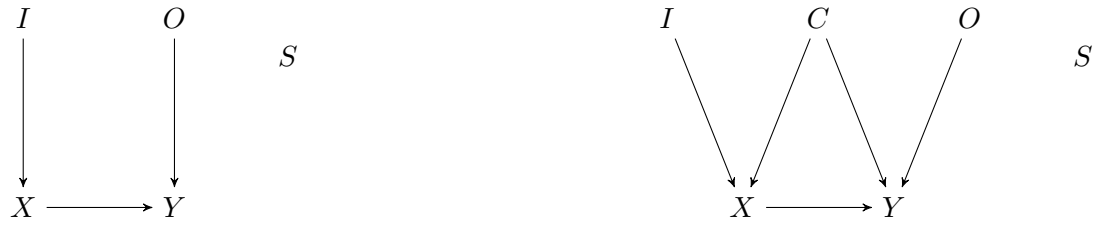
So far we assume the nuisance parameters μ^x and π to be known. In practice they are unknown and must be estimated. Section 3.4 relates AMR to AIPW through the bridge identity, establishes double robustness for $\hat{\theta}^0_{AMR}$, and gives a bias-aware expansion for $\hat{\theta}_{AMR}$. Before turning to these properties, we develop the intuition behind the proposed estimators.

3.3.2 Outcome-informed weighting

The proposed weighting assigns observations with the same value of Y (or Y^*) the same weight, smoothing variation in $h(X,Z)$ that is uninformative for the estimating equation. We illustrate this with Example 3.3.4. The derivation is detailed in Appendix A.3.

Example 3.3.4 (Explicit form of MR and AMR)**.** *Assume* $X\,|\,Z \sim$ *Bernoulli*$(\pi(Z))$ *and* $Y(x)\,|\,Z \sim \mathcal{N}(\mu^x(Z), \sigma^2)$ *for* $x \in \{0,1\}$*. Write* $\mathcal{N}(t; m, s^2)$ *for the density of* t *under a Gaussian distribution with mean* m *and variance* s^2*. We obtain:*

$$\begin{aligned}w(y) &= \frac{\mathbb{E}_Z\Big[\mathcal{N}(y;\mu^1(Z),\sigma^2) - \mathcal{N}(y;\mu^0(Z),\sigma^2)\Big]}{\mathbb{E}_Z\big[\pi(Z)\mathcal{N}(y;\mu^1(Z),\sigma^2) + \{1-\pi(Z)\}\mathcal{N}(y;\mu^0(Z),\sigma^2)\big]};\\ w^*(y^*) &= \frac{\mathbb{E}_Z\Big[\mathcal{N}(y^*;\pi(Z)\tau(Z),\sigma^2) - \mathcal{N}(y^*;-\{1-\pi(Z)\}\tau(Z),\sigma^2)\Big]}{\mathbb{E}_Z\big[\pi(Z)\mathcal{N}(y^*;\pi(Z)\tau(Z),\sigma^2) + \{1-\pi(Z)\}\mathcal{N}(y^*;-\{1-\pi(Z)\}\tau(Z),\sigma^2)\big]}\end{aligned}$$



(a) Without confounders. (b) With confounders.

Figure 3.1: Example causal DAGs.

where $\tau(z) \coloneqq \mu^1(z) - \mu^0(z)$.

Meanwhile, we can also derive the weights with plug-in nuisance estimators:

$$w^0(y) = \frac{\mathbb{E}_Z\left[\frac{\pi(Z)}{\hat{\pi}(Z)}\mathcal{N}(y;\mu^1(Z),\sigma^2) - \frac{1-\pi(Z)}{1-\hat{\pi}(Z)}\mathcal{N}(y;\mu^0(Z),\sigma^2)\right]}{\mathbb{E}_Z\left[\pi(Z)\mathcal{N}(y;\mu^1(Z),\sigma^2) + \{1-\pi(Z)\}\mathcal{N}(y;\mu^0(Z),\sigma^2)\right]};$$
$$w^{*0}(y^*) = \frac{\mathbb{E}_Z\left[\frac{\pi(Z)}{\hat{\pi}(Z)}\mathcal{N}(y^*;\mu^1(Z)-\hat{\mu}^*(Z),\sigma^2) - \frac{1-\pi(Z)}{1-\hat{\pi}(Z)}\mathcal{N}(y^*;\mu^0(Z)-\hat{\mu}^*(Z),\sigma^2)\right]}{\mathbb{E}_Z\left[\pi(Z)\mathcal{N}(y^*;\mu^1(Z)-\hat{\mu}^*(Z),\sigma^2) + \{1-\pi(Z)\}\mathcal{N}(y^*;\mu^0(Z)-\hat{\mu}^*(Z),\sigma^2)\right]}.$$

These differ in general from $w(y)$ *and* $w^*(y^*)$*, and coincide when* $\hat{\pi} = \pi$ *and* $\hat{\mu}^x = \mu^x$.

In Example 3.3.4, with no treatment effect, we obtain $w = w^* = 0$. The population MR and AMR then place zero weight on every observation, while inverse propensity score weighting keeps the raw $h(X, Z)$ and its full variance under weak overlap, though both target $\theta = 0$. This reflects how the projection averages out the variation in $h(X, Z)$ that is uninformative for the estimating equation, preventing extreme values from dominating.

3.3.3 Connection with covariate adjustment

MR and AMR use weights that are outcome-informed; here we describe their connection with covariate-adjustment methods. We give two examples in Figure 3.1, where I represents the instrumental variables, those correlated with the treatment X and affecting the outcome Y only through treatment; O contains the prognostic variables, those that predict the outcome without being related to the treatment; C represents the confounding variables that are correlated with both the treatment and the outcome; S consists of spurious variables, those representing random noise with no direct correlation with either the outcome or treatment. In what follows, we focus on the MR weights.

Proposition 3.3.5 (MR weights)**.** *Let the variables be as in Figure 3.1.*

(a) No confounders. *Suppose* $Z = \{I, O, S\}$ *as in Figure 3.1a, so* $\pi(Z) = \pi(I)$ *and* $Y \perp\!\!\!\perp I \,|\, X$. *Then*

$$\mathbb{E}\left[\frac{X - \pi(Z)}{\pi(Z)\{1 - \pi(Z)\}} \,\middle|\, Y\right] = \frac{\Pr(X = 1 \,|\, Y)}{\Pr(X = 1)} - \frac{\Pr(X = 0 \,|\, Y)}{\Pr(X = 0)}.$$

(b) With confounders. *Suppose* $Z = \{I, C, O, S\}$ *as in Figure 3.1b, with* $I \perp\!\!\!\perp Y \,|\, X, C$ *and the propensity score depending on* Z *only through* (I, C). *Then*

$$\begin{aligned}\mathbb{E}\left[\frac{X - \pi(Z)}{\pi(Z)\{1 - \pi(Z)\}} \,\middle|\, Y\right] = &\Pr(X = 1 \,|\, Y)\, \mathbb{E}\big[\Pr(X = 1 \,|\, C)^{-1} \,\big|\, Y, X = 1\big] \\ &- \Pr(X = 0 \,|\, Y)\, \mathbb{E}\big[\Pr(X = 0 \,|\, C)^{-1} \,\big|\, Y, X = 0\big].\end{aligned}$$

Proofs are in Appendix A.1.3. In both cases the instrumental variable I does not enter the weight through its own variation. In the no-confounder case the weight depends on the outcome through $\Pr(X = x \,|\, Y)$ and on treatment prevalence through $\Pr(X = x)$, but not on the propensity score $\pi(I)$; under confounding it depends on the propensity score only through $\Pr(X = x \,|\, C)$. Conditioning on Y averages the variation in I out of the weight.

This aligns with the principle underlying efficient covariate adjustment. An instrumental variable affects the outcome only through treatment and therefore lies outside the optimal adjustment set (Rotnitzky and Smucler, 2020; Henckel et al., 2022), as including it in the propensity score can increase the variance of inverse propensity score weighting. The projection of $h(X, Z)$ onto $\sigma(Y)$ achieves this automatically. Because $h(X, Z)$ enters only through $\mathbb{E}[h(X, Z) \,|\, Y]$, components of $h(X, Z)$ that are informative about treatment but not about the outcome do not contribute directly to the resulting weight.

3.3.4 Weights estimation

The weights w^0 and w^{*0} are unknown in practice and are estimated by (3.7) and (3.8). Proposition 3.3.6 gives a decomposition of the finite-sample mean squared error (MSE) cost and gain from replacing w^{*0} by $\hat{w}^*$.

Proposition 3.3.6 (MSE comparison of $\hat{\theta}_{AMR}$ and $\hat{\theta}_{AIPW}$)**.** *Treat* $\hat{\mu}^x, \hat{\pi}, \hat{w}^*$ *as fixed functions, and let* $\hat{\theta}_{AMR}, \hat{\theta}_{AIPW}$ *be computed from* n *i.i.d. samples of* $W \sim P$ *using these nuisances. With* $\Delta := (\hat{w}^* - w^{*0})(\hat{Y}^*)\hat{Y}^*$,

$$\mathbb{E}[(\hat{\theta}_{AMR} - \theta)^2] - \mathbb{E}[(\hat{\theta}_{AIPW} - \theta)^2] = B(B + 2b) + \frac{\mathrm{Var}\left(\hat{w}^*(\hat{Y}^*)\hat{Y}^*\right) - \mathrm{Var}\left(w^{*0}(\hat{Y}^*)\hat{Y}^*\right) - V_h}{n}, \tag{3.9}$$

where $b = \mathbb{E}[\hat{h}(X, Z)\hat{Y}^*] - \theta$, $B = \mathbb{E}\Delta$, *and* $V_h = \mathbb{E}[(\hat{Y}^*)^2 \,\mathrm{Var}(\hat{h}(X, Z) \mid \hat{Y}^*)] \geq 0$. *If* $\hat{w}^* = w^{*0}$, *the right-hand side equals* $-V_h/n \leq 0$.

The proof is in Appendix A.1.4. The term $b = \mathbb{E}[\hat{h}(X, Z)\hat{Y}^*] - \theta$ is the bias of AIPW under the fixed nuisances $(\hat{\mu}^x, \hat{\pi})$, while $B = \mathbb{E}\Delta$ is the additional bias introduced by estimating w^{*0} with $\hat{w}^*$. The term V_h is the analogue of the variance gap in Proposition 3.3.3 with Y^* replaced by $\hat{Y}^*$, so V_h/n is the reduction in variance from replacing $\hat{h}(X, Z)$ with its conditional expectation given $\hat{Y}^*$; it is large when $\hat{h}(X, Z)$ varies substantially conditional on $\hat{Y}^*$, as happens under weak overlap. The remaining term $\left[\mathrm{Var}(\hat{w}^*(\hat{Y}^*)\hat{Y}^*) - \mathrm{Var}(w^{*0}(\hat{Y}^*)\hat{Y}^*)\right]/n$ is the change in variance from estimating $\hat{w}^*$.

Equation (3.9) weighs the bias cost $B(B + 2b)$ against the variance gain V_h/n and the variance change from estimating w^{*0}. If nuisances are well estimated, b is small and the bias cost is of order B^2. The bias B comes from estimating w^{*0} by $\hat{w}^*$; as the regression function $w^{*0} = \mathbb{E}[\hat{h}(X, Z) \mid \hat{Y}^*]$ is one-dimensional, an estimate with small bias is less demanding to obtain and B is typically small. AMR therefore improves on AIPW once the variance gain V_h/n exceeds this bias cost. A regularised $\hat{w}^*$ can further shrink $\mathrm{Var}(\hat{w}^*(\hat{Y}^*)\hat{Y}^*)$ below $\mathrm{Var}(w^{*0}(\hat{Y}^*)\hat{Y}^*)$, making the variance change negative and the overall reduction larger than V_h/n.

We use kernel ridge regression (KRR; Saunders et al., 1998) with an RBF kernel to obtain the weight estimates $\hat{w}$ and $\hat{w}^*$ throughout the main text. Empirically, KRR has the smallest MSE and the highest stability among the methods we tried. Its formulation and the alternative methods for estimating the weights are discussed in Appendix A.2.

3.4 Asymptotic properties

We study the asymptotic properties of the oracle AMR estimator $\hat{\theta}^0_{AMR}$ and the feasible estimator $\hat{\theta}_{AMR}$. Write

$$\hat{\xi}^0 := (w^{*0}, \hat{\pi}, \hat{\mu}^x), \qquad \hat{\xi} := (\hat{w}^*, \hat{\pi}, \hat{\mu}^x),$$

for the nuisance functions used by $\hat{\theta}^0_{AMR}$ and $\hat{\theta}_{AMR}$, respectively. Throughout this section, estimators are computed using the K-fold cross-fitting procedure in Algorithm 3.1, with fixed K and folds satisfying $|D_k|/n \to 1/K$. Unless stated otherwise, estimated nuisance functions for observations in D_k are trained on the complementary fold D_{-k}, and fold indices are suppressed for readability.

The function w^{*0} is the $L_2(P)$ projection of $\hat{h}(X, Z)$ onto $\sigma(\hat{Y}^*)$, conditional on the training fold. Hence the tower property gives the bridge identity

$$\mathbb{E}[w^{*0}(\hat{Y}^*)\hat{Y}^*] = \mathbb{E}[\hat{h}(X, Z)\hat{Y}^*]. \tag{3.10}$$

The following regularity conditions are used throughout this section.

Assumption 3.4.1 (Overlap and moment conditions)**.** *There exists $\epsilon > 0$ such that*

$$\epsilon \leq \pi(z) \leq 1 - \epsilon \qquad \text{for all } z \in \mathcal{Z}.$$

Moreover, with probability tending to one,

$$\epsilon \leq \hat{\pi}(z) \leq 1 - \epsilon \qquad \text{for all } z \in \mathcal{Z}.$$

The outcome regressions μ^x, for each $x \in \{0, 1\}$, are uniformly bounded, as are their estimates $\hat{\mu}^x$ with probability tending to one, and $\mathbb{E}[Y^4] < \infty$.

Assumption 3.4.2 (Weight regularity)**.** *The feasible weight estimator $\hat{w}^*$ is uniformly bounded with probability tending to one, and $\left\|(\hat{w}^* - w^{*0})(\hat{Y}^*)\right\|_2 = o_P(1)$.*

3.4.1 Consistency

Theorem 3.4.3 (Consistency of $\hat{\theta}^0_{AMR}$ and $\hat{\theta}_{AMR}$)**.** *Suppose Assumption 2.1.2 and Assumption 3.4.1 hold. If either $\|\hat{\pi} - \pi\|_2 \xrightarrow{p} 0$ or $\|\hat{\mu}^x - \mu^x\|_2 \xrightarrow{p} 0$, for each $x \in \{0, 1\}$, then*

$$\hat{\theta}^0_{AMR} \xrightarrow{p} \theta.$$

If Assumption 3.4.2 additionally holds, then

$$\hat{\theta}_{AMR} \xrightarrow{p} \theta.$$

The proof is provided in Appendix A.1.5. The oracle estimator $\hat{\theta}^0_{AMR}$ inherits the usual double robustness of AIPW through the bridge identity (3.10). The feasible estimator $\hat{\theta}_{AMR}$ additionally requires consistent estimation of w^{*0}. Because w^{*0} is a one-dimensional regression function, the condition $\left\|(\hat{w}^* - w^{*0})(\hat{Y}^*)\right\|_2 = o_P(1)$ is plausible under standard smoothness and regularisation conditions for parametric or kernel-based regression. This is a one-dimensional estimation problem on the outcome residual, leaving the covariate-space nuisances $\hat{\pi}$ and $\hat{\mu}^x$ and their double robustness untouched. Preprocessing strategies such as representation learning, variable selection (Benkeser et al., 2020; Gui and Veitch, 2022), and the single-index construction of Christgau and Hansen (2024) instead add a high-dimensional, structural step in the covariate space, whose error propagates to bias and whose efficiency gain is conditional on correct specification.

3.4.2 Asymptotic distribution and inference

For clarity, write (3.2) as $\varphi_{AIPW}(W;\xi)$. The AMR estimating function is

$$\varphi_{AMR}(W;\xi) = w^*(Y^*)Y^* - \theta. \tag{3.11}$$

The oracle estimator $\hat{\theta}^0_{AMR}$ is asymptotically linear with φ_{AMR}, which therefore determines its limiting distribution. We develop this distribution and the inference it supports here, deferring to Section 3.4.3 the question of how its limiting variance should be interpreted.

Theorem 3.4.4 (Asymptotic normality of $\hat{\theta}^0_{AMR}$)**.** *Suppose Assumptions 3.4.1 and 2.1.2 hold. Further assume that, for each $x \in \{0,1\}$,*

(a) $\|\hat{\pi} - \pi\|_2 \|\hat{\mu}^x - \mu^x\|_2 = o_P(n^{-1/2})$;

(b) $\|\hat{\pi} - \pi\|_2 \xrightarrow{p} 0$ *and* $\|\hat{\mu}^x - \mu^x\|_2 \xrightarrow{p} 0$;

(c) The projection $w^*(y^*) = \mathbb{E}[h(X,Z) \,|\, Y^* = y^*]$ is bounded and continuous at P_{Y^*}-almost every point, and

$$\left\|(w^{*0} - w^*)(\hat{Y}^*)\right\|_2 \xrightarrow{p} 0.$$

Then, at any fixed P satisfying these conditions, the oracle estimator $\hat{\theta}^0_{AMR}$ satisfies

$$\sqrt{n}(\hat{\theta}^0_{AMR} - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2_{AMR,0}),$$

where $\sigma^2_{AMR,0} = \mathrm{Var}\left(\varphi_{AMR}(W;\xi)\right)$.

The proof is provided in Appendix A.1.6. The limiting variance $\sigma^2_{AMR,0}$ is never larger than that of AIPW; the two differ by the variance of the component that the outcome projection removes.

Corollary 3.4.5 (Influence-function variance comparison)**.**

$$\mathrm{Var}\left(\varphi_{AIPW}(W;\xi)\right) - \mathrm{Var}\left(\varphi_{AMR}(W;\xi)\right) = \mathrm{Var}\left(h(X,Z)Y^* - w^*(Y^*)Y^*\right) \geq 0.$$

This is the influence-function form of Proposition 3.3.3: φ_{AIPW} and φ_{AMR} differ from $h(X,Z)Y^*$ and $w^*(Y^*)Y^*$ by θ, and since $w^*(Y^*)Y^* = \mathbb{E}[h(X,Z)Y^* \,|\, Y^*]$, the gap equals $\mathrm{Var}(h(X,Z)Y^* - w^*(Y^*)Y^*)$, the variance of the component the outcome projection removes.

Under the conditions of Theorem 3.4.4, if w^{*0} were available, a consistent estimator of $\sigma^2_{AMR,0}$ would be

$$\hat{\sigma}^2_{AMR,0} = P_n\big(w^{*0}(\hat{Y}^*)\hat{Y}^* - \hat{\theta}^0_{AMR}\big)^2.$$

In practice w^{*0} is estimated by $\hat{w}^*$. The feasible estimator has an additional first-order bias term from estimating w^{*0}. Let B_n denote the fold-weighted conditional bias from estimating w^{*0}:

$$B_n \coloneqq \sum_{k=1}^{K} \frac{|D_k|}{n}\, \mathbb{E}\Big[(\hat{w}^*_{-k} - w^{*0}_{-k})(\hat{Y}^*_{-k})\, \hat{Y}^*_{-k} \,\Big|\, \mathcal{F}_{-k}\Big].$$

Here $\mathcal{F}_{-k} \coloneqq \sigma\big(\{W_i : i \notin D_k\}\big)$ is the σ-algebra generated by the observations outside fold D_k.

Proposition 3.4.6 (Bias-aware asymptotic normality of $\hat{\theta}_{AMR}$)**.** *Under the conditions of Theorem 3.4.4 and Assumption 3.4.2,*

$$\sqrt{n}\,(\hat{\theta}_{AMR} - \theta - B_n) \rightsquigarrow \mathcal{N}(0, \sigma^2_{AMR,0}).$$

The proof of Proposition 3.4.6 is given in Appendix A.1.7. Equivalently, $\hat{\theta}_{AMR}$ is asymptotically centred at $\theta + B_n$, not θ, so the estimation bias appears explicitly; recovering the usual centring at θ requires $\sqrt{n}\,B_n = o_P(1)$, which the next result shows is exactly the condition for asymptotic linearity.

Proposition 3.4.7 (Asymptotic linearity of $\hat{\theta}_{AMR}$)**.** *Under the conditions of Proposition 3.4.6,*

$$\sqrt{n}\,(\hat{\theta}_{AMR} - \theta) = \sqrt{n}\,P_n\big[\varphi_{AMR}(W;\xi)\big] + o_P(1),$$

if and only if $\sqrt{n}\,B_n = o_P(1)$. *In this case we have* $\sqrt{n}\,(\hat{\theta}_{AMR} - \theta) \rightsquigarrow \mathcal{N}(0, \sigma^2_{AMR,0})$.

There is no generic, feasible rate condition on the nuisance estimators in $\hat{\xi}$ that guarantees $\sqrt{n}\,B_n = o_P(1)$ (Appendix A.1.8). A practical alternative is the conservative interval based on AIPW, as provided below:

Proposition 3.4.8 (Conservative confidence interval for $\hat{\theta}_{AMR}$)**.** *Fix a nominal level* $\alpha \in (0,1)$ *and let* $q_{1-\alpha/2}$ *be the* $(1-\alpha/2)$*-quantile of the standard normal distribution. Let* $\sigma_0 \coloneqq \sqrt{\mathrm{Var}(\varphi_{AMR}(W;\xi))}$, $\sigma_X \coloneqq \sqrt{\mathrm{Var}(\varphi_{AIPW}(W;\xi))}$, *and* $\nu \coloneqq \sigma_X - \sigma_0$. *Suppose the conditions of Proposition 3.4.6 hold. Assume in addition that* $\sigma_0 > 0$, $\nu > 0$, *and that, for some constant* $\delta > 0$,

$$\sqrt{n}\,|B_n| \le q_{1-\alpha/2}\,\nu - \delta \quad \textit{with probability tending to one as } n \to \infty. \tag{3.12}$$

Then the interval

$$\left[\hat{\theta}_{AMR} - q_{1-\alpha/2}\frac{\hat{\sigma}'_{AMR}}{\sqrt{n}}, \hat{\theta}_{AMR} + q_{1-\alpha/2}\frac{\hat{\sigma}'_{AMR}}{\sqrt{n}}\right], \tag{3.13}$$

with $\big(\hat{\sigma}'_{AMR}\big)^2 = P_n\big(\hat{h}(X,Z)\hat{Y}^* - \hat{\theta}_{AMR}\big)^2$, *has asymptotic coverage at least* $1-\alpha$.

The proof is given in Appendix A.1.9. The bound (3.12) is not verifiable from data, so Proposition 3.4.8 gives a sufficient condition for conservativeness rather than a guarantee. It is weaker than the linearity condition of Proposition 3.4.7: it does not require $\sqrt{n}\,B_n \to 0$, only that $\sqrt{n}\,|B_n|$ stay below the margin $q_{1-\alpha/2}\,\nu$, where $\nu = \sigma_X - \sigma_0$ is the gain in standard error from AMR. This margin is larger under weak overlap, where σ_X inflates while σ_0 stays controlled. We show in our experiments that the interval can attain coverage above the nominal level, particularly under weak overlap (Section 3.5.1).

The interval centres at $\hat{\theta}_{AMR}$ but takes the AIPW standard error, the larger of the two by Corollary 3.4.5; the extra width absorbs B_n and yields coverage at least $1-\alpha$. It therefore matches the width of the usual AIPW interval while centring at the more stable $\hat{\theta}_{AMR}$, which under weak overlap avoids the heavy tails and inflated MSE of $\hat{\theta}_{AIPW}$ (Figure 3.5). The construction is informative only when AMR reduces variance, $\mathrm{Var}\left(h(X,Z)Y^* - w^*(Y^*)Y^*\right) > 0$; otherwise estimating w^* adds nuisance error for no gain.

3.4.3 Efficiency interpretation via the submodel

The variance gap of Corollary 3.4.5 might suggest $\hat{\theta}^0_{AMR}$ is more efficient. A smaller variance, however, is an efficiency gain only among regular estimators, so this would require φ_{AMR} to be the efficient influence function of θ. In the full model $\mathcal{M}$ it is not: since $\mathcal{M}$ is nonparametric, its tangent space is all of $L^0_2(P)$, and the influence function of any regular estimator is unique and equals φ_{AIPW}. As $\varphi_{AMR} \neq \varphi_{AIPW}$, an estimator with influence function φ_{AMR} cannot be regular in $\mathcal{M}$; $\hat{\theta}^0_{AMR}$ is superefficient there, which carries no efficiency guarantee. It is, however, a valid influence function relative to submodels whose scores are orthogonal to the difference between the two estimating functions, which we now characterise.

Define

$$r(W) \coloneqq \varphi_{AIPW}(W;\xi) - \varphi_{AMR}(W;\xi) = \{h(X,Z) - w^*(Y^*)\}Y^*.$$

The score directions orthogonal to r form the set

$$\mathcal{T}_{\max}(P) = \{s \in L^0_2(P) : \mathbb{E}[r(W)s(W)] = 0\}. \tag{3.14}$$

Theorem 3.4.9 (Influence-function characterisation)**.** *Let $\mathcal{M}' \subseteq \mathcal{M}$ be a regular submodel with tangent space $\mathcal{T}(P,\mathcal{M}')$. Then φ_{AMR} is an influence function for θ relative to $\mathcal{M}'$ if and only if*

$$\mathcal{T}(P,\mathcal{M}') \subseteq \mathcal{T}_{\max}(P).$$

The theorem identifies the submodels in which φ_{AMR} is a valid influence function, namely those whose scores are orthogonal to r. We now give a concrete submodel, in which φ_{AMR} is also the efficient influence function. Define

$$\mathcal{T}^*(P) = \{f(Y^*) : \mathbb{E}[f(Y^*)] = 0,\ f \in L_2(P)\}. \tag{3.15}$$

This is the tangent space generated by scores that depend only on Y^*. Let $\mathcal{M}^* \subseteq \mathcal{M}$ denote any regular submodel with tangent space $\mathcal{T}^*(P)$.

Corollary 3.4.10 (Efficient influence function relative to $\mathcal{M}^*$)**.** *Under strict overlap and $\mathbb{E}[(Y^*)^2] < \infty$, $\varphi_{AMR} \in L^0_2(P)$ and φ_{AMR} is the EIF of θ in $\mathcal{M}^*$.*

The proofs are given in Appendix A.1.10.

Remark 3.4.11 (Examples of score directions)**.** *The submodel $\mathcal{M}^*$ is one model satisfying Theorem 3.4.9 but not the only one: φ_{AMR} stays a valid influence function in any submodel whose scores are orthogonal to r. Scores through Y^* are of this kind, but $\mathcal{T}_{\max}(P)$ is larger, containing scores that are not functions of Y^*. For instance, write $\bar{r}(X,Z) = \mathbb{E}[r \mid X, Z]$ and take a mean-zero $s_0(X,Z)$ not proportional to $\bar{r}$. The score*

$$s(X,Z) = s_0(X,Z) - \frac{\mathbb{E}[s_0\,\bar{r}]}{\mathbb{E}[\bar{r}^2]}\,\bar{r}(X,Z)$$

is mean-zero with $\mathbb{E}[s\,r] = \mathbb{E}[s\,\bar{r}] = 0$, so it lies in $\mathcal{T}_{\max}(P)$. As a function of the covariates alone, s is not a function of $Y^* = Y - \mu^*(Z)$ unless constant, so $s \notin \mathcal{T}^*(P)$.

The variance reduction in Corollary 3.4.5 holds at any P, with no model restriction; it becomes an efficiency statement within a submodel in which φ_{AMR} is the EIF, such as $\mathcal{M}^*$ (Corollary 3.4.10); there $\hat{\theta}^0_{AMR}$ attains a smaller efficiency bound than AIPW. This is different from covariate-adjustment methods such as that of Christgau and Hansen (2024), which attain an efficiency gain only when the covariate-outcome relationship is correctly specified, whereas AMR requires no structural assumption on the covariates.

3.5 Experiments

We report simulation studies and one observational data analysis. Throughout, the nuisance functions and the weight function $\hat{w}^*$ are estimated by $K = 5$-fold cross-fitting, as in Algorithm 3.1. In each simulation, we run R Monte Carlo replicates and report bias $\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \theta)$, mean absolute error (MAE) $\frac{1}{R}\sum_{r=1}^{R}|\hat{\theta}^{(r)} - \theta|$, and root mean squared error (RMSE) $\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \theta)^2}$, where $\hat{\theta}^{(r)}$ denotes the estimate from replicate r.

The code is available at https://github.com/linyingyang/AMR.

3.5.1 Synthetic experiment results

We report empirical results on synthetic data and compare the performance of AMR and MR versus TMLE, CTMLE, CBPS, AIPW, IPW and DOPE (with the single-index form (3.4)); implementation details for all methods are given in Appendix A.4.

We employ the `causl` package built on the frugal parameterisation framework (Evans and Didelez, 2024) to generate synthetic data that precisely follows the intended causal distribution. In our simulation, each covariate in Z is independently drawn from the normal distribution $\mathcal{N}(0,1)$. We fix the dimensions of the confounders (C), prognostic variables (O), and spurious variables (S) at 5 each, and later assess estimator performance by varying the dimensionality p_I of the instrumental variables (I). Their roles are shown in the causal DAG in Figure 3.1b. The data generation process is as follows:

- $X|Z \sim \text{Bernoulli}\left(\pi(Z)\right)$, where $\pi(Z) = \text{expit}\left(\sum_{j=1}^{p_I} I_j + 0.5\sum_{j=1}^{5} C_j\right)$;
- $Y|X,Z \sim \mathcal{N}(5X + \mu^0(Z), 1)$;
- $\mu^0(Z) = 10\sin\left(\pi\sum_{j=1}^{5} O_j\right) + 20\left(\sum_{j=1}^{5} O_j\right)^2 + \left(\sum_{j=1}^{5} C_j \cos\left(\pi\sum_{j=1}^{5} O_j\right)\right)^2$.

We design the potential outcome, $\mu^0(Z)$, to follow a complex nonlinear form. As shown later, methods that rely on a priori assumptions like DOPE can be substantially biased when these assumptions are violated. The design also implies that increasing p_I makes the propensity score $\pi(Z)$ progressively more extreme.

The propensity score $\pi(Z)$ is estimated using logistic regression. For a fair comparison with DOPE, the potential outcome model is estimated via a feedforward neural network; configurations are detailed in Appendix A.4. The initial estimates for the propensity score and potential outcome, where needed, are identical across all estimators. We perform estimation over $R = 200$ replicates, varying the sample size $n \in \{400, 600, 1000\}$ and the instrumental variable dimension $p_I \in \{1, 5, 10, 15, 20\}$.

Figure 3.2 presents the bias, MAE, and RMSE of these estimators across settings; the corresponding numerical values are in Appendix A.5. AMR has the smallest or near-smallest MAE and RMSE in most configurations and is markedly more stable than AIPW; when sample size increases, TMLE and CTMLE become competitive with AMR. For small n, AMR has larger bias from the regularisation in the weights step, which becomes comparable to the others as n grows. Figure 3.3 repeats the comparison with IPW and AIPW removed, since they perform poorly relative to the other methods. DOPE is omitted from Figure 3.2 and Figure 3.3 because its errors are on a substantially larger scale in this setting, consistent with sensitivity to misspecification of the single-index form (3.4) for $\mu^0(Z)$.

Figure 3.4 presents the coverage and lengths of the 95% confidence intervals. The interval for AMR is conservative and comparable in length to that for AIPW, consistent with the discussion

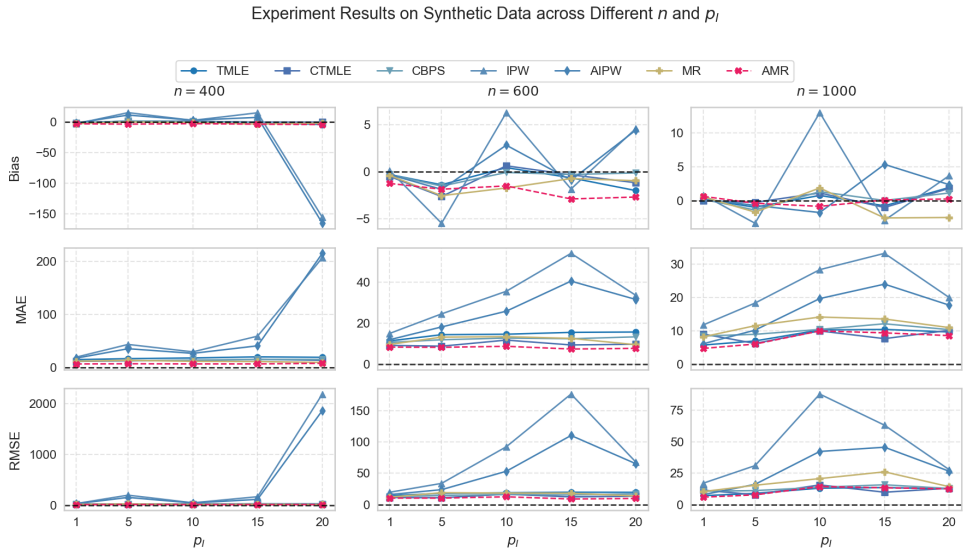


Figure 3.2: Bias, MAE and RMSE results on synthetic data, varying across p_I. IPW and AIPW are very unstable, especially when $n = 400$, $p_I = 20$.

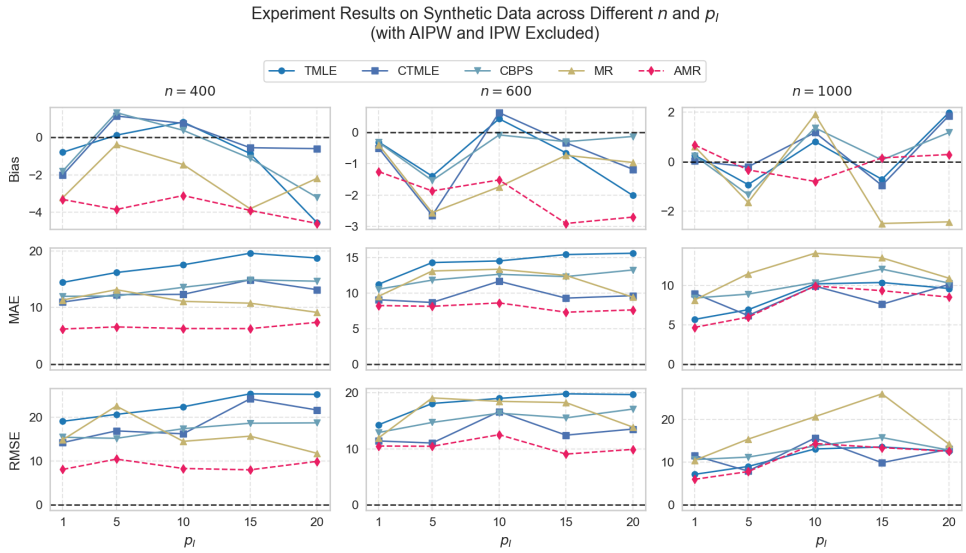


Figure 3.3: Bias, MAE and RMSE results on synthetic data, varying across p_I. Compared to Figure 3.2, we drop AIPW and IPW here for better visualisation. AMR shows better performance when sample size is small.

in Section 3.4. In this setting, TMLE and CTMLE yield intervals with under-coverage.

CBPS also shows competitive performance regarding precision as it relies on the estimated propensity score which is correctly specified in this setting. However, the confidence intervals of CBPS are much wider. This can be partly because, at the balancing stage, CBPS balances all covariates, not just the confounders, which brings more uncertainty. Given the poor performance of DOPE, it is excluded from the plot.

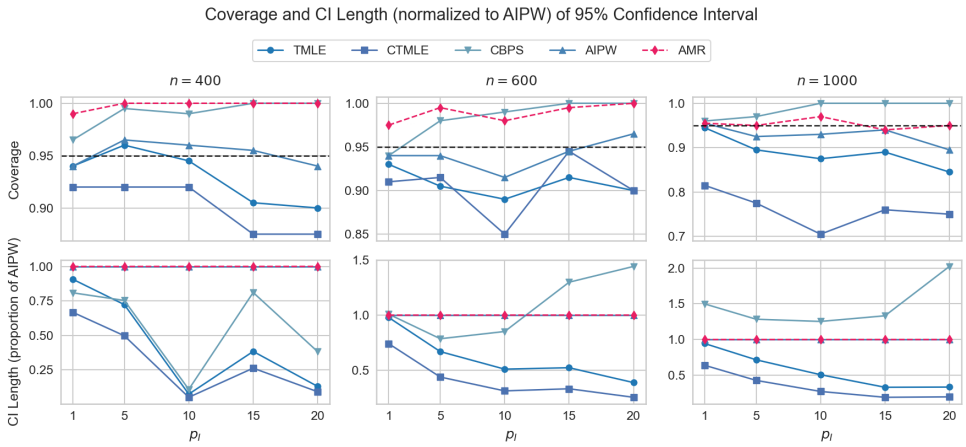


Figure 3.4: Coverage and length of confidence interval at nominal level of 95% on synthetic data.

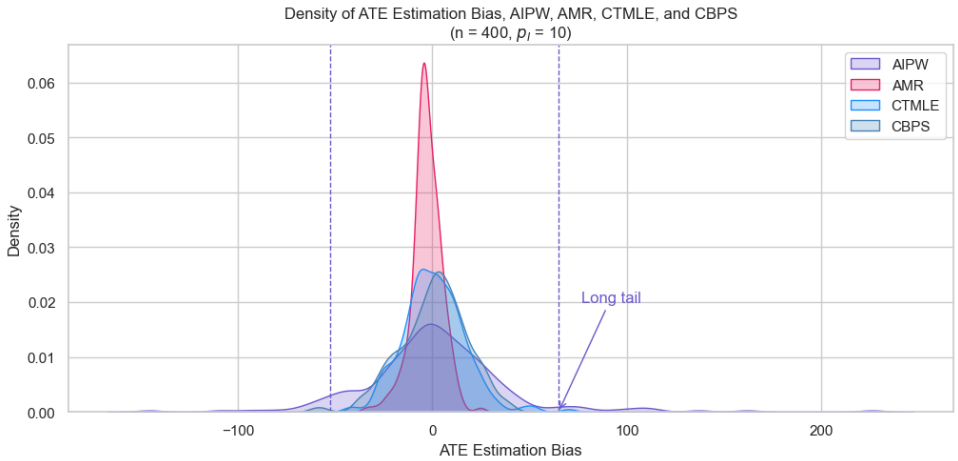


Figure 3.5: Empirical densities of estimation errors $\hat{\theta} - \theta$ for AIPW, AMR, CTMLE, and CBPS across $R = 200$ replicates on synthetic data, with $n = 400$ and $p_I = 10$.

As discussed in Section 3.4, the AMR confidence interval constructed using (3.13) is conservative and of similar width to that of $\hat{\theta}_{AIPW}$, but it is centred at $\hat{\theta}_{AMR}$, which is substantially more stable than $\hat{\theta}_{AIPW}$ in our simulations under weak overlap. When we shrink the nominal coverage to 60% (Figure 3.6), the AIPW interval under-covers while the AMR interval continues to attain its nominal level.

3.5.2 Application on NHANES data

The National Health and Nutrition Examination Survey (NHANES) is a U.S. nationwide program collecting health and nutritional information from a representative sample of the population. We use the mortality dataset collected by Lundberg et al. (2020), restricting to participants aged 40 to 65 and studying the association, interpreted under the measured-confounding as-

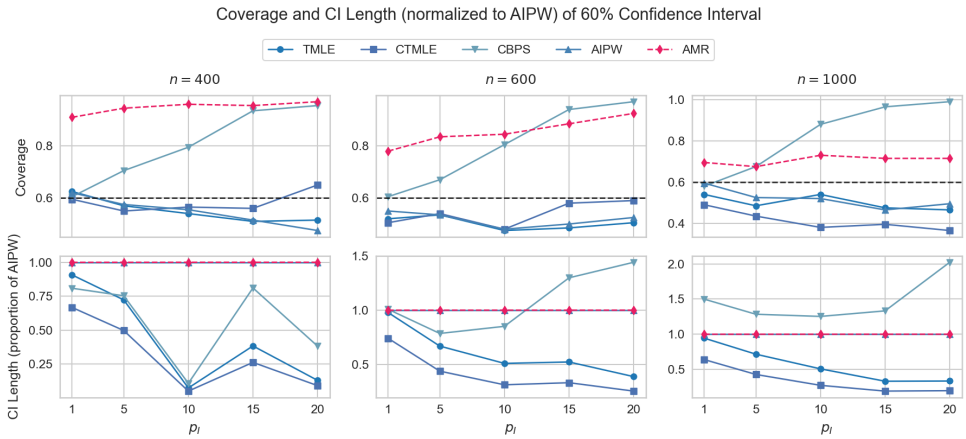


Figure 3.6: Coverage and length of confidence interval at nominal level of 60% on synthetic data.

sumption, between obesity, defined as body mass index (BMI) at least 30, and total cholesterol. We dichotomise BMI at 30 to obtain a binary treatment indicator $X = \mathbb{1}\{\text{BMI} \geq 30\}$; total cholesterol is the continuous outcome Y. Higher BMI has previously been linked with elevated total cholesterol (Freedman et al., 2001; Wormser et al., 2011; Flegal et al., 2016); the analysis here remains observational and conclusions depend on the validity of the measured-confounding assumption.

We follow the preprocessing steps in Lundberg et al. (2020) and Christgau and Hansen (2024), dropping covariates with more than 50% missing values. After restricting to participants aged 40–65, $n = 5985$ observations remain for analysis. To address potential confounding in this observational analysis, we adjust for an extensive set of covariates (61 in total) that capture a broad spectrum of health indicators. These include demographic and lifestyle factors such as gender, age and physical activity, and laboratory biomarkers including serum albumin, haemoglobin, etc.

The propensity score is estimated using logistic regression. We present the density of the propensity score estimates in Figure 3.7, which suggests weak overlap: a non-trivial fraction of observations have estimated treatment probability below 0.1. For comparison, we estimate the potential outcomes using two different models: a feedforward neural network (denoted NN) and linear regression (denoted LR) to see how sensitive each method is to the specific potential outcome model. The same set of estimators as in Section 3.5.1 is compared here. We provide estimates and confidence intervals, computed using Proposition 3.4.8, in Table 3.2.

All methods produce positive point estimates; only CBPS, CTMLE with potential outcome

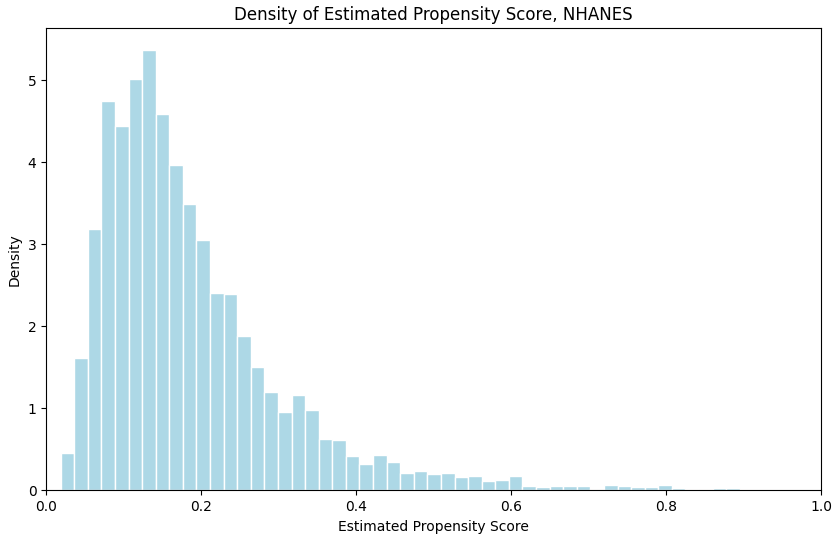


Figure 3.7: Density of estimated propensity score of exposure $X = \mathbb{1}\{\text{BMI} \geq 30\}$ on NHANES data.

Method	Estimation	95% CI
CBPS	4.70	$[0.426, 8.98]$
DOPE	230	$[-201, 662]$
CTMLE (NN)	2.59	$[-0.511, 5.68]$
CTMLE (LR)	3.90	$[0.815, 6.98]$
TMLE (NN)	3.41	$[-0.0175, 6.84]$
TMLE (LR)	3.15	$[-0.315, 6.62]$
AIPW (NN)	3.42	$[-0.00461, 6.85]$
AIPW (LR)	3.14	$[-0.326, 6.62]$
AMR (NN)	3.44	$[0.0114, 6.86]$
AMR (LR)	3.56	$[0.0906, 7.03]$

Table 3.2: ATE estimates and 95% confidence intervals on the NHANES data. For estimators requiring an outcome regression, both a feedforward neural network (NN) and linear regression (LR) are reported. IPW and MR are omitted from the table because confidence intervals are not computed for them; their point estimates are 3.90 and 4.70, respectively.

estimated by linear regression, and AMR under both potential outcome models have 95% confidence intervals whose lower endpoints lie above 0. In this analysis AMR also produces the smallest discrepancy between the linear regression and neural network outcome fits; this may reflect that AMR conditions on an outcome residual rather than directly on the full covariate vector.

3.6 Discussion

We propose AMR and revisit MR, two outcome-informed weighting estimators for the ATE. The key operation is a one-dimensional projection of the inverse propensity component onto the outcome, or its residual for AMR, which preserves the AIPW population identity through a bridge identity and reduces the variance of inverse propensity score weighting. Empirically, AMR is substantially more stable than AIPW under weak overlap; the feasible estimator can be sensitive to the quality of the outcome-residual regression when $\hat{\mu}^x$ is severely misspecified.

The same projection construction can be adapted to other causal estimands, including the Average Treatment Effect on the Treated (ATT), the Average Treatment Effect on the Control (ATC), and the value of a target policy as in Taufiq et al. (2024). Each extension requires a target-specific augmentation term and, for ATT and ATC, normalisation by the corresponding treatment-group probability; we leave their formal development to future work. In Section 3.3.3, we show how MR connects to efficient covariate adjustment. A comparable connection for AMR is less direct because the residual Y^*, the variable being conditioned on, depends on the propensity score; clarifying this connection is left for future work. We also considered a standard nonparametric bootstrap for $\hat{\theta}_{AMR}$. In these experiments the resulting intervals were substantially wider than the conservative interval of Proposition 3.4.8; one possible explanation is that resampling changes the empirical degree of overlap, which is already weak in these settings. Constructing shorter intervals with reliable coverage remains an open problem.

Chapter 4

Stable data fusion via equivalence testing

Chapter 3 discusses stable estimation from a single data source. This chapter turns to the fusion of several sources: pooling an external control sample with the current control can raise the power of the subsequent causal test, but the classic test-then-pool procedure can inflate its Type-I error by pooling an external control that is in fact incompatible with the current one. We propose an equivalence test to decide whether to fuse the data and a partial bootstrap or partial permutation for the causal test, and prove that the overall procedure controls the Type-I error and is consistent, with potential power gains. This chapter is based on joint work with Dr. Xing Liu and Professor Robin J. Evans (Yang et al., 2026).

4.1 Introduction

A central objective in causal inference is to determine whether an intervention produces a measurable effect in a target population. Although randomised controlled trials (RCTs) are the gold standard, they can be time-consuming and sometimes face ethical or feasibility constraints. Such challenges can usually affect the recruitment of patients into the placebo arm, inflating the variances of estimation and reducing the test power. At the same time, substantial amounts of clinical data are often available prior to the start of a study, particularly for control arms of similar trials (Viele et al., 2014). Thus, a common remedy is to supplement current data with individual patient data from earlier control arms, which would reduce costs, shorten trial duration, and ease recruitment (Schmidli et al., 2014; Galwey, 2017; Schmidli et al., 2020). This follows the valid incentive that, despite each trial's unique characteristics, multiple previous trials often share similar control treatments (Galwey, 2017).

Nevertheless, borrowing historical controls without accounting for potential discrepancies

between studies can introduce bias. Differences such as regional bias or assessment bias can be problematic (Burger et al., 2021). To address this, various methods have been proposed to examine the homogeneity before combining historical or external control data with current control arms in settings where fully powered randomised controlled trials are not feasible, including Bayesian approaches such as meta-analytic-predictive priors (Schmidli et al., 2014), frequentist techniques such as matching (Busgang et al., 2022) and hybrid strategies like Bayesian dynamic borrowing (Viele et al., 2014). We refer to Viele et al. (2014), Galwey (2017) and Kojima (2023) for the review of these methods.

Another mainstream approach is the test-then-pool (TTP) approach (Viele et al., 2014; Meng et al., 2020; Gao and Yang, 2023). This approach proceeds in two stages: a fusion test and a causality test. The fusion test evaluates whether historical and current data are compatible. If there is insufficient evidence to reject compatibility, the historical data are pooled with the current data, and a causal analysis is performed on the combined sample.

While relatively straightforward, a major criticism of the classic TTP is its limited power in identifying heterogeneity between historical and current controls. As a result, heterogeneous data may be inappropriately pooled, introducing additional bias, and the probability of such happening is *not* controlled by a test targeting the standard point hypotheses.

To mitigate this problem, Li et al. (2020) replaces the fusion test with an equivalence test. This modification provides the Type-I error control when concluding that the two controls are similar. However, since "similar" does not imply "identical", the subsequent causality test requires an appropriate adjustment to account for the differences between the control arms to guarantee the Type-I error control for the overall procedure. Such an adjustment was not explicitly derived or formally specified in their framework. Moreover, existing TTP methods remain limited to testing the similarity via mean, thereby ignoring the richer information contained in the full distributions. A detailed review of these approaches is provided in Section 4.2.3.

In this chapter, we propose a new TTP framework for fusing control arms, aimed at increasing the power to detect *distributional treatment effects* (DTE) while rigorously controlling the Type-I error rate. Our method addresses a common critique of classic TTP procedures that they can inflate Type-I errors by providing a formal guarantee of validity for the final treatment effect test. Unlike traditional mean-based equivalence tests, our approach captures the full distributional differences between arms via the *maximum mean discrepancy* (MMD), offering a more comprehensive evaluation of treatment effect heterogeneity. We further provide

practical guidance for choosing the equivalence radius to achieve desired power levels.

Our key contributions include:

- Extension of TTP to distributional testing, allowing detection of treatment effects beyond mean differences and enabling broader applications.
- Partial bootstrap and partial permutation procedures for valid approximation of the null distribution when fused control arms contain differences.
- Equivalence-based fusion test using MMD, enabling principled pooling decisions.
- Overall TTP validity proof, establishing that the treatment effect test after fusion retains the nominal Type-I error.

In Section 4.2, we provide background review on MMD and two-sample testing based on it, as well as existing TTP procedures. The proposed TTP method is detailed in Section 4.3. In Section 4.4 we present synthetic experiment results, evaluating the performance of the proposed TTP on Type-I error control and demonstrating the effect of different hyperparameter choices. We present the application on a program, *Prospera*, in Section 4.5. We conclude the chapter with potential directions in Section 4.6.

4.2 Background

Throughout this chapter we work with outcome distributions and their samples, with individual observations written y.[8] Let Q_c, Q_h, Q_t be probability measures on $\mathbb{R}$, denoting the outcome distributions of the current control, historical control, and treatment groups. We observe i.i.d. random variables $\{y_i^c\}_{i=1}^m \sim Q_c$, $\{y_i^h\}_{i=1}^\ell \sim Q_h$, and $\{y_i^t\}_{i=1}^n \sim Q_t$. For any distribution Q, its empirical distribution based on i.i.d. samples $\{y_i\}_{i=1}^n \sim Q$ is $Q^n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, where δ_y is the Dirac measure at y. In particular, $Q_c^m = \frac{1}{m}\sum_{i=1}^m \delta_{y_i^c}$, and Q_t^n, Q_h^ℓ are defined analogously. Our main goal is to test for a statistically significant difference between the treatment group Q_t and the control group Q_c, possibly after augmenting by the historical control Q_h. This is phrased as testing the DTE by setting $\mathcal{H}_0 : Q_c = Q_t$. We refer to it as the *causality* test. For two sets of observations, $\{y_i^c\}_{i=1}^m \sim Q_c$ and $\{y_i^h\}_{i=1}^\ell \sim Q_h$, the empirical distribution of their pooled sample, $\{y_i^c\}_{i=1}^m \bigcup \{y_i^h\}_{i=1}^\ell$, is denoted $Q_f^{m+\ell} = \frac{m}{m+\ell}Q_c^m + \frac{\ell}{m+\ell}Q_h^\ell$.

[8] Observations are written y, matching the outcome of Section 2.5; the capitals X, Y, Z inside expectations are generic integration variables, each a draw from the distribution indicated.

4.2.1 Distributional treatment effect

While the ATE is the most widely studied causal estimand, it summarises impact only through differences in means. This focus can obscure important heterogeneity: two treatments may have the same mean effect but very different impacts on the distribution of outcomes. Thus, ATE provides only limited insights. In contrast, DTEs address this limitation by characterising how an intervention shifts the entire outcome distribution, rather than just its first moment. Increasing attention is being paid to distributional treatment effects, which offer more comprehensive information beyond the mean (Abadie, 2002; Kennedy et al., 2023).

A growing body of work develops nonparametric methods for estimating DTEs. One strand leverages kernel embeddings of probability distributions, which represent distributions in a reproducing kernel Hilbert space (RKHS). This approach allows treatment effects to be evaluated through differences between embedded distributions, making it possible to capture complex shifts without imposing strong parametric assumptions. Kernel-based methods also provide a natural way to compare entire distributions, and can be integrated with flexible machine learning models (Muandet et al., 2017, 2021; Park et al., 2021).

By moving beyond comparing only the first moments, estimation of this kind offers a more flexible and informative framework, particularly in settings where heterogeneity or tail behaviour is crucial, or when the outcomes are high-dimensional.

Remark 4.2.1 (Connection with ATE)**.** *An appealing feature of the proposed kernel two-sample testing framework is its flexibility. By choosing a characteristic kernel, the method captures full distributional differences and thus enables inference on distributional treatment effects. At the same time, if we choose the linear kernel $k(x, y) = x^\top y$, the framework reduces to a test for mean differences, recovering the classical average treatment effect. The same construction therefore covers both ATE and DTE inference, according to the choice of kernel.*

4.2.2 MMD and two-sample testing

Our TTP framework requires a notion of dissimilarity that quantifies distributional differences. We use the maximum mean discrepancy (MMD; Gretton et al., 2006, 2012), which we review next.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. Denote by $\mathcal{H}_k$ its associated RKHS (Berlinet and Thomas-Agnan, 2011), equipped with the RKHS inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$. The MMD between

two distributions Q and P is the maximal difference between their expectations of all functions in the unit ball of $\mathcal{H}_k$, defined as

$$D(Q,P) \coloneqq \sup_{f\in\mathcal{H}_k:\|f\|_{\mathcal{H}_k}\leq 1} \left|\mathbb{E}_{X\sim Q} f(X) - \mathbb{E}_{Y\sim P} f(Y)\right|,$$

where $\|f\|_{\mathcal{H}_k} = \langle f, f\rangle_{\mathcal{H}_k}^{1/2}$ is the induced RKHS norm. Under regularity conditions, the MMD can be rewritten as the RKHS-norm of the difference between the so-called *kernel mean embeddings* of the two distributions Q and P. For a probability measure R, we define its kernel mean embedding as $\xi_R \coloneqq \mathbb{E}_{X\sim R}\, k(\cdot, X)$.[9] Throughout, we will assume the kernel k satisfies the following condition, which guarantees the existence of ξ_Q, ξ_P (Gretton et al., 2012, Lemma 3).

Assumption 4.2.2 (Finite second moment)**.** *The kernel k satisfies $\mathbb{E}_{X\sim R} k(X,X) < \infty$, for both $R = Q$ and $R = P$.*

Under Assumption 4.2.2, the *squared* MMD admits the following equivalent forms (Gretton et al., 2012, Lemma 4, 6)

$$D^2(Q,P) = \|\xi_Q - \xi_P\|_{\mathcal{H}_k}^2 = \mathbb{E}_{X,X'\sim Q} k(X,X') + \mathbb{E}_{Y,Y'\sim P} k(Y,Y') - 2\mathbb{E}_{X\sim Q, Y\sim P} k(X,Y).$$

We further assume that k is *characteristic* (Sriperumbudur et al., 2011) unless stated otherwise, meaning that the mapping $Q \mapsto \xi_Q$ is injective. In this case, the associated MMD is a statistical metric. In particular, MMD separates distributions, meaning that $D(Q,P) = 0$ if and only if $Q = P$ (Gretton et al., 2012, Theorem 5). Most common kernels, such as Gaussian kernels, inverse multi-quadric (IMQ) kernels, and Matérn kernels, are characteristic. The former two kernels also satisfy Assumption 4.2.2 for all Q, P, since they are bounded.

The above identities also suggest a natural finite-sample estimator for MMD using Monte-Carlo estimation. Given random samples $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$ drawn independently from Q and P, respectively, the MMD can be estimated by the following biased but consistent estimator, formed by replacing the expectations with empirical averages

$$D^2(Q^n, P^m) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(X_i,X_j) + \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(Y_i,Y_j) - \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m k(X_i,Y_j). \tag{4.1}$$

It can be shown that (4.1) can be rewritten as a *two-sample V-statistic of degree* $(2,2)$ (Gretton

[9] Within this chapter and Chapter 6, ξ denotes the kernel mean embedding; elsewhere it is the nuisance parameter of Chapter 3 and the exogenous noise of Chapter 5. See Table 2.1.

et al., 2012; Shekhar et al., 2022), whose asymptotic properties are well-known (Serfling, 2009, Chapter 5).

Remark 4.2.3 (U-statistics)**.** *An alternative estimator in the literature is the following estimator*

$$U_{nm} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} k(X_i, X_j) + \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j \neq i}}^{m} k(Y_i, Y_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(X_i, Y_j). \quad (4.2)$$

Similarly to (4.1)*, the estimator above can be rewritten as a two-sample U-statistic (Hoeffding, 1948). U-statistics share similar statistical properties as V-statistics and have the advantage of being unbiased (Serfling, 2009, Chapter 5). For these reasons,* (4.2) *is the more commonly used estimator in the MMD literature. However, U-statistics can take negative values, whereas V-statistics are always non-negative. As will become clear in Section 4.3.1, non-negativity is essential for the fusion test used in our proposed method. Therefore, we will focus on V-statistics in this work, and defer a discussion on extensions to U-statistics to Appendix B.4.*

MMD has been widely applied in two-sample testing for the homogeneity of two distributions (Gretton et al., 2006, 2012; Schrab et al., 2023; Shekhar et al., 2022; Chau et al., 2024). To test the null hypothesis $\mathcal{H}_0 : Q = P$, MMD checks first the separation property to reformulate the null hypothesis to the equivalent form $\mathcal{H}_0 : D(Q, P) = 0$. It then rejects the null hypothesis if the estimate $D(Q^n, P^m)$ exceeds a critical value, which is the $(1-\alpha)$-quantile of the distribution of $D(Q^n, P^m)$ under $\mathcal{H}_0$, for a prescribed significance level α.

One major challenge in MMD tests is that the critical value is typically intractable and thus needs to be estimated. Common approaches include permutation and bootstrapping, of which both aim to construct empirical samples from the null distribution. The permutation method proceeds by randomly permuting the pooled sample $\{Z_\ell\}_{\ell=1}^{n+m} = \{X_i\} \cup \{Y_j\}$ to form the permuted sample $\{Z_\ell^\pi\}_{\ell=1}^{n+m}$, and computing the permuted MMD statistic

$$T_\pi := D^2\Big(\frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i^\pi}, \frac{1}{m}\sum_{i=n+1}^{n+m} \delta_{Z_i^\pi}\Big). \quad (4.3)$$

Repeating this over B random permuted samples yields the empirical quantity $\{T_{\pi^{(b)}}\}_{b=1}^{B}$. The critical value $\hat{q}_{1-\alpha}^{m,n}$ is then given by the Monte Carlo estimate based on this sample.

Alternatively, the critical value can be estimated using bootstrapping approaches, with the most common method being wild bootstrapping (Shao, 2010; Schrab et al., 2023), which

constructs the bootstrap samples using Rademacher weights. In our work, we instead follow Liu and Gandy (2026) to use Efron's bootstrapping, where the bootstrapped statistics are formed by drawing samples from $\{Z_l\}_{l=1}^{n+m}$ with replacement. The critical value is then the $(1-\alpha)$-quantile based on the bootstrapped statistics. As argued in Liu and Gandy (2026, Section 2.2), Efron's bootstrapping gives similar empirical performance while allowing existing theoretical tools to be used to prove statistical properties for their equivalence MMD test.

4.2.3 Existing mean-based TTP

We provide a more detailed review of existing TTP methods. As introduced in Section 4.1, TTP is a general two-stage framework consisting of a fusion test and a causality test. The fusion test determines whether a historical control group with outcome distribution Q_h is sufficiently compatible with the current control group Q_c to justify pooling. If compatibility is not rejected, the external and current controls are merged, and the resulting pooled control is compared with the treatment group Q_t in the causal stage.

The classic TTP does so by performing a statistical test targeting $\check{\mathcal{H}}_0^f : \mu_c = \mu_h$, where μ_c, μ_h are the means of Q_c, Q_h, respectively. If the fusion test fails to be rejected, then the external control group is merged with the current control group to form a pooled group. The causality test then compares the resulting control group, either alone or pooled, with the treatment group, Q_t, to assess the presence of a treatment effect. This is formalised by performing the causality test targeting $\check{\mathcal{H}}_0 : \mu_c = \mu_t$.

While this approach is simple, it is often insufficient to conclude equivalence between the control groups. Indeed, failure to reject $\check{\mathcal{H}}_0^f$ does not imply that μ_c and μ_h are equivalent; instead, it can simply be a result of lack of statistical power. Consequently, the TTP framework can incorrectly merge heterogeneous historical control arms into the causal analysis, especially when the current control size m is small relative to the treatment size n. This, in turn, can then lead to inflated Type-I error in the causality test (Viele et al., 2014).

To address this limitation, Li et al. (2020) proposed to replace the fusion test with an equivalence test, targeting instead the null hypothesis $\check{\mathcal{H}}_0^f : |\mu_c - \mu_h| \geq \delta$ for some pre-specified margin $\delta > 0$. Equivalence testing is more natural for assessing the similarity between the control arms, because it provides probabilistic guarantee on the error rate of falsely merging heterogeneous historical controls. However, this equivalence-based TTP test still relies solely on *mean* comparisons and does not account for differences in the broader outcome distributions,

which can be important when the goal is to assess effects beyond the mean.

4.2.4 Warm-up: distributional TTP

The TTP framework is general in nature: any statistical test can be used in the fusion stage. To align with the classic TTP procedure and facilitate comparison, we extend it here from mean-based to distribution-based testing, allowing the framework to target the DTE instead of the ATE. This formulation is not proposed in prior work; rather, we introduce it to formalise a "classic TTP" analogue at the distributional level, which will serve as a conceptual baseline and benchmark for our later developments. Specifically, we define the causal null hypothesis as

$$\mathcal{H}_0 : Q_c = Q_t \quad \text{against} \quad \mathcal{H}_1 : Q_c \neq Q_t.$$

When using an MMD two-sample test with a characteristic kernel, this is equivalent to

$$\mathcal{H}_0 : D(Q_c, Q_t) = 0 \quad \text{against} \quad \mathcal{H}_1 : D(Q_c, Q_t) > 0.$$

Without merging from external control group, the causality test uses merely Q_c^m and Q_t^n. When one sample size is small (here we assume $m \ll n$), the test may have insufficient power to detect differences in treatment effects. This motivates fusing the current control Q_c^m with an external or historical control, whose response distribution Q_h is sufficiently similar to Q_c.

The fusion test assesses

$$\tilde{\mathcal{H}}_0^f : Q_c = Q_h \quad \text{against} \quad \tilde{\mathcal{H}}_1^f : Q_c \neq Q_h.$$

or equivalently,

$$\tilde{\mathcal{H}}_0^f : D(Q_h, Q_c) = 0 \quad \text{against} \quad \tilde{\mathcal{H}}_1^f : D(Q_h, Q_c) > 0. \tag{4.4}$$

The corresponding test function for the fusion stage is

$$\phi_f^0(Q_c^m, Q_h^\ell) = \mathbb{1}\Big\{ D^2(Q_c^m, Q_h^\ell) > \hat{q}_{1-\alpha_f}^{m,\ell} \Big\}, \tag{4.5}$$

where α_f is the prescribed significance level. Here, $\hat{q}_{1-\alpha_f}^{m,\ell}$ is the permutation-based $(1-\alpha_f)$-quantile of the test statistic.

The intuition is straightforward: if $\tilde{\mathcal{H}}_0^f$ is rejected, Q_h^ℓ is excluded from pooling, and the causality test remains to be $\phi^0(Q_c^m, Q_t^n)$. Otherwise, the fused control-arm response distribution is $Q_f^{m+\ell} = \frac{m}{m+\ell}Q_c^m + \frac{\ell}{m+\ell}Q_h^\ell$, which is then compared to Q_t^n in the causality testing stage using an MMD test based on $D^2(Q_f^{m+\ell}, Q_t^n)$.

We denote the overall TTP decision rule by $\phi^0(Q_c^m, Q_h^\ell, Q_t^n) : \mathbb{R}^m \times \mathbb{R}^\ell \times \mathbb{R}^n \to \{0, 1\}$, where $\phi^0 = 1$ indicates rejection of $\mathcal{H}_0$. The function ϕ^0 is defined conditional on the fusion stage test function ϕ_f^0, written as

$$\phi^0(Q_c^m, Q_h^\ell, Q_t^n) = \begin{cases} \mathbb{1}\left\{D^2(Q_c^m, Q_t^n) > \hat{q}_{1-\alpha}^{m,n}\right\}, & \phi_f^0(Q_c^m, Q_h^\ell) = 1, \\ \mathbb{1}\left\{D^2(Q_f^{m+\ell}, Q_t^n) > \hat{q}_{1-\alpha}^{m+\ell,n}\right\}, & \phi_f^0(Q_c^m, Q_h^\ell) = 0. \end{cases}$$

In both cases, the causality test decision at significance level α is made via the permutation-based test.

This formulation generalises the classical TTP to a distributional setting. However, the fundamental issue remains: the fusion test may be underpowered when the current control sample size m is small, leading to an incorrect merge and inflated Type-I error in the causal stage. This limitation motivates the development of our proposed method.

4.3 Method

To address the aforementioned problems, we propose a new TTP procedure in this chapter. For the fusion stage, we replace the classic two-sample test with an equivalence test. Conditional on the fusion test outcome, the causality test is conducted using either a traditional permutation test if the historical sample is not fused, or a partial bootstrap/permutation test otherwise. The complete procedure is given in Algorithm 4.1.

Unlike the equivalence-based TTP proposed by Li et al. (2020), our method flexibly detects treatment effects at the distributional level, is not restricted to normal distribution, and does not require the variance to be known. In addition, the proposed partial permutation and partial bootstrap procedures ensure the asymptotic validity and consistency of the entire framework, even when the fused control arms are not identical. These properties provide formal theoretical guarantees for Type-I error control of the overall framework, which were not established in Li et al. (2020).

Algorithm 4.1 Equivalence TTP

Input: Data Q_c^m, Q_h^ℓ and Q_t^n; equivalence radius $\rho > 0$; test levels α for the causality test and α_f for the fusion test.

1: Conduct equivalence test $\mathcal{H}_0^f : D(Q_h, Q_c) \geq \rho$ at level α_f.
2: **if** $\mathcal{H}_0^f$ is **not** rejected **then**
3: Conduct a permutation test between Q_c^m and Q_t^n for $\mathcal{H}_0$ at level α.
4: **else**
5: Form the pooled control sample

$$Q_f^{m+\ell} = \frac{m}{m+\ell} Q_c^m + \frac{\ell}{m+\ell} Q_h^\ell.$$

6: Conduct a partial bootstrap or partial permutation causality test between $Q_f^{m+\ell}$ and Q_t^n for $\mathcal{H}_0$ at level α.
7: **end if**

4.3.1 Fusion test: MMD equivalence testing

We introduce the fusion test in this section. We propose to test

$$\mathcal{H}_0^f : D(Q_c, Q_h) \geq \rho \quad \text{against} \quad \mathcal{H}_1^f : D(Q_c, Q_h) < \rho, \tag{4.6}$$

where $\rho > 0$ is the *equivalence radius*.

The benefits of testing the above hypotheses are two-fold. First, unlike (4.4) where the null hypothesis is that Q_c and Q_h are identical, we employ an equivalence testing setting, where the null hypothesis is that they differ by at least ρ in MMD. With this formulation, we reject only when there is statistical evidence for $D(Q_c, Q_h) < \rho$. This ensures control over the probability of an "incorrect merge", where an external control Q_h is distinct from Q_c but the fusion test still suggests merging (namely, not rejecting the $\mathcal{H}_0^f$ in (4.4)) due to insufficient power. Meanwhile, we use MMD as a notion of dissimilarity between Q_c and Q_h, which captures not only differences in their means, but in their entire distributions.

To test (4.6), we employ the MMD equivalence test proposed in Liu and Gandy (2026, Section 4.2). This test uses the test statistic

$$\Delta_{m,\ell}^f \coloneqq \rho - D(Q_c^m, Q_h^\ell),$$

and rejects $\mathcal{H}_0^f$ in (4.6) if $\Delta_{m,\ell}^f > q_{1-\alpha_f}^{m,\ell}$, where $q_{1-\alpha_f}^{m,\ell}$ is the $(1-\alpha_f)$-quantile of the distribution of $S_{m,\ell} \coloneqq D(Q_c^m, Q_c) + D(Q_h^\ell, Q_h)$.

The intuition behind $S_{m,\ell}$ is that, under $\mathcal{H}_0^f : D(Q_h, Q_c) \geq \rho$, it upper bounds $\Delta_{m,\ell}^f$ through a triangle inequality. It can also be shown that $S_{m,\ell}$ is close to 0 with high probability under $\mathcal{H}_0^f$. Hence, the MMD equivalence test rejects the equivalence null hypothesis if $\Delta_{m,\ell}^f$ is sufficiently large, or, equivalently, if $D(Q_c^m, Q_h^\ell)$ is sufficiently smaller than ρ. Using this intuition, Liu and Gandy (2026, Theorem 8) establishes the validity and consistency of this equivalence MMD test. We restate their result below with slight adaptations to suit our setting. Its proof is in Appendix B.2.1.

Theorem 4.3.1 (Calibration and consistency of MMD equivalence test)**.** *Let $\alpha_f \in (0,1)$. Denote by $q_{1-\alpha_f}^{m,\ell}$ the $(1-\alpha_f)$-quantile of the distribution of $S_{m,\ell} = D(Q_c^m, Q_c) + D(Q_h^\ell, Q_h)$. Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h$, and Assumption 4.3.2 is satisfied, so that $m, \ell \to \infty$ as $n \to \infty$. Then there exists $a \in (0, \alpha_f]$ such that*

$$\lim_{n\to\infty} \Pr(\Delta_{m,\ell}^f > q_{1-\alpha_f}^{m,\ell}) = \begin{cases} 0, & D(Q_h, Q_c) > \rho, \\ a, & D(Q_h, Q_c) = \rho, \\ 1, & D(Q_h, Q_c) < \rho. \end{cases}$$

However, a challenge is that $q_{1-\alpha_f}^{m,\ell}$ is often intractable, because the distribution of $S_{m,\ell}$ is typically unknown. Liu and Gandy (2026) proposed to use a bootstrap procedure to approximate this unknown quantile. Define

$$D_W^2(Q_c^m) \coloneqq \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m (W_i - 1)(W_j - 1) k(y_i, y_j),$$

where $W = (W_1, \ldots, W_m) \sim \text{Multinomial}(m; 1/m, \ldots, 1/m)$, and $D_W^2(Q_h^\ell)$ is defined analogously with weights drawn from $\text{Multinomial}(\ell; 1/\ell, \ldots, 1/\ell)$. We then define the bootstrap statistic

$$S_{m,\ell}^b \coloneqq D_{W^b}(Q_c^m) + D_{\widetilde{W}^b}(Q_h^\ell), \tag{4.7}$$

where $\{W^b\}_{b=1}^B$ and $\{\widetilde{W}^b\}_{b=1}^B$ are drawn independently from $\text{Multinomial}(m; 1/m, \ldots, 1/m)$ and $\text{Multinomial}(\ell; 1/\ell, \ldots, 1/\ell)$, respectively. Liu and Gandy (2026, Lemma 7) shows that $q_{1-\alpha_f}^{m,\ell}$ can be approximated by the $(1-\alpha_f)$-quantile of the distribution of the bootstrap sample $\{S_{m,\ell}^b\}_{b=1}^B$, namely

$$q^{m,\ell}_{1-\alpha_f,B} = \inf\Big\{u \in \mathbb{R} : \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{S^b_{m,\ell} \le u\} \ge 1-\alpha_f\Big\}. \tag{4.8}$$

To summarise, the MMD equivalence test rejects $\mathcal{H}_0^f$ if $\Delta^f_{m,\ell} > q^{m,\ell}_{1-\alpha_f,B}$. We will denote this test by $\phi_f : \mathbb{R}^m \times \mathbb{R}^\ell \to \{0,1\}$, where $\phi_f(Q_c^m, Q_h^\ell) = \mathbb{1}\Big\{\Delta^f_{m,\ell} > q^{m,\ell}_{1-\alpha_f,B}\Big\}$.

4.3.2 Causality test: partial bootstrap and partial permutation

Let $\phi_* : \mathbb{R}^m \times \mathbb{R}^\ell \times \mathbb{R}^n \to \{0,1\}$ denote the causality test for $\mathcal{H}_0 : D(Q_c, Q_t) = 0$, so that $\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1$ if $\mathcal{H}_0$ is rejected, and 0 otherwise. In particular, conditional on $\phi_f(Q_c^m, Q_h^\ell) = 0$, no merging is performed, and hence the causality test $\phi_*(Q_c^m, Q_h^\ell, Q_t^n)$ does *not* depend on Q_h^ℓ. When merging is performed, the control and historical samples are pooled into $Q_f^{m+\ell}$, and we write $\phi_*(Q_f^{m+\ell}, Q_t^n)$ for the same test applied to this merged sample.

One of the key steps in the causal stage is to approximate the null distribution of the test statistic under $\mathcal{H}_0$, so that we can estimate the critical value. Unlike traditional TTP, the Q_h^ℓ we decide to merge can be different from Q_c^m (by at most ρ). Consequently, the merged control-arm distribution $Q_f^{m+\ell}$ may also differ from Q_c^m. This creates a problem for directly applying a permutation test between $Q_f^{m+\ell}$ and Q_t^n, as such permutation approximates the distribution of $D^2(Q_f^{m+\ell}, Q_t^n)$ when $Q_f = Q_t$, where Q_f is the population-level mixture

$$Q_f = \frac{m}{m+\ell}Q_c + \frac{\ell}{m+\ell}Q_h.$$

The null distribution of $D^2(Q_f^{m+\ell}, Q_t^n)$ coincides with that under $Q_f = Q_c$ only when $Q_h = Q_c$. If $Q_h \neq Q_c$, even by a small margin, the direct permutation approach no longer yields the correct null distribution, leading to potential size distortion. We thus propose the following methods to mitigate this problem. For the convenience of the following discussion, we impose the assumption of fixed ratios:

Assumption 4.3.2 (Fixed sample ratios)**.** *The samples Q_c^m, Q_h^ℓ and Q_t^n are independent with $c_1 \coloneqq \frac{m}{n} \in (0,1]$ and $c_2 \coloneqq \frac{\ell}{n} \in (0,1]$, where c_1 and c_2 are constants.*

Remark 4.3.3 (Generalisation of fixed ratios)**.** *Although we assume fixed sample size ratios ℓ/n and m/n, the results below extend to cases where these ratios converge to finite limits, as long as the limits are greater than 0. The proofs remain valid under this relaxation by adding error terms controlling the convergence of ratios. We provide a simple example in Appendix B.2.11.*

Algorithm 4.2 Partial bootstrap.

Input: Data Q_c^m, Q_h^ℓ, Q_t^n; number of bootstrap replicates B; test level α.

1: **for** $b \leftarrow 1$ to B **do**
2: Bootstrap m, n samples from Q_c^m to get $Q_{c,b}^m$, $Q_{t,b}^n$.
3: Bootstrap ℓ samples from Q_h^ℓ to get $Q_{h,b}^\ell$.
4: Form $Q_{f,b}^{m+\ell} = \frac{m}{m+\ell} Q_{c,b}^m + \frac{\ell}{m+\ell} Q_{h,b}^\ell$.
5: Calculate $T_{n,m,\ell}^* \coloneqq D^2(Q_{t,b}^n, Q_{f,b}^{m+\ell})$ and $\widetilde{T}_{n,m,\ell}^* \coloneqq D^2(Q_{c,b}^m, Q_{f,b}^{m+\ell})$.
6: **end for**
7: **return** The empirical $(1-\alpha)$-quantile

$$\hat{q}_{1-\alpha,B}^{m+\ell,n} = \inf\left\{ u \in \mathbb{R} : \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{ \sqrt{n}\Big[D^2(Q_{t,b}^n, Q_{f,b}^{m+\ell}) - D^2(Q_{c,b}^m, Q_{f,b}^{m+\ell}) \Big] \le u \right\} \ge 1-\alpha \right\}. \tag{4.9}$$

4.3.2.1 Partial bootstrap

To approximate the distribution of $D^2(Q_f^{m+\ell}, Q_t^n)$ under $\mathcal{H}_0$ in the fused setting, we propose the *partial bootstrap* algorithm. This procedure resamples the current control and treatment groups from the same empirical source distribution Q_c^m under the null, while independently resampling the historical control group from Q_h^ℓ. This preserves the correct dependence structure of the fused control under $\mathcal{H}_0$ even when $Q_h \neq Q_c$. The details are provided in Algorithm 4.2.

The test statistic in partial bootstrap is

$$\Delta \coloneqq \sqrt{n}\Big(D^2(Q_f^{m+\ell}, Q_t^n) - D^2(Q_f^{m+\ell}, Q_c^m) \Big),$$

and we reject $\mathcal{H}_0$ at test level α when $\Delta > \hat{q}_{1-\alpha,B}^{m+\ell,n}$, with $\hat{q}_{1-\alpha,B}^{m+\ell,n}$ defined in (4.9).

For notational brevity, we write $T_{n,m,\ell} = D^2(Q_f^{m+\ell}, Q_t^n)$ and $\widetilde{T}_{n,m,\ell} = D^2(Q_f^{m+\ell}, Q_c^m)$, so that $\Delta = \sqrt{n}(T_{n,m,\ell} - \widetilde{T}_{n,m,\ell})$. Furthermore, let $T_{n,m,\ell}^* \coloneqq D^2(Q_{f,b}^{m+\ell}, Q_{t,b}^n)$ be the partial bootstrap statistic constructed as in Algorithm 4.2, and let $\widetilde{T}_{n,m,\ell}^* \coloneqq D^2(Q_{f,b}^{m+\ell}, Q_{c,b}^m)$ be its centering. In particular, $\widetilde{T}_{n,m,\ell}$ is "partially" bootstrapped because, although $Q_{f,b}^{m+\ell}$ is sampled with replacement from $Q_f^{m+\ell}$, the other sample $Q_{t,b}^n$ is bootstrapped from the current control arm Q_c^m, instead of the standard choice Q_t^n. The latter construction is crucial to ensure that the proposed bootstrapping procedure can approximate the null distribution of Δ, even if $Q_c \neq Q_h$. Let Pr denote probability under the data-generating distribution, and $\Pr^*$ the bootstrap probability conditional on the observed data $\mathcal{D}_{n,m,\ell} \coloneqq \{y_i^t\}_{i=1}^n \cup \{y_i^c\}_{i=1}^m \cup \{y_i^h\}_{i=1}^\ell$ that governs the resampling steps. Define

$$\Delta^* = \sqrt{n}(T^*_{n,m,\ell} - \widetilde{T}^*_{n,m,\ell}), \quad \text{and} \quad \sigma_c^2 \coloneqq \mathrm{Var}_{X\sim Q_c}\left[\mathbb{E}_{Y\sim Q_f} k(X,Y) - \mathbb{E}_{Z\sim Q_c} k(X,Z)\right].$$

Proposition 4.3.4 (Weak convergence of the partial bootstrap under non-identical control groups)**.** *Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, and Assumption 4.3.2 is satisfied. Under the causal null $\mathcal{H}_0 : Q_c = Q_t$ and assuming $\sigma_c^2 > 0$, we have*

$$\Delta \rightsquigarrow \mathcal{N}(0, 4(1 + 1/c_1)\sigma_c^2).$$

Moreover,

$$\sup_{z\in\mathbb{R}} \left|\Pr^*\{\Delta^* \le z\} - \Pr\{\Delta \le z\}\right| \xrightarrow{a.s.} 0.$$

We provide the proof in Appendix B.2.2. Note that the ratio $c_1 = m/n$ enters the asymptotic variance of Δ through the factor $1 + 1/c_1$: larger values of c_1 yield a less variable distribution. This matches the intuition that a larger proportion of current control observations reduces the sampling variability of the test statistic Δ. Meanwhile, Q_f and $Q_f^{m+\ell}$ affect the limiting distributions of Δ and Δ^* via σ_c^2 as shown in its definition.

With Proposition 4.3.4, we obtain the validity and consistency of partial bootstrap, presented in Theorems 4.3.5 and 4.3.6. Their proofs are deferred to Appendices B.2.3 and B.2.4, respectively.

Theorem 4.3.5 (Partial bootstrap validity under non-identical control groups)**.** *Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, and Assumption 4.3.2 is satisfied. Let $\hat{q}^{m+\ell,n}_{1-\alpha,B}$ be the partial bootstrap critical value defined in* (4.9)*. Under the causal null hypothesis $\mathcal{H}_0 : Q_c = Q_t$ and assuming $\sigma_c^2 > 0$, the test has asymptotic level α:*

$$\limsup_{n\to\infty} \lim_{B\to\infty} \Pr\Big\{\Delta > \hat{q}^{m+\ell,n}_{1-\alpha,B} \,\Big|\, \mathcal{H}_0\Big\} \le \alpha.$$

Theorem 4.3.5 requires $\sigma_c^2 > 0$. For a continuous control distribution under an analytic characteristic kernel, this holds whenever $Q_c \neq Q_h$; it can fail only in degenerate cases, for instance when $Q_c = Q_h$ or Q_c has finite support, where the limiting distribution is trivial. In such cases, Figure 4.2 suggests that the resulting critical values are greater than or equal to the true ones, so the test remains conservative but valid. In practice, the case $Q_h = Q_c$ is less common than $Q_t = Q_c$, since temporal shifts or changes in the environment or population typically separate

Q_h from Q_c. A theoretical analysis of this degenerate case remains an interesting direction for further investigation.

We next provide the consistency of the partial bootstrap test, irrespective of whether Q_c and Q_h are identical or not. Let $\gamma \coloneqq c_1/(c_1+c_2)$, which equals $m/(m+\ell)$ under Assumption 4.3.2. We denote $\cos\beta \coloneqq \langle \xi_h - \xi_c\,,\, \xi_t - \xi_c\rangle_{\mathcal{H}_k}/(\|\xi_h - \xi_c\|_{\mathcal{H}_k}\|\xi_t - \xi_c\|_{\mathcal{H}_k})$ when $Q_c \neq Q_h$.

Theorem 4.3.6 (Partial bootstrap test consistency)**.** *Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, and Assumption 4.3.2 is satisfied. Suppose $Q_t \neq Q_c$. Regardless of whether $Q_c = Q_h$, we have*

$$\lim_{n\to\infty}\lim_{B\to\infty}\Pr\Big\{\Delta > \hat{q}^{m+\ell,n}_{1-\alpha,B} \,|\, \mathcal{H}_1\Big\} = 1,$$

if and only if $Q_c = Q_h$, or $Q_c \neq Q_h$ and $2(1-\gamma)D(Q_h, Q_c)\cos\beta < D(Q_c, Q_t)$.

Note the consistency comes with conditions regarding Q_c and Q_h. We provide a geometric interpretation of these conditions in Section 4.3.2.4.

Remark 4.3.7. *In the later discussion, we assume $B \to \infty$, i.e., we consider the population bootstrap quantile $\hat{q}^{m+\ell,n}_{1-\alpha,\infty}$, so that we can drop $\lim_{B\to\infty}$ in Theorem 4.3.5 and Theorem 4.3.6 for brevity.*

4.3.2.2 Normal approximation

We introduce partial bootstrap to approximate the asymptotic distribution of the test statistic Δ. Since we have the form of limiting distribution of Δ in Proposition 4.3.4, it is straightforward to have an alternative approach to approximate the limiting normal distribution. We estimate the variance σ_c^2 with empirical observed samples:

$$\hat{\sigma}_c^2 = \frac{(1-\gamma)^2}{m-1}\sum_{i=1}^{m}\Big(\hat{g}_i - \frac{1}{m}\sum_{i=1}^{m}\hat{g}_i\Big)^2,$$
$$\text{where} \qquad \hat{g}_i = \frac{1}{\ell}\sum_{j=1}^{\ell}k(y_i^c, y_j^h) - \frac{1}{m-1}\sum_{\substack{j=1\\ j\neq i}}^{m}k(y_i^c, y_j^c).$$

We then take the $(1-\alpha)$-quantile, $\hat{q}'_{1-\alpha}$, from $\mathcal{N}(0, 4(1+1/c_1)\hat{\sigma}_c^2)$ as the critical value, and we reject $\mathcal{H}_0$ if $\Delta > \hat{q}'_{1-\alpha}$. Since we directly take the critical value by estimating the limiting distribution, the asymptotic validity and consistency of such normal approximation is straightforward. The proofs are similar to Appendix B.2.3 and Appendix B.2.4 excluding the discussions

associated with bootstrap. However, we still recommend partial bootstrap for better approximation of the null distribution, if computation cost is not a concern. Details are provided in Section 4.3.2.4.

4.3.2.3 Partial permutation

Alternatively, one may treat Q_h^ℓ as an ancillary sample entering only through the test statistic, and adapt the classical kernel two-sample test with a permutation procedure. We thus propose an alternative algorithm named partial permutation, which replaces the step 2 of Algorithm 4.2 by a permutation step:

1. Permute the pooled observations from Q_c^m and Q_t^n to obtain Q_{c,π^b}^m and Q_{t,π^b}^n, respectively.
2. Form the fused control $Q_{f,\pi^b}^{m+\ell} = \frac{m}{m+\ell} Q_{c,\pi^b}^m + \frac{\ell}{m+\ell} Q_h^\ell$.

Denote $T_{n,m,\ell}^b \coloneqq D^2\Big(Q_{f,\pi^b}^{m+\ell}, Q_{t,\pi^b}^n\Big)$. Similar to classic permutation test, we use $T_{n,m,\ell}$ as the test statistic, which is different from Δ used in the proposed partial bootstrap. The critical value $\hat{c}_{1-\alpha,B}^{m+\ell,n}$ is set to be

$$\hat{c}_{1-\alpha,B}^{m+\ell,n} = \inf\Big\{u : \frac{1}{B+1}\sum_{b=0}^{B} \mathbb{1}\{T_{n,m,\ell}^b \le u\} \ge 1-\alpha\Big\},$$

where $T_{n,m,\ell}^0 = T_{n,m,\ell}$.

In the next result, we show the validity and consistency results of partial permutation.

Theorem 4.3.8 (Partial permutation validity)**.** *Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, and Assumption 4.3.2 is satisfied. Under the null hypothesis $H_0 : Q_c = Q_t$,*

$$\Pr\Big\{T_{n,m,\ell} \ge \hat{c}_{1-\alpha,B}^{m+\ell,n}\Big\} \le \alpha.$$

The proof is provided in Appendix B.2.5. Note that in partial permutation, we no longer require Q_c and Q_h to be non-identical for the validity to hold.

Writing $\lambda \coloneqq 1/(1+c_1)$, which equals $n/(m+n)$ under Assumption 4.3.2, we obtain the consistency of partial permutation:

Theorem 4.3.9 (Partial permutation consistency)**.** *Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, and Assumption 4.3.2 is satisfied. When $Q_t \neq Q_c$, with the critical value*

$\hat{c}^{m+\ell,n}_{1-\alpha,B}$ *at the test level* α*, we have*

$$\lim_{n\to\infty} \Pr\Big\{T_{n,m,\ell} > \hat{c}^{m+\ell,n}_{1-\alpha,B} \,\Big|\, \mathcal{H}_1\Big\} = 1,$$

if and only if $Q_c = Q_h$*, or* $Q_c \neq Q_h$ *and* $2D(Q_h, Q_c)\cos\beta < \Big(\frac{1}{1-\gamma} + \lambda\Big)D(Q_t, Q_c)$.

The proof can be found in Appendix B.2.6.

Remark 4.3.10. *We show that the merged samples remain valid and consistent under* $\mathcal{H}_0$*, regardless of the difference between* Q_c *and* Q_h *in the partial permutation or bootstrap algorithms. However, empirically (Section 4.4) and theoretically, we find that merging* Q_h *distributions that are sufficiently close to* Q_c *improves both test consistency and power. Therefore, we choose a small* ρ *so that only* Q^ℓ_h *distributions closely matching* Q^m_c *are merged, making the fusion step essential.*

4.3.2.4 Comparison between partial bootstrap and partial permutation

The partial bootstrap and partial permutation procedures generate treatment- and control-arm samples under the causal null from the same empirical source distribution Q^m_c. The former samples from Q^m_c to form the bootstrap samples $Q^n_{t,b}$. This ensures that, under $\mathcal{H}_0$, the resampled $Q^n_{t,b}$ and the fused $Q^{m+\ell}_{f,b}$ have the correct joint distribution, so the empirical distribution of the test statistic converges to its true null distribution $Q_c = Q_t$ when $Q_c \neq Q_h$. The key distinction between partial bootstrap and partial permutation lies in what role Q^ℓ_h plays in them. When Q^m_c and Q^n_t are drawn from different distributions, treating Q^ℓ_h merely as an auxiliary sample can lead to a poor approximation of the null distribution of $T_{n,m,\ell}$. The issue arises because permutation also implicitly enforces exchangeability between Q^m_c and Q^ℓ_h , effectively breaking the natural distance between them. In doing so, the permutation step ignores the distributional gap that must be preserved under the null, resulting in inflated critical values and weaker consistency. By contrast, partial bootstrap maintains the correct dependence structure, leading to a more reliable approximation. A summary of the comparison is provided in Table 4.1.

Figure 4.1 shows a geometric interpretation of the results, which uses the definition that the kernel mean embedding maps a probability measure to a point in the RKHS $\mathcal{H}_k$ (Song et al., 2013). In our setting, the condition for consistency depends on $D(Q_h, Q_c)$. When $\cos\beta \leq 0$, there is no restriction. However, when $\cos\beta > 0$ specifically, we require the distance $D(Q_h, Q_c)$ to be relatively small, which is controllable via setting the value of the radius (margin) ρ.

	Partial bootstrap	Partial permutation
Validity	Asymptotic	Finite exact
Consistency	Requires $\xi_h = \xi_c$ or $2D(Q_h, Q_c)\cos\beta < \frac{1}{1-\gamma} D(Q_t, Q_c)$	Requires $\xi_h = \xi_c$ or $2D(Q_h, Q_c)\cos\beta < \left(\frac{1}{1-\gamma} + \lambda\right) D(Q_t, Q_c)$
Degeneracy ($Q_t = Q_c$)	✓(if $\sigma_c^2 > 0$)	✓
Null distribution approximation	✓	Fail when $Q_t \neq Q_c$

Table 4.1: Comparison between partial bootstrap and partial permutation.

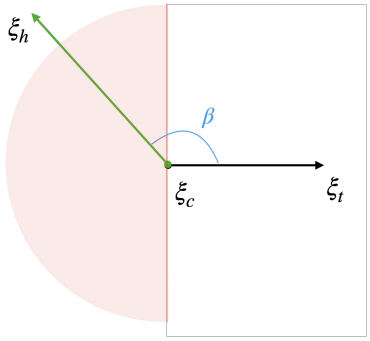


Figure 4.1: Geometric explanation of $\cos\beta$. The region with $\cos\beta \leq 0$ is pink.

Remark 4.3.11 (Conditional consistency)**.** *The consistency of partial permutation and partial bootstrap holds only under certain conditions involving the unknown quantities $D(Q_h, Q_c)$, $D(Q_t, Q_c)$, and $\cos\beta$. Since these quantities are not observable in practice, it is unclear whether the conditions are satisfied. A conservative check can be performed by estimating $D(Q_h, Q_c)$ and $D(Q_t, Q_c)$ and taking the largest plausible value $\cos\beta = 1$. Bounding $\cos\beta$ by 1 in Theorem 4.3.6 yields a sufficient condition for consistency: $2(1-\gamma)D(Q_h, Q_c) < D(Q_c, Q_t)$. This condition provides intuition about the family of consistent alternatives: the causality test $\mathcal{H}_0$ is consistently rejected if Q_t is sufficiently "further away" (more different) from Q_c than the merged distribution Q_h by a factor of $2(1-\gamma)$, reflecting the cost of merging samples from potentially different distributions. In particular, when $Q_h = Q_c$, no extra bias is introduced, and the test remains consistent against all Q_t. In practice, a more intuitive approach is to choose a relatively small margin ρ that upper bounds $D(Q_h, Q_c)$ for the fused control; when γ is relatively large, a larger ρ may be permissible.*

Figure 4.2 illustrates the true null distribution of the test statistic and its approximations using either partial bootstrap or partial permutation under different data generating processes. We consider the setting where $Q_c = \mathcal{N}(0, 1)$. When $Q_t \neq Q_c$, we set $Q_t = \mathcal{N}(-1, 1)$. When $Q_c \neq Q_h$, we set $Q_h = \mathcal{N}(2, 1)$. We use sample sizes $n = 600$ and $m = \ell = 300$. The MMD is

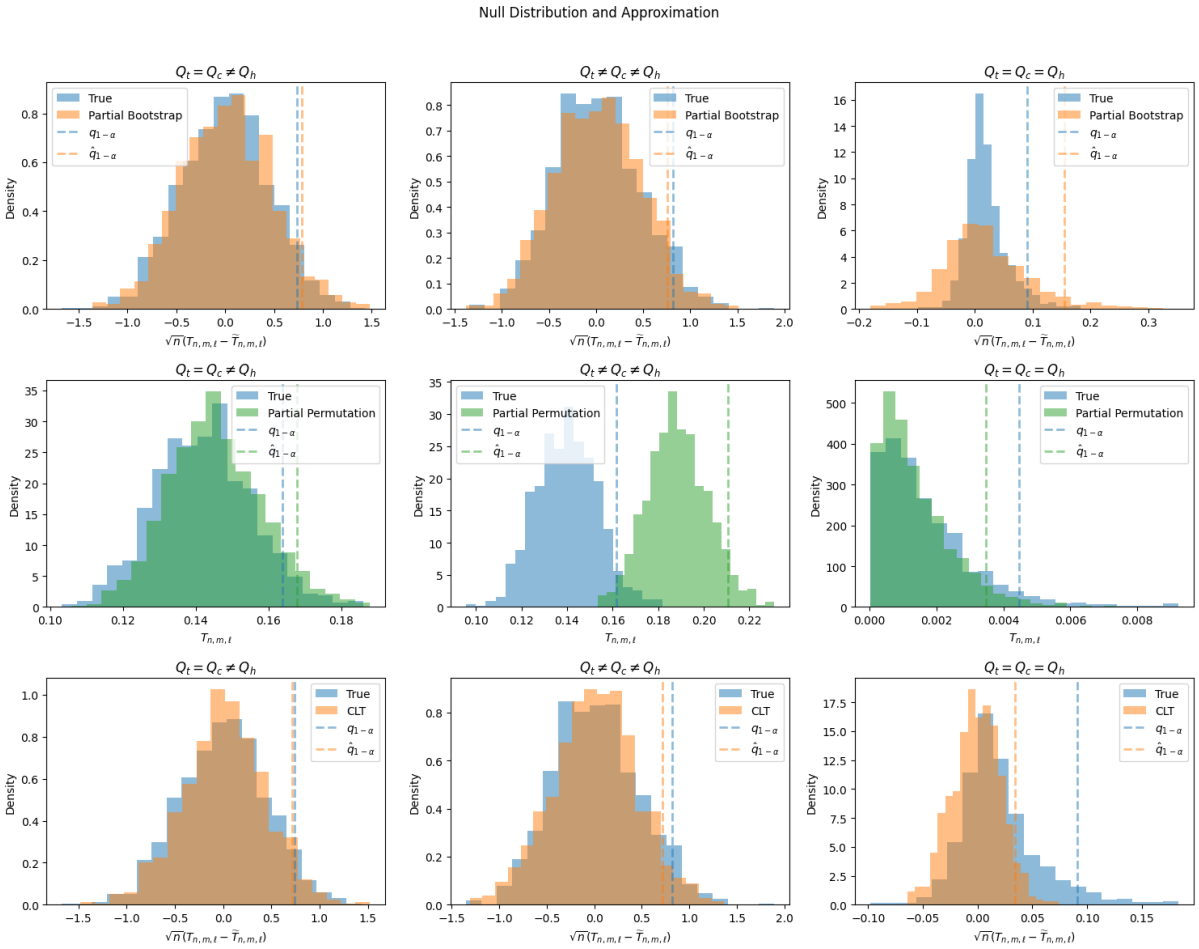


Figure 4.2: Comparison of the true and approximate distributions of the test statistic under partial bootstrap (first row), partial permutation (middle row) and normal approximation (last row). Results from 1000 simulations. Within each simulation, $B = 1000$ bootstraps/permutations are performed.

computed using the RBF kernel with the bandwidth chosen by the median heuristic, defined in (4.17). The middle plot of the second row highlights the weaker power of partial permutation, which produces unnecessarily large critical values when $Q_t \neq Q_c$.

The last row of Figure 4.2 shows the distribution obtained via the normal approximation. While the approximation is valid, it performs slightly worse than the partial bootstrap, as the asymptotic distribution becomes accurate only when sample sizes are sufficiently large. We provide more experiment results with n growing in Appendix B.3. In contrast, partial bootstrap achieves a much closer approximation to the true distribution of the test statistic. This comparison highlights a key distinction: while consistency and validity are addressed in the asymptotic behaviour, in finite samples partial bootstrap can deliver substantially better power than the normal approximation.

4.3.3 The overall procedure

Having introduced the proposed fusion and causality tests, we now show the validity of the overall TTP testing procedure.

The following result shows that the overall equivalence TTP testing procedure is well-calibrated, so long as the causality test ϕ_* and the fusion test ϕ_f are "well-behaved". Its proof is in Appendix B.2.7.

Theorem 4.3.12. *Let $\alpha, \alpha_f \in (0,1)$. Consider the equivalence test-then-pool procedure in Algorithm 4.1. Given probability measures Q_c, Q_h, Q_t, assume the fusion test ϕ_f satisfies*

$$\lim_{n\to\infty} \Pr\Big(\phi_f(Q_c^m, Q_h^\ell) = 1\Big) = \begin{cases} 0, & D(Q_h, Q_c) > \rho, \\ a, & D(Q_h, Q_c) = \rho, \\ 1, & D(Q_h, Q_c) < \rho, \end{cases} \tag{4.10}$$

for some $a \in (0, \alpha_f]$. Moreover, assume the causality test ϕ_* satisfies

$$\limsup_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,|\, \phi_f(Q_c^m, Q_h^\ell) = 0\Big) \le \alpha, \quad \text{if } Q_c = Q_t, \tag{4.11}$$

and

$$\limsup_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,|\, \phi_f(Q_c^m, Q_h^\ell) = 1\Big) \le \alpha, \quad \text{if } Q_c = Q_t \text{ and } D(Q_h, Q_c) \le \rho. \tag{4.12}$$

Then the overall equivalence TTP test satisfies

$$\limsup_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,|\, \mathcal{H}_0\Big) \le \alpha.$$

Remark 4.3.13. *Equation* (4.10) *is a condition on the asymptotic validity and consistency of the fusion test. It is met by, for example, the MMD equivalence test reviewed in Section 4.3.1, as shown in Theorem 4.3.1.*

Remark 4.3.14. *Theorem 4.3.12 does not require the significance level of the fusion test, α_f, to coincide with that of the causality test, α. This flexibility follows from the assumptions* (4.11) *and* (4.12)*, which require the causality test to achieve the correct asymptotic level when applied to the unmerged and merged samples, respectively.*

Remark 4.3.15. *Conditions* (4.11) *and* (4.12) cannot *be shown by directly applying the validity results for partial bootstrap (Theorem 4.3.5) and partial permutation (Theorem 4.3.8). This is because* (4.11) *and* (4.12) *are* conditional *on the outcome of the fusion test, whereas Theorem 4.3.5 and Theorem 4.3.8 only concern the* unconditional *Type-I error, which is equivalent to always merging the samples. Nonetheless, we show that these conditions still hold for partial bootstrap and partial permutation, with proofs provided in Appendix B.1.2, thus establishing the validity of our proposed TTP procedure. Similarly, as we discuss below, we extend the consistency results, Theorems 4.3.6 and 4.3.9, to the setting conditioning on the result of the fusion test.*

Next, we show that the proposed TTP procedure is consistent whenever the causality tests are. The proof is in Appendix B.2.8.

Theorem 4.3.16. *Consider the equivalence test-then-pool procedure in Algorithm 4.1. Given probability measures Q_c, Q_h, Q_t, assume that the fusion test satisfies* (4.10)*, and that*

$$\lim_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \phi_f(Q_c^m, Q_h^\ell) = 0\Big) = 1, \tag{4.13}$$

and

$$\lim_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \phi_f(Q_c^m, Q_h^\ell) = 1\Big) = 1, \quad \text{if } D(Q_h, Q_c) \le \rho. \tag{4.14}$$

Then the overall equivalence TTP test satisfies

$$\lim_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_1\Big) = 1.$$

Theorems 4.3.12 and 4.3.16 work with arbitrary fusion and causality tests, and outline sufficient conditions on these components for such consistency to hold for the corresponding TTP procedure. We argue in Appendices B.2.9 and B.2.10 that these regularity conditions hold for the proposed MMD equivalence fusion test and the causality test with either partial bootstrap or partial permutation, which gives us Corollaries 4.3.17 and 4.3.18. Notably, the significance level of the fusion test, α_f, need *not* be identical to that of the causality test, α.

Corollary 4.3.17. *Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, and Assumption 4.3.2 is satisfied. Let $\alpha, \alpha_f \in (0,1)$, and consider the equivalence TTP test described in Algorithm 4.1, where the fusion test has level α_f. Under $\mathcal{H}_0 : Q_c = Q_t$,*

1. *Using the partial bootstrap causality test with level α, the equivalence TTP test satisfies the following asymptotic level if $\sigma_c^2 > 0$ and $D(Q_c, Q_h) \neq \rho$:*

$$\limsup_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_0\Big) \leq \alpha. \tag{4.15}$$

2. *Using the partial permutation causality test with level α, the equivalence TTP test satisfies* (4.15) *for all Q_h with $D(Q_c, Q_h) \neq \rho$.*

Next, we show that the proposed TTP procedure consistently rejects alternatives whenever the causality test is capable of doing so.

Corollary 4.3.18. *Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, and Assumption 4.3.2 is satisfied. Let $\alpha, \alpha_f \in (0,1)$, and consider the equivalence TTP test described in Algorithm 4.1, where the fusion test has level α_f. Under $\mathcal{H}_1 : Q_c \neq Q_t$,*

1. *With the partial bootstrap causality test, the equivalence TTP test satisfies*

$$\lim_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_1\Big) = 1, \tag{4.16}$$

   if $Q_c = Q_h$, or if $Q_c \neq Q_h$ and $2(1-\gamma)D\big(Q_h, Q_c\big)\cos\beta < D(Q_c, Q_t)$.

2. *With the partial permutation causality test, the equivalence TTP test satisfies* (4.16) *if $Q_c = Q_h$, or if $Q_c \neq Q_h$ and $2D(Q_h, Q_c)\cos\beta < \Big(\frac{1}{1-\gamma} + \lambda\Big)D(Q_t, Q_c)$.*

Remark 4.3.19 (Lack of full consistency)**.** *In Corollary 4.3.18, the conditions on the historical control arm Q_h are inherited from Theorems 4.3.6 and 4.3.9; we refer to Remark 4.3.11 for a discussion. In particular, when $Q_h \neq Q_c$, the proposed TTP procedure is not consistent against all alternatives Q_t. Intuitively, this is because, when $Q_h \neq Q_c$, merging Q_h^ℓ introduces bias into the MMD statistic used in the causality test, and hence it can no longer consistently reject Q_t unless the difference between Q_t and Q_c is substantially larger than that between Q_h and Q_c.*

One practical mitigation is to merge with smaller probability as the sample sizes grows. This is sensible because, when the current control arm is sufficiently large, the causality test based only on the current control already has adequate power, making merging unnecessary. One way of achieving this is to let the margin ρ depend on the sample size, namely taking $\rho = \rho_n \in o(1)$. In Section 4.4.2, we assess how varying ρ affects the performance of the proposed TTP procedure. A rigorous treatment of validity and consistency for decaying sequences ρ_n remains an open problem.

Although both the fusion and the causality tests in our proposed TTP framework are built on MMD, the underlying principle extends naturally to other statistical metrics, such as the Kolmogorov-Smirnov (KS) metric and the Wasserstein metric. Computationally, MMD has a quadratic cost of $O((n+m+\ell)^2)$, whereas the KS and Wasserstein metrics require only $O((m+\ell)\log(m+\ell) + n\log n)$. However, one major advantage of MMD is the flexibility in the kernel selection. By designing suitable kernels, the fusion and the causality tests can down-weight or up-weight certain forms of distributional discrepancy, such as mean shifts or tail behaviour; see the next section for a discussion. In contrast, the standard KS statistic does not offer this type of tuning and, because it is based on the supremum norm between empirical distribution functions, it is less sensitive to discrepancies in the tails or to isolated outliers (Bryson, 1974; Wang et al., 2014), making it more restrictive in practice. We leave the discussion of using Wasserstein metric to Section 4.6.

4.3.4 Choosing the kernel and bandwidth

Our framework allows flexible choices of kernels and bandwidth, which can be selected according to specific objectives. In general, we recommend employing the RBF (radial basis function) kernel $k(x,y) = \exp(-\frac{1}{2\zeta^2}\|x-y\|^2)$ with the bandwidth ζ[10] set according to the median heuristic (Gretton et al., 2012), defined as

$$\zeta^2_{\text{med}} \coloneqq \text{Median}\{\|v-v'\|_2^2 : v, v' \in \mathcal{D}, v \neq v'\}, \tag{4.17}$$

where $\mathcal{D} = \mathcal{D}_{n,m,\ell}$ if the historical sample is merged, and $\mathcal{D} = Q_c^m \cup Q_t^n$ otherwise.

It is also possible to tailor the kernel to target particular aspects of distributions. For example, if one is only interested in testing equality of the first moment, namely the ATE, then a linear kernel $k(x,y) = x^\top y$ can be used. Moreover, heavy-tailed or even unbounded kernels, such as the inverse multi-quadric kernel (Sriperumbudur et al., 2011, Example 3) or polynomial kernels, are more sensitive to differences in the tails (Sadhanala et al., 2019, S.13.2). However, linear and polynomial kernels are not characteristic: for example, a linear kernel cannot distinguish differences in variances (Sriperumbudur et al., 2010, Example 2). One possible solution is to combine such a kernel with a characteristic kernel, such as the RBF, and use $k(x,y) = x^\top y + \epsilon \exp(-\frac{1}{2\zeta^2}\|x-y\|^2)$ for some $\epsilon > 0$. This kernel remains characteristic and thus ensures that the MMD can distinguish all distributions, while still up-weighting first-moment

[10] Within this chapter ζ denotes the kernel bandwidth, distinct from the engression noise of Chapter 5.

differences.

In our experiments, we find that bounded kernels, such as the RBF kernel, typically yield better performance, whereas unbounded ones such as the linear kernel often require large sample sizes for the asymptotic properties to hold accurately, thus leading to poor Type-I error rate control with small samples.

4.3.5 Choosing the equivalence radius

In our setup, ρ controls the tolerance for differences between the historical and current data. A larger ρ allows greater heterogeneity, leading to a higher merging rate. Although, under the conditions of Corollary 4.3.17, the choice of ρ does not affect the validity of the proposed testing procedure, it influences consistency. As discussed in Section 4.4.2, a large ρ may reduce power when Q_c, Q_h, and Q_t satisfy certain geometric conditions. Therefore, we recommend choosing a small ρ to ensure a power gain. Nevertheless, the key advantage of our method lies in its robustness: it guarantees Type-I error control regardless of how different the merged and original datasets are.

Note that ρ is a hyperparameter that must be specified a priori. In Remark 4.3.19, we discuss a potential extension of choosing decaying ρ_n as a function of sample size n. Another alternative is to adjust ρ after observing the test results, but doing so may violate Type-I error control for $\mathcal{H}_0^f$ and, consequently, inflate the Type-I error for $\mathcal{H}_0$. A promising line of future work is to explore adaptive or sequential updating of ρ as the function of observed data, for example, via e-values (Ramdas and Wang, 2025).

4.4 Synthetic experiment

We conduct synthetic experiments to compare the proposed equivalence-based TTP with the classic TTP for detecting DTE, as defined in Section 4.2.3. The experiments are designed mainly to demonstrate the validity of our method and the power improvement achieved through data fusion.

We set both α_f and α to be 0.05. All MMD metrics are computed using the RBF kernel, where the bandwidth is selected by the median heuristic as in (4.17). We present results from 1000 simulations with $B = 1000$ bootstrap or permutation replicates per simulation.

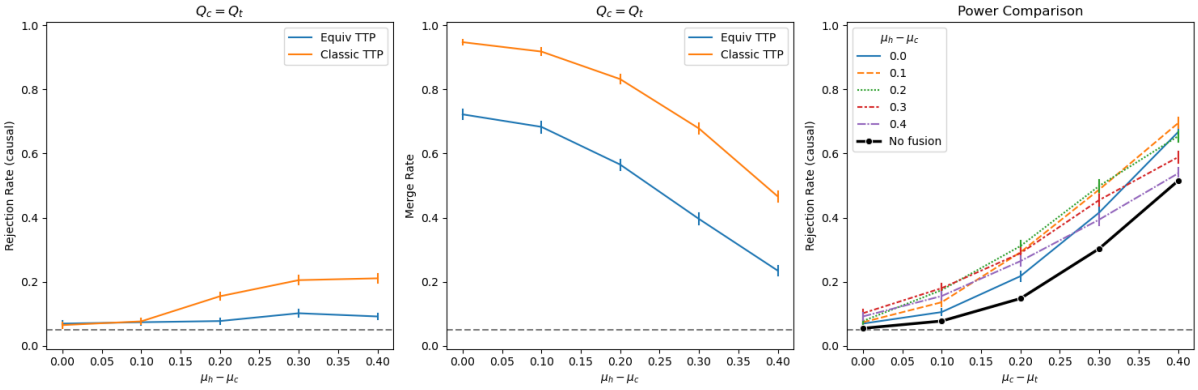


Figure 4.3: Type-I error (leftmost panel) and power (rightmost panel) under distributional discrepancy induced by mean shifts. The middle panel shows the proportion of the 1000 simulations in which historical controls are merged. The parameter ρ is fixed at 0.4. The RBF kernel and partial bootstrap procedure are used.

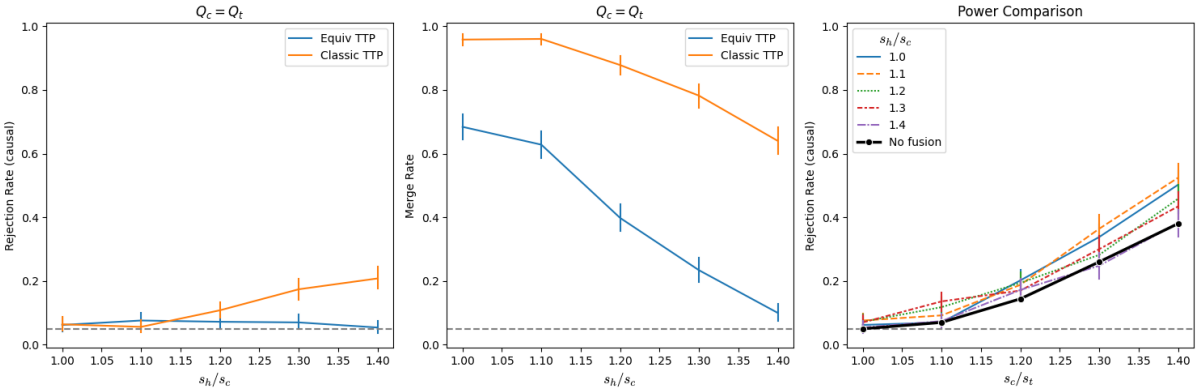


Figure 4.4: Type-I error (leftmost panel) and power (rightmost panel) under distributional discrepancy induced by variance shifts. The middle panel shows the proportion of the 1000 simulations in which historical controls are merged. Type-I error and power results when distribution discrepancy is produced by variance shifts. ρ is fixed as 0.4. The RBF kernel and partial bootstrap procedure are used.

4.4.1 Equivalence TTP controls Type-I error and improves power

Figure 4.3 and Figure 4.4 compare the performance of the proposed equivalence TTP, the classic TTP, and a test without fusing historical controls under mean-shift and variance-shift settings, respectively. In these experiments, we set $n = 100$, $m = 50$ and $\ell = 100$.

In the mean-shift scenario (Figure 4.3), $Q_t = \mathcal{N}(0, 1)$, $Q_c = \mathcal{N}(\mu_c, 1)$, and $Q_h = \mathcal{N}(\mu_h, 1)$, with the axes varying $\mu_h - \mu_c$ and $\mu_c - \mu_t$. The variance-shift scenario (Figure 4.4) is the analogue with common mean 0 and shifted scales: $Q_t = \mathcal{N}(0, 1)$, $Q_c = \mathcal{N}(0, s_c^2)$, and $Q_h = \mathcal{N}(0, s_h^2)$, with the axes varying the scale ratios s_h/s_c and s_c/s_t ($s_t = 1$).

The results demonstrate the effectiveness of the proposed equivalence TTP method. Using

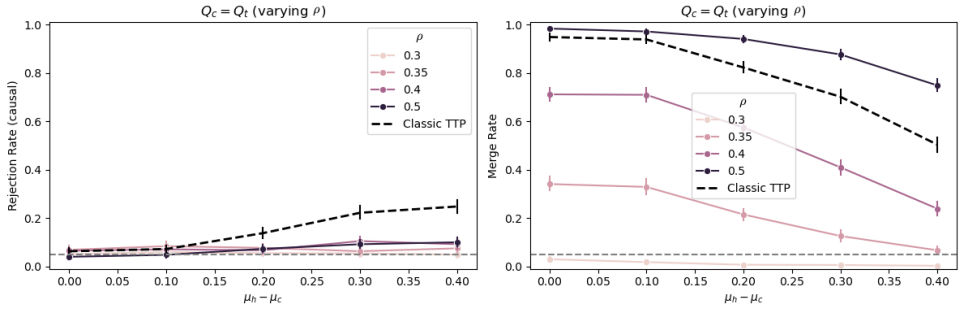


Figure 4.5: Type-I error (left panel) is controlled across varying ρ, despite differing merge rates (right panel). In these experiments, $Q_t = Q_c = \mathcal{N}(0, 1)$, $Q_h = \mathcal{N}(\mu_h, 1)$. The RBF kernel is used with the partial bootstrap procedure.

$\rho = 0.4$, the procedure successfully fuses historical controls selectively, as shown in the middle plots, while maintaining control of the Type-I error rate. This highlights one of the main advantages of our proposed method, as it differs from the classic TTP which exhibits clear Type-I error inflation (left panels). At the same time, our method achieves substantial power gains relative to not fusing controls, confirming the benefits of incorporating historical data when it is sufficiently similar to the current control group.

4.4.2 Effects of varying ρ

We also investigate how varying the choice of ρ affects both the merge rate and the power of the causality test under different "directions" of distribution shifts. We follow the same experiment set up used in Figure 4.3. As shown in Figure 4.5, larger values of ρ increase the probability of merging even when Q_h and Q_c differ more substantially. Nevertheless, the Type-I error of the causality test remains close to the nominal significance level $\alpha = 0.05$, indicating that validity is preserved across a range of margins.

Figure 4.7 examines the effect of ρ on statistical power. Here we allow $\mu_c - \mu_t$ to take negative values. Since $\mu_h - \mu_c > 0$, this setting corresponds to Q_t and Q_h moving in the same direction away from Q_c, i.e., $\cos\beta > 0$. In this case, merging Q_h reduces the apparent difference between Q_t and Q_c which can lead to reduced power. Indeed, for sufficiently large ρ, we observe power dropping below the no-fusion baseline due to this cancellation effect. We highlight this result as a cautionary note: overly permissive choices of ρ can harm power, reinforcing the importance of choosing ρ conservatively in practice.

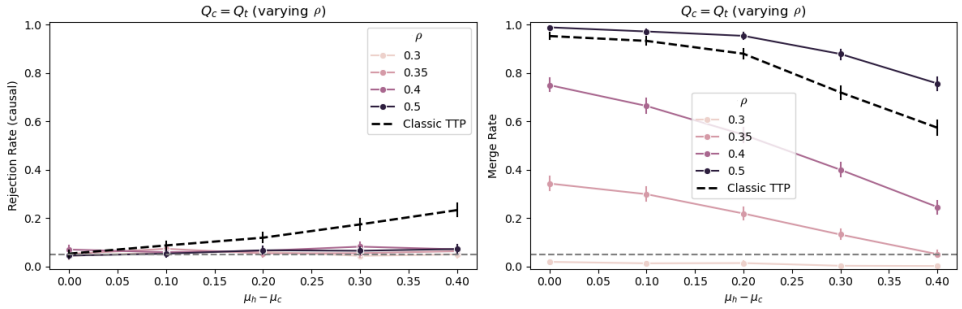


Figure 4.6: Type-I error (left panel) is controlled across varying ρ, despite differing merge rates (right panel). In these experiments, $Q_t = Q_c = \mathcal{N}(0, 1)$, $Q_h = \mathcal{N}(\mu_h, 1)$. The RBF kernel is used with the partial permutation procedure.

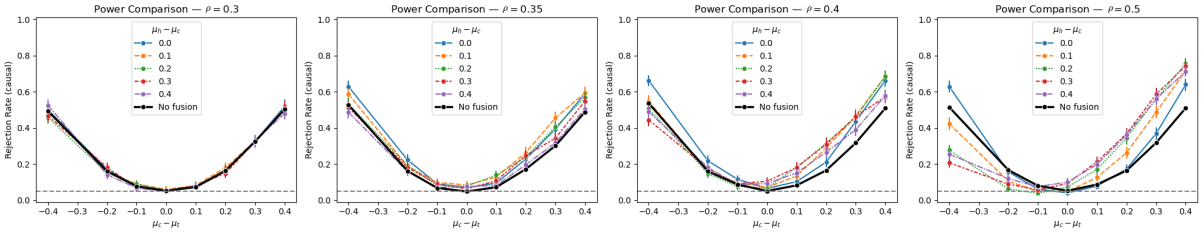


Figure 4.7: Power comparison with ρ varying. In these experiments, $Q_t = \mathcal{N}(0, 1)$ with $\mu_t = 0$, $Q_c = \mathcal{N}(\mu_c, 1)$, and $Q_h = \mathcal{N}(\mu_h, 1)$. The RBF kernel and partial bootstrap are used.

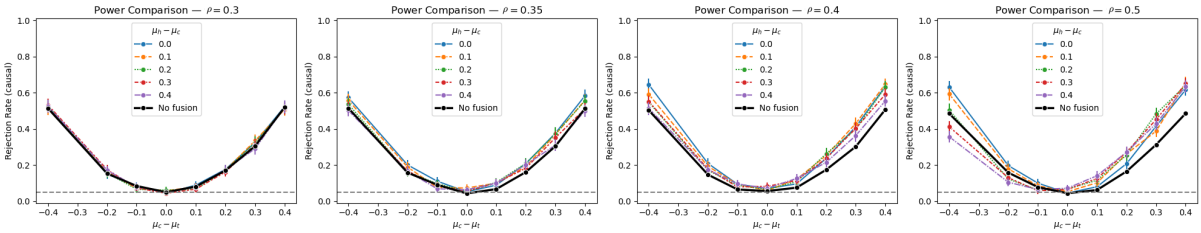


Figure 4.8: Power comparison with ρ varying. In these experiments, $Q_t = \mathcal{N}(0, 1)$ with $\mu_t = 0$, $Q_c = \mathcal{N}(\mu_c, 1)$, and $Q_h = \mathcal{N}(\mu_h, 1)$. The RBF kernel and partial permutation are used.

4.4.3 Comparison between partial bootstrap and partial permutation

In Figures 4.6 and 4.8, we present results from partial permutation under the same data generation mechanism with mean shift, as in Figure 4.5. As expected, the Type-I error and the power follow similar patterns to those in the partial bootstrap setting. Meanwhile, as we suggest above, since partial permutation fails to approximate the null distribution when $Q_c \neq Q_t$, the power from partial permutation is indeed smaller than what we have in partial bootstrap; we have the same observation in results of Table 4.2.

Bandwidth in RBF Kernel	$\mu_c - \mu_t = -0.4$		$\mu_c = \mu_t$		$\mu_c - \mu_t = 0.4$	
	Merge	**Causal**	**Merge**	**Causal**	**Merge**	**Causal**
Median heuristic	0.285	0.577 (0.513)	0.307	0.064 (0.051)	0.272	0.572 (0.535)
0.5	0.319	0.528 (0.511)	0.342	0.070 (0.063)	0.354	0.536 (0.566)
1	0.673	0.572 (0.559)	0.638	0.074 (0.053)	0.694	0.711 (0.629)
1.5	0.824	0.581 (0.566)	0.807	0.074 (0.065)	0.832	0.739 (0.677)

Table 4.2: Merge rates and causality test rejection rates by bandwidth and $\mu_c - \mu_t$ blocks by partial bootstrap, averaged over 1000 simulations, with 1000 bootstraps / permutations in each simulation. $\alpha_f = \alpha = 0.05$, ρ is fixed at 0.4, $\mu_h - \mu_c = 0.2$. For the causality test, rejection rates from the partial bootstrap are reported directly, with the corresponding partial permutation results given in parentheses. All experiments are done with RBF kernels.

4.4.4 Comparison of varying kernel bandwidth

We evaluate the performance of the proposed equivalence TTP using partial bootstrap and partial permutation under varying kernel bandwidths of the selected RBF kernel, with $n = 100$, $m = 50$, and $\ell = 100$ (Table 4.2). For fixed ρ, the kernel bandwidth influences the merge rate. Nevertheless, equivalence TTP consistently controls the Type-I error. Moreover, the partial bootstrap method generally achieves higher power than partial permutation, in agreement with our earlier discussion.

4.4.5 Comparison with linear kernel

As discussed earlier, our method accommodates various kernel choices, which should be selected according to the task. The linear kernel effectively captures differences in first moments but may fail to detect higher-order discrepancies such as variance shifts. In Figure 4.9, we apply the linear kernel to the variance shift setting from Figure 4.4. Although the underlying distributions differ in variance, the MMD two-sample tests fail to detect this difference, regardless of whether data fusion is applied.

4.5 Prospera program

In this section, we apply the equivalence TTP procedure on the *Prospera* program. Originally launched in 1997 as Progresa, Mexico's conditional cash transfer (CCT) program provides cash to poor households conditional on school attendance, health visits, and participation in nutrition

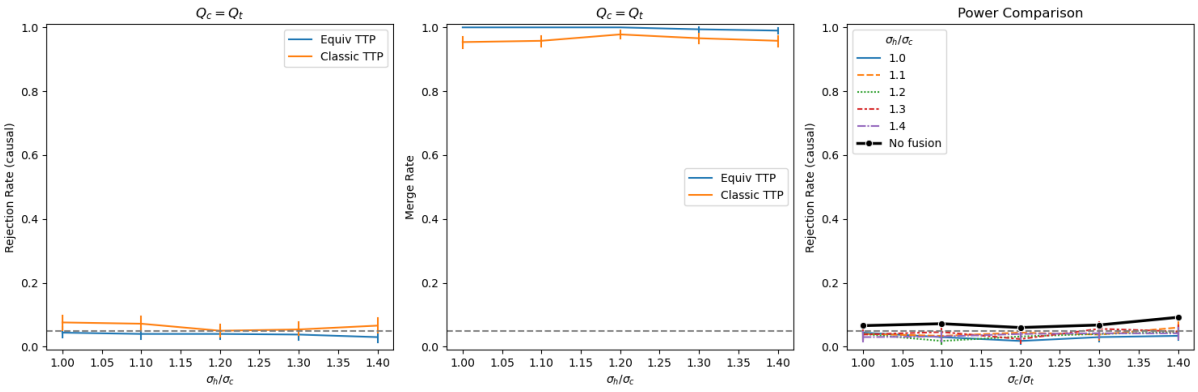


Figure 4.9: Experiment results when distributional discrepancy produced by shifting variance. ρ is fixed at 1. The linear kernel and partial bootstrap are used.

programs (Skoufias, 2005). The initial evaluation was conducted as a randomised controlled trial across 505 villages in seven poor states: 320 villages were assigned to treatment (eligible poor households received stipends conditional on compliance) and 185 to control (no program). We use village-level school enrollment rates from 1998 as post-treatment outcomes, with the 1997 baseline serving as pre-treatment information.

We first examine whether the program improved school enrollment. A standard two-sample t-test applied to the complete 1998 data rejects equality of mean enrollment rates between treatment and control villages at the $\alpha = 0.05$ level ($p = 0.0118$). This confirms that Prospera significantly increased school enrollment, consistent with prior findings on the program's effectiveness.

To evaluate performance in settings with limited contemporary controls, we randomly subsample 50 control villages from the 1998 control group and use their corresponding 1997 baseline data as historical controls. We also randomly sample 200 villages from the treatment group for testing. Figure 4.10 displays the enrollment rate distribution of treatment, control, and historical groups in one such simulation. We repeat this procedure for 1000 simulations in total.

We set the fusion test level as $\alpha_f = 0.05$ and the significance level of the causality test to $\alpha = 0.05$. As before, the kernel bandwidth ζ is chosen by the median heuristic (4.17) for the RBF kernel. Under these settings, the proposed equivalence TTP procedure accepts fusing the current control group with historical control group (1997) at a margin of $\rho = 0.5$ in 99.4% of the 1000 replications.

Table 4.3 reports the rejection rates of the causality test against $\mathcal{H}_0$ across different methods.

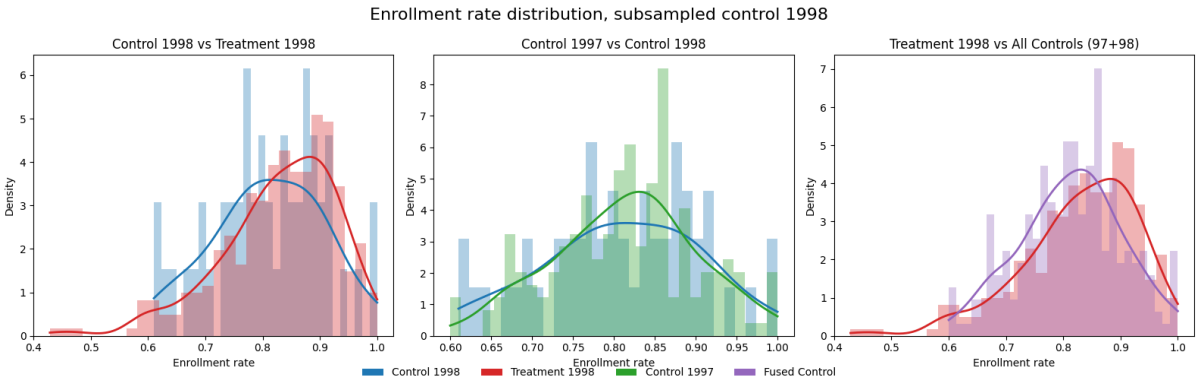


Figure 4.10: Distribution of enrollment rates across groups.

Method	Rejection Rates
Equivalence TTP	0.61
Distribution test (without fusion)	0.40
Mean test (with fusion)	0.36
Mean test (without fusion)	0.23

Table 4.3: Rates of rejecting $\mathcal{H}_0$ of 1000 simulations.

The proposed TTP achieves the highest power, outperforming a classical distribution test that does not fuse control groups. Furthermore, compared with their mean-based counterparts, distributional tests demonstrate superior power, as the differences in distributions additional to the mean are captured. Together, these findings validate the effectiveness of the proposed equivalence TTP framework.

4.6 Discussion

In this chapter, we focus on the fusion of control arms (and, symmetrically, treatment arms) within randomised clinical trials, and therefore make no assumptions such as unconfoundedness. In practice, however, it is increasingly common to consider fusing real-world observational data with experimental data. Extending our framework to such settings requires additional assumptions, most notably ignorability and the comparability of populations. Lack of ignorability in the observational data may induce confounding bias; even under ignorability, substantial differences in covariate distributions between datasets may alter the target estimand.

Our proposed method can be adapted to address these concerns. For instance, MMD-based tests can be applied to assess distributional equivalence of covariates between observational and experimental samples, providing a principled diagnostic for population alignment in high-

dimensional settings. Similarly, when imbalance is detected, our framework could be combined with covariate adjustment or weighting strategies to mitigate bias before fusion. In this way, the same equivalence-testing tools that validate control-arm pooling in trials can also serve as safeguards when extending pooling to broader causal inference contexts.

The choice of radius ρ presents an interesting direction for future work. As discussed in Section 4.3.5, intuitively, ρ should not be too large, so that the merged historical control distribution Q_h^ℓ remains similar to the current control distribution Q_c^m; however, setting ρ too small may make the fusion test overly conservative. We provide ideas on potential extensions in Section 4.3.5, Remark 4.3.19 and Section 4.4.2.

We discuss at the end of Section 4.3.3 on using other distributional distance metrics, and one of the potential directions is to use the Wasserstein metric. Different from the Kolmogorov-Smirnov metric, the Wasserstein distance can in principle be made more flexible through an appropriate choice of cost function, but we do not pursue this direction here, as both the equivalence testing employed in this chapter and most existing methods for distributional treatment effects are formulated in terms of MMD. Extending the proposed TTP framework to Wasserstein-based test statistics is therefore a natural direction for future work.

Another potential extension involves merging multiple data sources. Instead of relying on binary "accept or reject" decisions, one could assign weights based on their similarity to the current control group. Specifically, the kernel-based distance $D(Q_c, Q_h)$ from the kernel equivalence test could be used to construct a smooth weighting function, allowing closer sources to contribute more heavily to the fused control distribution while down-weighting or excluding more distant sources. Combining information from several candidate sources in this manner would require developing procedures for joint calibration.

Finally, while this work is motivated by causal inference, the proposed framework has potential applications beyond causal settings, such as domain adaptation, transfer learning, and robust statistical testing in machine learning contexts.

Chapter 5

Stability evaluation via faithful causal simulation

Chapter 3 and Chapter 4 establish their stability guarantees analytically. Assessing and comparing such methods empirically requires a benchmark with known causal structure, which real data rarely provide. This chapter constructs one: a generative model that reproduces the observed and interventional distributions, centring the parameterisation on the target causal estimand. The resulting model can be used both for causal estimation and for controlled simulation. This chapter is a revised and extended version of joint work with Professor Xinwei Shen and Professor Robin J. Evans (Yang et al., 2025).

5.1 Introduction

The use of machine learning tools has given causal inference a new lease of life, enabling complex models to be used with principled causal estimators and guarantees about statistically important quantities (Wager and Athey, 2018; Chernozhukov et al., 2018; Hahn et al., 2020). To build trustworthy causal models, however, we also need to understand when these methods may be more or less reliable, or perhaps fail completely. This implies that causal inference needs a set of good benchmarking tools. Unfortunately, real-world datasets are not ideal for this task, because they cannot give us access to the ground truth other than in a few very special circumstances. In particular, they rarely provide the counterfactual outcomes we care about, and the distribution we want to evaluate often differs from the one that produced the observations. Well-designed simulations can address this discrepancy (Neal et al., 2020; Parikh et al., 2022); they allow us to choose a ground truth, stress-test new methods, compare their generalisability and stability, and expose failure modes before deployment. In trial design, simulations also guide sample size calculations, making patient recruitment more efficient. Valid benchmarks have therefore

become central to method development in the AI era. Faithful simulation is no longer optional; it is a core step in building trustworthy causal models (Friedrich and Friede, 2024; Pezoulas et al., 2024; Gamella et al., 2025).

Generating data from a marginal interventional distribution, rather than merely reproducing observational conditional relationships, poses a distinct modelling problem (Havercroft and Didelez, 2012). We allow (Z, X, Y) to be multivariate, with $Z \in \mathbb{R}^{d_z}$, $X \in \mathbb{R}^{d_x}$, $Y \in \mathbb{R}^{d_y}$. Inspired by Evans and Didelez (2024), the joint distribution P of (Z, X, Y) can be decomposed into the distribution of the past P_{ZX}, the causal margin $P_{Y(x)}$, and a remaining component that completes the joint distribution. This decomposition isolates $P_{Y(x)}$ as a primary object.

The difficulty is that $P_{Y(x)}$ cannot, in general, be obtained from the structural equation of the outcome alone: while $Y \mid X, Z$ is specified by a structural model, the interventional margin additionally depends on the distribution of Z under intervention. Thus, modelling $P_{Y(x)}$ requires capturing observational structure while enabling controlled generation under interventions.

We introduce *frengression*, a generative approach that models the marginal interventional distribution directly. The central idea is to represent $P_{Y(x)}$ via a reparameterised independent noise variable, yielding a generative mechanism that supports sampling under user-specified interventions.

5.1.1 Contribution

Frengression directly targets marginal causal quantities while also generating complex benchmark datasets with user-specified marginal causal distributions. This unified framework guarantees a precisely defined marginal effect while closely reflecting real-world conditions. Consistent with the frugal parameterisation, frengression consists of variation independent components to isolate and control specific causal properties, such as the strength of confounding.

We highlight some key properties of frengression:

- **Generative model for causal margin.** Frengression is a flexible deep generative model for distributional regression that represents the entire outcome distribution while explicitly specifying the marginal interventional distribution of interest. Isolating this margin as a variation independent component lets it be controlled directly.
- **Broad applicability.** Our framework supports both estimation and simulation for a

Method	Binary Treatment	Continuous Treatment	Longitudinal Data	Distributional Treatment Effects	Extrapolation	Modelling Causal Margin
TARNet CFRNet Dragonnet CEVAE	✓	×	×	×	×	×
CausalEGM CausalBGM	✓	✓	×	×	×	×
RealCause Credence	✓	×	×	×	×	×
DeepACE	✓	×	only static treatment regime	×	×	×
Deep LTMLE	✓	×	✓	×	×	×
Frugal Flow	✓	✓	×	×	×	✓
Frengression	✓	✓	✓	✓	✓	✓

Table 5.1: Comparison of methods.

variety of (complex) settings, including distributional treatment effects, continuous treatments, longitudinal settings, and survival data.

- **Flexibility and robustness.** Frengression factorises the joint distribution according to a frugal, variation independent parameterisation, enabling each component to be modelled with any suitable method (not limited to neural networks). The model fitting does not require extensive hyperparameter tuning, enhancing overall stability.
- **Theoretical guarantees.** We provide a consistency guarantee for models trained with the empirical energy score, along with extrapolation guarantees under continuous treatments.

5.1.2 Related work

We review existing approaches to generative modelling for causal estimation and simulation and how they motivate our frengression framework. We summarise this comparison between frengression and other methods in Table 5.1.

5.1.2.1 Deep generative models for the causal margin

Deep learning models have shown remarkable flexibility for causal estimation in high-dimensional, complex settings. TARNet and CFRNet learn shared representations balancing treated and control groups (Shalit et al., 2017); Dragonnet incorporates propensity score estimation into the outcome network (Shi et al., 2019). However, since they enforce covariate balance across treatment arms, these architectures are restricted to binary treatments and static covariates, and do not extend to longitudinal data.

Structural causal models (Section 2.1.2) provide a natural generative framework: one builds

a generative model directly for the structural assignments of each variable. Holovchak et al. (2025) applied this idea in the nonparametric instrumental variable setting, estimating the full (conditional) interventional distribution under latent confounding. Other generative approaches to causal effect estimation include CEVAE (Louizos et al., 2017), which uses variational autoencoders to recover the joint distribution of observed and latent confounders from purely observational data; CausalEGM (Liu et al., 2024), which encodes confounders into a low-dimensional latent space via an encoder–generator architecture, accommodating continuous treatments (e.g. dosage levels); and its extension CausalBGM (Liu and Wong, 2025), which adds a Bayesian neural network returning both the mean and variance of individual treatment-effect posteriors under a Gaussian assumption.

All these methods model the *conditional* distribution. Modelling the *marginal interventional* distribution generatively is less direct, since the causal margin is not in general induced by the structural equation of the outcome alone. Concretely, if $Y = f(X, Z, \varepsilon_Y)$ with ε_Y exogenous, sampling from $P_{Y(x)|Z=z}$ requires only drawing ε_Y and computing $f(x, z, \varepsilon_Y)$; sampling from the marginal $P_{Y(x)}$ additionally requires simulation from the interventional distribution of $Z(x)$. The generative mechanism inducing the causal margin therefore does not coincide with the original structural form. Frengression instead constructs a generative model for the causal margin directly, via a reparameterised independent noise variable. The variation independence of its components permits flexible specification of the margin while preserving causal validity.

5.1.2.2 Modelling longitudinal data

Causal estimation in longitudinal settings remains comparatively underdeveloped. Building on LTMLE (longitudinal targeted maximum likelihood estimation, van der Laan and Gruber, 2012; Lendle et al., 2017) and iterative g-computation, Frauen et al. (2023) proposed DeepACE, which uses recurrent neural networks to model patient trajectories and estimate average risk ratios; it handles only static treatment regimes. Shirakawa et al. (2024) introduced Deep LTMLE, extending LTMLE and DeepACE through transformer architectures to estimate counterfactual mean outcomes under dynamic treatment policies. Both methods estimate marginal causal quantities by empirical averaging of conditional potential outcomes, require careful hyperparameter tuning, and do not support data simulation.

By contrast, frengression extends directly to longitudinal data, accommodating marginal and conditional estimands under both static and dynamic regimes, and supporting data simulation

with little tuning.

5.1.2.3 Modelling distributional treatment effects

As mentioned in Chapter 4, most causal estimation methods target the average treatment effect, but DTEs have become increasingly important for richer insights beyond the mean (Abadie, 2002; Bitler et al., 2017). Existing approaches include: semiparametric models (Díaz, 2017; Kennedy et al., 2023), which require correct model specification; kernel-based methods, which avoid parametric assumptions but are difficult to interpret and scale poorly in high dimensions (Muandet et al., 2017; Park et al., 2021); neural network methods (Holovchak et al., 2025), which are flexible but do not admit user-specified causal quantities; and Bayesian methods (Venturini et al., 2015; Xu et al., 2018), which are computationally expensive. Each trades off some combination of flexibility, interpretability, and efficiency.

Frengression bridges this gap through its modular, variation independent parameterisation and distributional regression models, which directly target specific distributional quantities while preserving both transparency and efficiency in estimation and data simulation.

5.1.2.4 Generative models for causal data simulation

Several generative models simulate data for benchmarking causal inference methods. Athey et al. (2024) use Wasserstein generative adversarial networks (GANs) in a Monte Carlo simulation framework. Neal et al. (2020) developed RealCause, based on neural autoregressive flows (Huang et al., 2018). Parikh et al. (2022) proposed Credence, a variational autoencoder factorising the joint distribution into the covariate marginal, the treatment given covariates, and the outcome given covariates and treatment. The conditional and marginal outcome distributions are specified only indirectly through a post-hoc penalty in the training loss, and the model is restricted to binary treatments.

None of these methods allows direct control of the marginal causal effect. Frugal flows (de Vassimon Manela et al., 2024) address this by combining the frugal parameterisation with normalising flows to model the marginal interventional distribution directly. Their reliance on copula flows (Kamthe et al., 2021) restricts them to univariate outcomes or static regimes, requires hierarchical copulas for multivariate outputs, and is sensitive to hyperparameters.

Compared to density-estimation generative models, which are restricted either to invertible

transformations (frugal flows and RealCause, via normalising flows) or to tractable decoder distributions (CausalEGM, CEVAE, and Credence, via variational autoencoders with Gaussian or Bernoulli decoders), frengression is easier to fit and allows direct control of the causal margin while keeping data simulation tractable.

5.1.3 Structure

We introduce the frengression method in Section 5.2, the architecture in Section 5.2.3, and the estimation procedure and identifiability results in Section 5.2.4. We provide consistency results for fitting the model in Section 5.3. Frengression exhibits strength with respect to treatment extrapolation, which is explored in Section 5.4. Its flexibility for longitudinal data and survival analysis is demonstrated in Section 5.5. We then provide empirical results to show frengression's comprehensive capability in synthetic experiments (Section 5.6) and real trial data LEADER (Section 5.7). We conclude with potential directions in Section 5.8.

5.2 Method

In this section we describe the necessary notation and basic method for frengression.

5.2.1 Notation

We use a hat to mark samples generated from a (fitted) frengression model, e.g. $\widehat{Z}$, $\widehat{X}$, $\widehat{Y}$. The Euclidean norm of a vector $x \in \mathbb{R}^d$ is denoted $\|x\|$. We write P_V for the distribution of a random variable V, $P_{V \mid W}$ for the conditional distribution of V given W, and p_V for the corresponding density or mass function. We use $\overline{W}_k$ to denote $(W_0, \ldots, W_k)$, the values of a time-varying variable W up to time point k. As with the frugal parameterisation (Evans and Didelez 2024, Remark 1.4), our method fits within any causal inference framework, including Pearl's "$do(\cdot)$" operator $\Pr(Y = y \,|\, do(X = x))$. We follow the potential outcomes setup of Definition 2.1.1, writing $P_{Y(x)}$ for the distribution of $Y(x)$.

In this chapter, we consider a general setting depicted in Figure 5.1(a), which describes a simple time-varying confounding scenario. We partition the treatment vector $X = (X_0, X_1)$, where X_0 denotes the intervention on the treatments causally prior to Z, and X_1 denotes interventions on the treatments causally subsequent to Z; the respective supports are $\mathcal{X}_0$ and $\mathcal{X}_1$.

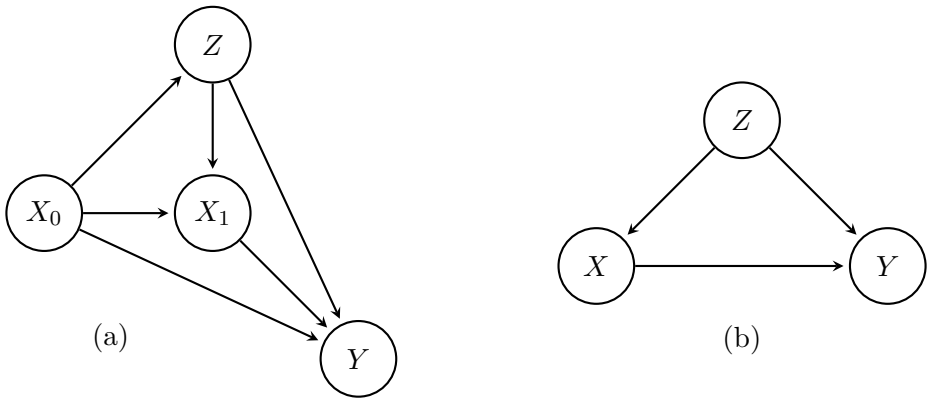


Figure 5.1: (a) General causal DAG considered in this chapter. This time-dependent confounding structure is commonly seen in longitudinal data. (b) Causal DAG without X_0. This structure is more common in time-fixed settings.

In such a setting, when we write $Y(x)$, it is equivalent to $Y(x_0, x_1)$. This structure is commonly seen in longitudinal analysis (Havercroft and Didelez, 2012) and serves as the motivating challenge for simulating data that satisfies marginal requirements in Evans and Didelez (2024). In the absence of any interventions on the covariates, the causal DAG reduces to the simpler form shown in Figure 5.1(b), which is more typical in static settings.

5.2.2 Review of key concepts

We have provided a brief review of the frugal parameterisation in the introduction. As hinted by its name, frengression employs engression as the generative model. Before we proceed further to more details, we provide a review of concepts related to engression here.

Parameterised by a neural network, engression aims to estimate the full conditional distribution of a response Y given predictors W. This predictive distribution $Y \,|\, W = w$ can be taken from a general model class produced by engression and be written as $\mathcal{M} = \{f(w, \zeta)\}$, where $f : (w, \zeta) \mapsto y$ belongs to a function class and ζ[11] is a random vector with a pre-specified distribution, independent of the covariates. A typical choice is that ζ follows the multivariate standard Gaussian distribution. We write $P_f(y \,|\, w)$ for the distribution of $f(w, \zeta)$, with the randomness coming from ζ, and $\mathcal{L}(P, P_0)$ for the loss function of a distribution P given a reference P_0. Engression specifically concerns a solution to:

[11] Within this chapter ξ, η, and ζ denote noise variables. Elsewhere ξ is the nuisance parameter of Chapter 3 and the kernel mean embedding of Chapter 4, η a learned representation in Chapter 3, and ζ a kernel bandwidth in Chapter 4.

$$\widetilde{f} = \operatorname*{argmin}_{f \in \mathcal{M}} \mathbb{E}_{(Y,W)\sim P}\, \mathcal{L}(P_f(y \mid W), Y).$$

Although different loss functions can be used in engression (Shen and Meinshausen, 2024, Appendix D), it shows strong, reliable performance using the energy loss $\mathcal{L}(P_f(y \mid W), Y) = -\operatorname{ES}(P_f(y \mid W), Y)$, where the energy score (ES) is defined as

$$\operatorname{ES}(P, u) = \frac{1}{2}\mathbb{E}_{U,U'\sim P}\|U - U'\| - \mathbb{E}_{U\sim P}\|U - u\| \tag{5.1}$$

given a distribution P and an observation u (Gneiting and Raftery, 2007); the U, U' are taken to be independent. The energy score is associated with the energy distance (Székely and Rizzo, 2013), a distributional divergence that is defined as

$$\operatorname{ED}(P, P') = 2\mathbb{E}_{U\sim P, V\sim P'}\|U - V\| - \mathbb{E}_{U,U'\sim P}\|U - U'\| - \mathbb{E}_{V,V'\sim P'}\|V - V'\|. \tag{5.2}$$

This is a special case of the squared maximum mean discrepancy (MMD; see Section 4.2.2) distance, denoted by $\operatorname{MMD}^2(P, P')$, with kernel $k(u, v) = \frac{1}{2}(\|u\| + \|v\| - \|u - v\|)$ (Sejdinovic et al., 2013). Energy loss can be estimated directly from samples, providing computational simplicity. It differs from the energy distance only by a positive scaling and an additive constant independent of the generator parameters, so the two share the same minimiser. For compatibility with the relevant literature regarding distributional distance (Briol et al., 2019), we prove the consistency of models trained aiming at minimising energy distance $\operatorname{ED}(P_f, P)$ in Section 5.3.

Another speciality of engression is that its model class $\mathcal{M}$ contains the pre-additive noise models (pre-ANMs). Define function classes $\mathcal{F}$ and $\mathcal{G}$. The pre-ANMs can be written as

$$f(w + g(\zeta)) : f \in \mathcal{F}, g \in \mathcal{G}, \tag{5.3}$$

where $g(\zeta)$ represents the pre-additive noise; f and g are to be learned. Shen and Meinshausen (2024) provided detailed discussions on the extrapolation properties provided by pre-ANMs, allowing one to relax the causal effect identification assumptions; details are given in Section 5.4.

5.2.3 Model and parameterisation

The architecture of frengression is shown in Figure 5.2. A frengression model contains three components, each corresponding to one component in the frugal parameterisation:

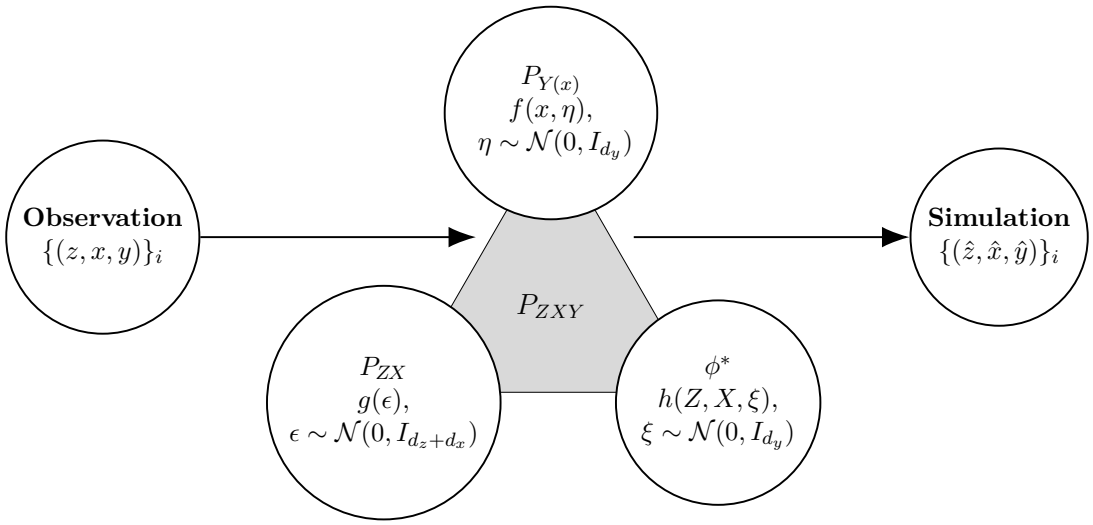


Figure 5.2: Estimation and simulation architecture of frengression.

- For the joint distribution P_{ZX}, we build a generative model:

$$g(\epsilon), \tag{5.4}$$

  where $\epsilon \sim \mathcal{N}(0, I_{d_z+d_x})$.

- Given a specified form of the (marginal) causal effect $P_{Y(x)}$, we use

$$Y = f(x, \eta) \tag{5.5}$$

  as the generative form, where $\eta \sim \mathcal{N}(0, I_{d_y})$. For example, if $P_{Y(x)}$ is $\mathcal{N}(x, 1)$, it can be specified in a generative form as $f(x, \eta) = x + \eta$.

- Association remainder $\phi^*_{Y|ZX}$. Given the causal margin in (5.5), the remaining part to fully specify the distribution $P_{Y|ZX}$ is a conditional distribution of $\eta \,|\, Z = z, X = x$. For this, we build a conditional generative model

$$\widetilde{\eta} = h(z, x, \xi), \tag{5.6}$$

  where $\xi \sim \mathcal{N}(0, I_{d_y})$. Similar to the copula measure chosen by the original frugal parameterisation, this term captures the association between Y and (Z, X) in a way that is compatible with the specified causal margin.

Frengression parameterises the joint (observed) distribution via generative models, preserving the same properties as the original frugal parameterisation. First, the three components

jointly specify the whole distribution P_{ZXY}. To see this, note that model (5.4) specifies the distribution for (Z, X), while models (5.5) and (5.6) together specify a distribution of $Y \mid Z, X$. Furthermore, the three components can be chosen to be variation independent. The first component is about the distribution of (Z, X) whereas in the other two components, X and Z are either given or conditioned on; thus, g is variation independent of (f, h). Note that the marginal distribution of η is always fixed as the standard Gaussian and is thus not a model component. For the variation independence between f and h, note that model (5.5) specifies the causal margin $P_{Y(x)}$ only through the function f with a fixed value of x and η drawn from its (fixed) marginal distribution, whereas model (5.6) is about the conditional distribution of $\eta \mid Z, X$ with a constraint of a normal margin. Thus, these two components are variation independent. More details are given in Remark 5.2.1.

We write a frengression model as $\varphi = (g, f, h)$. While we specify the distributions of ϵ, η and ξ in this chapter, their distributions can be non-Gaussian, provided that they are sampled independently of the observed data. Note that even though we draw η from a standard normal distribution, because it is subsequently passed through the nonlinear function $f(\cdot)$, the interventional distribution induced by frengression need not be Gaussian.

Remark 5.2.1 (Interventional distribution $P_{Y(x)}$ versus conditional distribution $P_{Y|X=x}$)**.** *Model* (5.5) *with a fixed x and η following its marginal distribution induces the interventional distribution of $P_{Y(x)}$, whereas model* (5.5) *in combination with* (5.6)*, that is $f(X, h(Z, X, \xi))$, induces the conditional distribution of $P_{Y|ZX}$. Note that in the latter case, X and the noise term $h(Z, X, \xi)$ are correlated, whereas in the "marginal generative causal model"* (5.5)*, x is fixed and so is independent from the noise term η.*

Remark 5.2.2 (Target estimand)**.** *The target interventional distribution need not be the fully marginal distribution $P_{Y(x)}$; it can be conditional, $P_{Y(x)\,|\,W}$, where W is a chosen set of variables that are held fixed rather than intervened upon. In the static setting of Figure 5.1(b), a common example is $W = Z$.*

Remark 5.2.3 (Privacy)**.** *Our generative framework factorises the simulation process into three variation independent modules, so each can be specified or replaced on its own. In contexts where the variables Z or X contain sensitive information, the model to generate them ($g(\cdot)$) can be replaced by any differentially private generative mechanism. By the post-processing theorem of differential privacy (Dwork et al., 2006), any further mapping applied to a differentially private algorithm remains differentially private. The entire simulated dataset also preserves that privacy*

guarantee, making our approach naturally extendable to privacy-sensitive settings.

5.2.4 Estimation

As in engression, parameters of each component in frengression are typically estimated by minimising the energy loss, for its simplicity and convenience. However, other loss functions can be used, such as KL divergence.

On the population level, the objective function for model (5.4) is

$$\mathcal{L}_{ZX}(g) := \mathcal{L}(P_{ZX}, P_g) = \mathbb{E}\Big[\|(Z,X) - g(\epsilon)\| - \frac{1}{2}\|g(\epsilon) - g(\epsilon')\|\Big]. \tag{5.7}$$

The objective for models (5.5) and (5.6) jointly is

$$\mathcal{L}_{Y|ZX}(f,h) := \mathbb{E}\Big[\|Y - f(X,\widetilde{\eta})\| - \frac{1}{2}\|f(X,\widetilde{\eta}) - f(X,\widetilde{\eta}')\|\Big] + \mathbb{E}\Big[\|\eta - \bar{\eta}\| - \frac{1}{2}\|\bar{\eta} - \bar{\eta}'\|\Big], \tag{5.8}$$

where $\widetilde{\eta} = h(Z, X, \xi)$, $\widetilde{\eta}' = h(Z, X, \xi')$, $\bar{\eta} = h(\bar{Z}, \bar{X}, \xi'')$, and $\bar{\eta}' = h(\bar{Z}', \bar{X}, \xi''')$ with $\eta \sim \mathcal{N}(0, I_{d_y})$, $\xi, \xi', \xi'', \xi''' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_y})$, $(Z,X) \sim P_{ZX}$, $\bar{X} = (\bar{X}_0, \bar{X}_1) \sim P_X$, and, conditional on $\bar{X}_0$, $\bar{Z}, \bar{Z}' \overset{\text{i.i.d.}}{\sim} P_{Z \,|\, X_0 = \bar{X}_0}$. This is chosen because these two models should be estimated such that:

- the object $f(X, \widetilde{\eta})$, when $\widetilde{\eta}$ is drawn from its conditional distribution given (Z, X), matches the observed conditional distribution $P_{Y|ZX}$, which is ensured by the first term in (5.8);
- the second term in (5.8) calibrates h so that the noise it feeds into f follows the reference distribution $\mathcal{N}(0, I_{d_y})$; the precise statement, conditional on X_0 and for P_X-almost every x, is given in Proposition 5.2.7.

As for sampling from $P_{Z|X_0}$, if X_0 does not exist (i.e. the confounder Z is not affected by any treatment as in Figure 5.1(b)), we do it independently from the marginal distribution P_Z; otherwise, we need to learn an auxiliary model $e^*(x_0, \zeta)$ that matches the distribution of $P_{Z|X_0}$ via the standard engression approach

$$e^* \in \operatorname*{argmin}_e \mathbb{E}\Big[\|Z - e(X_0,\zeta)\| - \frac{1}{2}\|e(X_0,\zeta) - e(X_0,\zeta')\|\Big],$$

where $\zeta, \zeta' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_z})$.

We define the population version of frengression, denoted by $\varphi^* = (g^*, f^*, h^*)$, as the solution to minimising the objective functions:

$$g^* = \operatorname*{argmin}_g \mathcal{L}_{ZX}(g)$$
$$(f^*, h^*) = \operatorname*{argmin}_{f,h} \mathcal{L}_{Y|ZX}(f, h).$$

To show the population guarantees that frengression learns each component correctly and recovers the true data generating process of (Z, X, Y), we impose the following assumptions:

Assumption 5.2.4 (Correct specification)**.** *We write the function classes $\mathcal{G} = \{g(\epsilon)\}$, $\mathcal{F} = \{f(x, \eta)\}$, and $\mathcal{H} = \{h(z, x, \xi)\}$. There exist $g' \in \mathcal{G}$, $f' \in \mathcal{F}$, and $h' \in \mathcal{H}$ such that*

(a) $g'(\epsilon) \sim P_{ZX}$ with $\epsilon \sim \mathcal{N}(0, I_{d_x+d_z})$;

(b) $f'(x, h'(z, x, \xi)) \sim P_{Y|Z=z,X=x}$ for P_{ZX}-almost every (z, x), with $\xi \sim \mathcal{N}(0, I_{d_y})$;

(c) for P_X-almost every $x = (x_0, x_1)$, $h'(Z_x, x, \xi) \sim \mathcal{N}(0, I_{d_y})$ when $Z_x \sim P_{Z \mid X_0=x_0}$ and $\xi \sim \mathcal{N}(0, I_{d_y})$.

The following identification conditions adapt those of Assumption 2.1.2.

Assumption 5.2.5 (Identification assumptions regarding DAG in Figure 5.1(a))**.**

(a) ***Positivity:*** *The joint support of (X_0, Z, X_1) is full, that is,*

$$p_{X_0ZX_1}(x_0, z, x_1) > 0$$

for almost all $(x_0, z, x_1) \in \mathcal{X}_0 \times \mathcal{Z} \times \mathcal{X}_1$.

(b) ***Unconfoundedness:*** *The potential outcome $Y(x)$ is independent of the observed treatment given past covariates and treatment:*

$$Y(x) \perp\!\!\!\perp X_1 \mid Z, X_0, \quad \text{and} \quad Y(x) \perp\!\!\!\perp X_0, \quad \forall\, x \in \mathcal{X}.$$

(c) ***Consistency:***

If $X = x$ then the observed outcome equals the potential outcome at that level:

$$X = x \quad \text{implies} \quad Y = Y(x), \quad \forall\, x \in \mathcal{X}.$$

When Assumption 5.2.5 holds, the potential outcome distribution $P_{Y(x)}$ is identifiable from the observed joint distribution of (Z, X, Y) in Figure 5.1(a):

Lemma 5.2.6 (Identifiability)**.** *When Assumption 5.2.5 is satisfied, we have:*

$$p_{Y(x)}(y) = \int_{\mathcal{Z}} p_{Y|X_0ZX_1}(y \mid x_0, z, x_1)\, p_{Z|X_0}(z \mid x_0)\, \mathrm{d}z.$$

The proof can be found in Appendix C.2.1. In Section 5.4 we show that a pre-additive noise structure extrapolates the median of the causal margin beyond the observed treatment support, relaxing positivity for this functional.

Assumption 5.2.4 asserts the existence of functions (g', f', h'), while Assumption 5.2.5 guarantees unique potential outcome distributions. With these assumptions, we introduce the following proposition, which provides the population guarantee that frengression correctly learns the true distributions when the model is correctly specified.

Proposition 5.2.7 (Correctness)**.** *When Assumptions 5.2.4 and 5.2.5 hold, the frengression solution $\varphi^* = (g^*, f^*, h^*)$ satisfies, for P_{ZX}-almost every (z, x) and P_X-almost every x,*

$$\begin{aligned} g^*(\epsilon) &\sim P_{ZX}, \\ f^*(x, h^*(z, x, \xi)) &\sim P_{Y|Z=z,X=x}, \\ f^*(x, \eta) &\sim P_{Y(x)}, \end{aligned}$$

where $\epsilon \sim \mathcal{N}(0, I_{d_z+d_x})$, $\eta \sim \mathcal{N}(0, I_{d_y})$, and $\xi \sim \mathcal{N}(0, I_{d_y})$.

The proof of Proposition 5.2.7 is provided in Appendix C.2.2.

5.2.4.1 Finite-sample frengression

φ^* solves the population objectives in (5.7) and (5.8). In practice, we only have access to finite samples $\{(Z_i, X_i, Y_i)\}_{i=1}^n \sim P_{ZXY}(z, x, y)$; the empirical version of frengression is

$$\begin{aligned} \hat{g} &= \operatorname*{argmin}_g \widehat{\mathcal{L}}_{ZX}(g) \\ (\hat{f}, \hat{h}) &= \operatorname*{argmin}_{f,h} \widehat{\mathcal{L}}_{Y|ZX}(f, h), \end{aligned}$$

where the empirical loss function estimates are

$$\widehat{\mathcal{L}}_{ZX}(g) := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{j=1}^{m}\left\|(X_i, Z_i) - g(\epsilon_{i,j})\right\| - \frac{1}{2m(m-1)}\sum_{j=1}^{m}\sum_{\substack{j'=1\\ j'\neq j}}^{m}\left\|g(\epsilon_{i,j}) - g(\epsilon_{i,j'})\right\|\right], \tag{5.9}$$

$$\begin{aligned}\widehat{\mathcal{L}}_{Y|ZX}(f,h) := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{j=1}^{m}\left\|Y_i - f(X_i, \widetilde{\eta}_{i,j})\right\| - \frac{1}{2m(m-1)}\sum_{j=1}^{m}\sum_{\substack{k=1\\ k\neq j}}^{m}\left\|f(X_i, \widetilde{\eta}_{i,j}) - f(X_i, \widetilde{\eta}_{i,k})\right\|\right]\\ + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{j=1}^{m}\left\|\eta_i - \bar{\eta}_{i,j}\right\| - \frac{1}{2m(m-1)}\sum_{j=1}^{m}\sum_{\substack{k=1\\ k\neq j}}^{m}\left\|\bar{\eta}_{i,j} - \bar{\eta}_{i,k}\right\|\right],\end{aligned} \tag{5.10}$$

respectively. Here, for $i = 1, \ldots, n$ and $l = 1, \ldots, m$, we simulate $\epsilon_{i,l} \sim \mathcal{N}(0, I_{d_z+d_x})$, $\eta_i \sim \mathcal{N}(0, I_{d_y})$ and $\xi_{i,l} \sim \mathcal{N}(0, I_{d_y})$, and set $\widetilde{\eta}_{i,l} = h(Z_i, X_i, \xi_{i,l})$ and $\bar{\eta}_{i,l} = h(Z'_{i,l}, X_i, \xi_{i,l})$. The draws $Z'_{i,l}$ follow the same conditional distribution as $\bar{Z}$ in (5.8): if X_0 is absent (the common time-fixed setting of Figure 5.1(b)), no auxiliary model is needed and $Z'_{i,l}$ is taken from a random permutation of the observed Z's, which estimates P_Z; only when X_0 is present are the draws taken from the auxiliary model, $Z'_{i,l} = e(X_{0i}, \zeta_{i,l})$, estimating $P_{Z \mid X_0 = X_{0i}}$. Only in the latter case, the auxiliary model e for $P_{Z \mid X_0}$ is fitted by minimising

$$\begin{aligned}\widehat{\mathcal{L}}_{Z|X_0}(e) := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{j=1}^{m}\left\|Z_i - e(X_{0i}, \zeta_{i,j})\right\|\right.\\ \left. - \frac{1}{2m(m-1)}\sum_{j=1}^{m}\sum_{\substack{j'=1\\ j'\neq j}}^{m}\left\|e(X_{0i}, \zeta_{i,j}) - e(X_{0i}, \zeta_{i,j'})\right\|\right],\end{aligned} \tag{5.11}$$

where $\zeta_{i,l} \sim \mathcal{N}(0, I_{d_z})$, for $i = 1, \ldots, n$, $l = 1, \ldots, m$.

The empirical losses (5.9) and (5.10) estimate the corresponding population objectives; the second term of (5.10) is unbiased when the auxiliary draws are taken from the true conditional distribution $P_{Z \mid X_0}$, which (5.11) estimates when X_0 is present. In Section 5.3, we present theoretical guarantees that, by minimising the empirical losses, the fitted generative distribution converges to the distribution within the model class that is closest to the target in energy distance.

In practice, we parameterise g, f and h with neural networks. The resulting empirical objectives are therefore almost everywhere differentiable with respect to every model parameter, allowing us to optimise them with stochastic gradient descent to obtain the fitted frengression model.

5.2.5 Sampling

Given the generative nature of our model, sampling from a fitted frengression is straightforward. We provide the sampling procedures from the population version φ^* for four common simulation scenarios below:

1. Simulate $(\widehat{Z}, \widehat{X})$.
   Sample $\epsilon \sim \mathcal{N}(0, I_{d_z+d_x})$ and set $(\widehat{Z}, \widehat{X}) = g^*(\epsilon)$.

2. Simulate $\widehat{Y}(x)$.
   For a treatment value x, draw $\eta \sim \mathcal{N}(0, I_{d_y})$ and compute $\widehat{Y}(x) = f^*(x, \eta)$.

3. Simulate $(\widehat{Z}, \widehat{X}, \widehat{Y})$ from the joint distribution of (Z, X, Y).

   (i) Generate $(\widehat{Z}, \widehat{X})$ as in step 1.

   (ii) Draw $\xi \sim \mathcal{N}(0, I_{d_y})$ and obtain $\widetilde{\eta} = h^*(\widehat{Z}, \widehat{X}, \xi)$.

   (iii) Produce the response via $\widehat{Y} = f^*(\widehat{X}, \widetilde{\eta})$.

4. Simulate observed $(\widehat{Z}, \widehat{X})$ together with a potential outcome $\widehat{Y}(x')$ (as in single world intervention graphs, Richardson and Robins, 2013):

   (i) Generate $(\widehat{Z}, \widehat{X})$ and obtain $\widetilde{\eta} = h^*(\widehat{Z}, \widehat{X}, \xi)$.

   (ii) Produce the response via $f^*(x', \widetilde{\eta})$ with a specified x'.

The (g^*, f^*, h^*) can be replaced by model $(\hat{g}, \hat{f}, \hat{h})$ fitted using finite samples. Taken together, these sampling schemes allow us to generate synthetic observations from any component or the full joint distribution of a model fitted on real data.

5.2.6 Sampling-based inference

Frengression allows sampling-based inference. For instance, a typical quantity of interest is the average potential outcome $\mathbb{E}Y(x)$ under a specific intervention x. A fitted frengression model contains $\hat{f}(x, \eta)$, and this estimates the marginal outcome distribution. Thus, for any given intervention value x, we sample $\eta_i \sim \mathcal{N}(0, I_{d_y})$ for $i = 1, \ldots, m$. Then we obtain a sample $\{Y_i = \hat{f}(x, \eta_i), i = 1, \ldots, m\}$ from the estimated distribution $\hat{P}_{Y(x)}$. The expected potential outcome can thus be estimated using the empirical mean $\frac{1}{m}\sum_{i=1}^{m} \hat{f}(x, \eta_i)$. Besides the interventional mean, any other quantity can also be estimated; for instance, the α-quantile of $P_{Y(x)}$ is consistently estimated by the α-quantile of this sample.

5.3 Model consistency

In this section we first provide a general consistency result for generative modelling with the empirical energy distance as the loss function, and then adapt it to the components of frengression. The generative model g_θ (distinct from the component generator g of frengression), parameterised by $\theta \in \Theta$,[12] takes a random noise input ν, independent of the data and drawn from a fixed reference distribution; we assume it is a standard Gaussian, though any other distribution under which the assumptions below hold may be used instead. The model returns the sample $g_\theta(\nu)$, and we write P_θ for its distribution, the distribution it generates. The aim is to learn an unknown data-generating distribution Q, and since frengression is parameterised by a neural network, θ collects the network weights.

In this chapter, we train the generative model g_θ by minimising the energy distance between P_θ and Q. Following the definition in (5.2), we write the population energy distance as

$$\mathrm{ED}(P_\theta, Q) = 2\mathbb{E}_{X\sim P_\theta, Y\sim Q}\|X - Y\| - \mathbb{E}_{X,X'\sim P_\theta}\|X - X'\| - \mathbb{E}_{Y,Y'\sim Q}\|Y - Y'\|. \tag{5.12}$$

We make the following assumption:

Assumption 5.3.1 (Existence of minimisers)**.** *We assume that every energy-distance minimisation over Θ in this section attains its minimum: there exists a population minimiser*

$$\theta^* = \arg\min_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q),$$

and the empirical objectives $\mathrm{ED}(P_\theta, Q^n)$ and $\mathrm{ED}(P_\theta^m, Q^n)$ introduced below likewise attain their minima, denoted $\hat{\theta}_n$ and $\hat{\theta}_{m,n}$.

Here P_{θ^*} is the distribution in the model class $\{P_\theta : \theta \in \Theta\}$ closest to Q in energy distance.

In practice, it is not common to have direct access to Q; instead, we generally have access to empirical observations from Q: $\{y_1, \ldots, y_n\}$, forming the empirical measure $Q^n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. When g_θ is analytically accessible, an estimator could be found as

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q^n),$$

[12] Within this chapter θ denotes the parameters of the generative model, distinct from the average treatment effect θ of Chapter 3. Accordingly, P_θ is indexed by this parameter rather than being the distribution of a random variable (Section 2.5).

where

$$\mathrm{ED}(P_\theta, Q^n) = \frac{2}{n}\sum_{j=1}^{n}\mathbb{E}_{X\sim P_\theta}\|X - y_j\| - \mathbb{E}_{X,X'\sim P_\theta}\|X - X'\| - \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j\neq j'}}^{n}\|y_j - y_{j'}\|.$$

If g_θ is not directly accessible, but samples can be independently drawn from it, the estimator then becomes

$$\hat{\theta}_{m,n} = \arg\min_{\theta\in\Theta}\mathrm{ED}(P_\theta^m, Q^n),$$

where

$$\mathrm{ED}(P_\theta^m, Q^n) = \frac{2}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}\|x_i - y_j\| - \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{\substack{i'=1\\ i'\neq i}}^{m}\|x_i - x_{i'}\| - \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{\substack{j'=1\\ j\neq j'}}^{n}\|y_j - y_{j'}\|.$$

$\mathrm{ED}(P_\theta^m, Q^n)$ aligns with the empirical objective functions defined in Section 5.2.4.1.

Under the assumption that the minimiser $\theta^* = \mathrm{argmin}_{\theta\in\Theta}\,\mathrm{MMD}(P_\theta, Q)$ exists and when the kernel used in the MMD is bounded, Briol et al. (2019) proved the consistency of $\hat{\theta}_n$ and $\hat{\theta}_{m,n}$, i.e. $\lim_{n\to\infty}\hat{\theta}_n = \theta^*$ and $\lim_{m,n\to\infty}\hat{\theta}_{m,n} = \theta^*$. However, frengression is trained using energy distance, which is a special case of MMD employing an unbounded, Euclidean distance-based kernel. Although it is reasonable to argue that a bounded data space may lead to a bounded kernel in practice, our goal is to extend the consistency guarantee of Briol et al. to the unbounded kernels used in energy distance, under tail conditions on the data, at the level of the generated distribution rather than the parameter. Here we provide the consistency result under the following alternative two assumptions.

Assumption 5.3.2 (Sub-exponential tails)**.** *There exist constants $C, \kappa > 0$ such that, for every probability measure P appearing in the analysis, a random variable $X \sim P$ satisfies*

$$\Pr(\|X\| > t) \leq C\exp(-\kappa t), \qquad \forall t > 0.$$

We first present the convergence result based on energy distance:

Theorem 5.3.3. *Suppose:*

(a) Assumption 5.3.1 and Assumption 5.3.2 hold;

(b) the model is correctly specified, that is, $P_{\theta^*} \overset{d}{=} Q$.

Then

$$\mathrm{ED}\left\{P_{\hat{\theta}_n}, Q\right\} \to 0 \quad \textit{in probability,} \qquad \textit{and hence} \qquad P_{\hat{\theta}_n} \rightsquigarrow Q \quad \textit{in probability.}$$

If, in addition,

(c) Θ *is compact, and*

(d) for each ν*, the map* $\theta \mapsto g_\theta(\nu)$ *is Lipschitz on* Θ*, that is,* $\|g_\theta(\nu) - g_{\theta'}(\nu)\| \leq L(\nu)\,\|\theta - \theta'\|$ *for all* $\theta, \theta' \in \Theta$*, where* $L(\nu)$ *is measurable with* $\mathbb{E}[\sqrt{L(\nu)}] < \infty$*,*

then the same conclusions hold for the finite-sample estimator $\hat{\theta}_{m,n}$ *as* $m, n \to \infty$*.*

Theorem 5.3.3 transfers to the frengression components through the integrated objectives (5.9) and (5.10): under conditions (a)–(d), the empirical minimiser approaches the set of population minimisers, each of which induces the true conditional and causal-margin distributions by Proposition 5.2.7.

Corollary 5.3.4 (Convergence of frengression components)**.** *Let* $(\hat{g}, \hat{f}, \hat{h})$ *be the frengression estimators obtained by minimising the empirical objectives* (5.9) *and* (5.10)*. Suppose Assumptions 5.2.4 and 5.2.5 hold, together with conditions (a)–(d) of Theorem 5.3.3. Then, as* $m, n \to \infty$*, the following weak convergences hold in probability, for* P_{ZX}*-almost every* (z, x) *and* P_X*-almost every* x*:*

$$\begin{aligned} \hat{g}(\epsilon) &\rightsquigarrow P_{ZX}, \\ \hat{f}\big(x, \hat{h}(z, x, \xi)\big) &\rightsquigarrow P_{Y \mid Z=z, X=x}, \\ \hat{f}(x, \eta) &\rightsquigarrow P_{Y(x)}, \end{aligned}$$

where $\epsilon \sim \mathcal{N}(0, I_{d_z+d_x})$*,* $\eta \sim \mathcal{N}(0, I_{d_y})$ *and* $\xi \sim \mathcal{N}(0, I_{d_y})$ *are drawn independently of the data.*

The proofs can be found in Appendices C.2.3 and C.2.4.

Because the model is parameterised by a neural network, the minimiser θ^* need not be unique: distinct weights can induce the same P_θ. We therefore state consistency at the level of distributions, as in Theorem 5.3.3, rather than for the parameter θ itself. Correct specification enters only to identify the limit as Q; without it, the same argument gives

$$\mathrm{ED}(P_{\hat{\theta}_n}, Q) - \inf_{\theta \in \Theta} \mathrm{ED}(P_\theta, Q) \to 0 \quad \text{in probability,}$$

so a fitted frengression attains the smallest energy distance to Q achievable within the model class.

5.4 Extrapolation

For simplicity, in this section, we consider the case with univariate treatment X and outcome Y but allow for multiple covariates Z.

Assumption 5.4.1 (Pre-ANM)**.** *The causal margin can be parameterised via a pre-additive noise model*

$$Y = \widetilde{f}(X + \widetilde{\eta}),$$

where $\widetilde{f}$ is any strictly monotone and continuously differentiable function and $\widetilde{\eta}$ has a continuous, positive density on its support and is independent of X. Without loss of generality, assume $\widetilde{\eta}$ has a median 0 and is supported on $[-\widetilde{\eta}_0, \widetilde{\eta}_0]$ for some $\widetilde{\eta}_0 > 0$; we let $\widetilde{\eta}_0 = \infty$ when $\widetilde{\eta}$ has an unbounded support.

We study the extrapolation performance of frengression for median estimation of the causal margin. Let $\widetilde{m}(x)$ denote the true median of $Y(x)$ for any x. When Assumption 5.4.1 holds, the frengression model class can be smaller for Assumption 5.2.4 to hold. Specifically, we reparameterise the model for the causal margin as $f(X + f_1(\eta))$, where f and f_1 are both strictly monotone and continuously differentiable functions, with the normalisation $f_1(0) = 0$, without loss of generality (Shen and Meinshausen, 2024). Denote the median of the population frengression solution as $m^*(x)$, which is defined as the median of $f^*(x, \eta)$.

The following proposition indicates that frengression can recover the true median of causal margin for x outside of the support. The proof can be found in Appendix C.2.5.

Proposition 5.4.2 (Extrapolation)**.** *Let the support of X be $\mathcal{X} = [x_{\min}, x_{\max}]$. Under Assumptions 5.2.4, 5.2.5, and 5.4.1, we have*

$$m^*(x) = \widetilde{m}(x)$$

for all $x \in [x_{\min} - \widetilde{\eta}_0, x_{\max} + \widetilde{\eta}_0]$.

The experimental results provided in Figure 5.4 show that frengression maintains accurate estimates even when extrapolating beyond the observed values of treatment X.

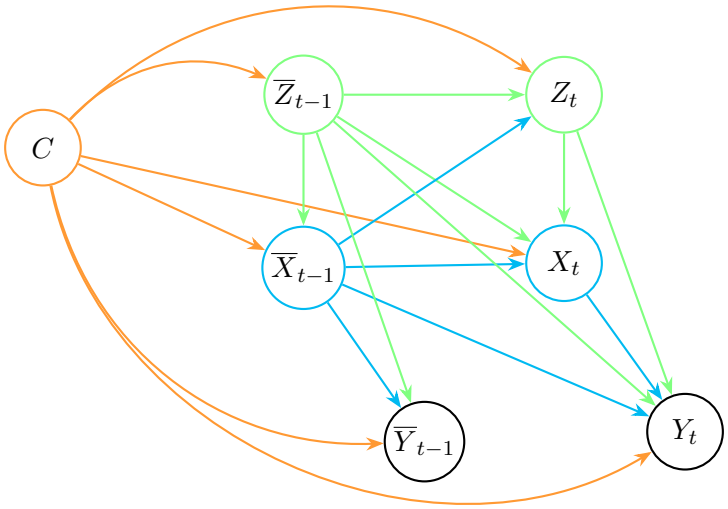


Figure 5.3: Example diagram of longitudinal data. For simplicity we present one time point t as an example, and we show that (Z_t, X_t, Y_t) are connected with all previous time steps. As with the static case, there are no dimensional constraints on the variables.

5.5 Frengression on sequential data

The discussion in Sections 5.2 and 5.3 introduced the basic frengression machinery in the static case. Many scientific questions, however, involve time-varying decision points, such as longitudinal studies and survival analysis, where treatment decisions may change over time in response to evolving patient histories. The frugal parameterisation is particularly helpful in longitudinal analysis, as the time-varying covariates are marginalised out. Leveraging this idea, especially the recursive/nested variant of frugal parameterisation introduced in Section 6 of Evans and Didelez (2024), frengression extends naturally to the longitudinal context.

An illustrative causal diagram is given in Figure 5.3. Let $C \in \mathbb{R}^{d_c}$ denote the static, pre-treatment baseline covariates; for the ease of notation, we use Z, X, Y to denote the vector: $Z = (Z_0, \ldots, Z_{T-1})$ with $Z_t \in \mathbb{R}^{d_z}$ the time-varying covariates; $X = (X_0, \ldots, X_{T-1})$, with $X_t \in \mathbb{R}^{d_x}$ the sequence of interventions; and $Y = (Y_0, \ldots, Y_{T-1})$, $Y_t \in \mathbb{R}^{d_y}$ the observed outcomes. In the special case of survival-valued outcomes, $Y_t \in \{0, 1\}$ with $Y_t = 1$ indicating an event occurrence; simulation stops after such an event. Other than survival variables, all can assume continuous or discrete values.

5.5.1 Model

We illustrate the following sections with an example where the interventional distribution of interest is $P_{Y(x)|C}$.

For data as described above, we write the sequential frengression as φ_s. We first build a model to generate the baseline covariates C using the unconditional engression $g_c(\varsigma)$, where $\varsigma \sim \mathcal{N}(0, I_{d_c})$. Then we build the sequential frengression $(\varphi_0, \ldots, \varphi_{T-1})$, where $\varphi_t = (g_t, f_t, h_t)$ for $t = 0, \ldots, T-1$. For each $t \in \{0, \ldots, T-1\}$:

- When $t = 0$, build $g_0(c, \epsilon)$ for the joint distribution $P_{Z_0 X_0|C}$; otherwise $g_t(c, \overline{z}_{t-1}, \overline{x}_{t-1}, \epsilon)$ for the joint distribution $P_{Z_t X_t|C\overline{Z}_{t-1}\overline{X}_{t-1}}$. In both cases, $\epsilon \sim \mathcal{N}(0, I_{d_z+d_x})$.
- $Y = f_t(c, \overline{x}_t, \eta)$, $\eta \sim \mathcal{N}(0, I_{d_y})$ for the interventional distribution $P_{Y_t(\overline{x}_t)|C}$ of treatment values up to time t given baseline covariates C.
- $\widetilde{\eta} = h_t(c, \overline{z}_t, \overline{x}_t, \xi)$ where $\xi \sim \mathcal{N}(0, I_{d_y})$; this term is used for modelling the association.

The sequential frengression consists of all the components above: $\varphi_s = (g_c(\varsigma), \varphi_0, \varphi_1, \ldots, \varphi_{T-1})$. For convenience, we also denote $g = (g_0, \ldots, g_{T-1})$, $f = (f_0, \ldots, f_{T-1})$ and $h = (h_0, \ldots, h_{T-1})$, then φ_s can also be expressed as $\varphi_s = (g_c, g, f, h)$.

Survival data require one further adjustment, because the effective sample size shrinks over time as events occur. Accordingly, each φ_t is trained on the subpopulation that is still at risk at time t. This renormalisation does not affect the generative procedure, because at simulation time we censor a synthetic trajectory as soon as an event occurs and remove it from all subsequent steps, ensuring that the simulated data follow the same at-risk pattern as the training cohort. In this way frengression extends to right-censored settings while preserving its generative flexibility.

5.5.2 Estimation

The population objective for g_c is the analogue of (5.7) with (Z_i, X_i) replaced by C_i. Estimation objectives of g, f, h are the same as in (5.7) and (5.8) respectively, only now the outcome variable contains T time steps of dimension $d_y \times T$, and here we specify causal estimand of interest to be conditional on baseline covariates, which can be customised according to the specific question of interest. More formally, the population solution to φ_s can be expressed as:

$$g_c^* := \operatorname*{argmin}_{g_c} \mathbb{E}\Big[\|C - g_c(\epsilon_c)\| - \frac{1}{2}\|g_c(\epsilon_c) - g_c(\epsilon_c')\|\Big]$$
$$g^* := (g_0^*, \dots, g_{T-1}^*) = \operatorname*{argmin}_{g} \mathbb{E}\Big[\|(Z, X) - g(C, \epsilon)\| - \frac{1}{2}\|g(C, \epsilon) - g(C, \epsilon')\|\Big]$$
$$(f^*, h^*) := (f_0^*, \dots, f_{T-1}^*, h_0^*, \dots, h_{T-1}^*)$$
$$= \operatorname*{argmin}_{f,h} \mathbb{E}\Big[\|Y - f(C, X, \widetilde{\eta})\| - \frac{1}{2}\|f(C, X, \widetilde{\eta}) - f(C, X, \widetilde{\eta}')\|\Big] + \mathbb{E}\Big[\|\eta - \bar{\eta}\| - \frac{1}{2}\|\bar{\eta} - \bar{\eta}'\|\Big]$$

Here $g(C, \epsilon)$ denotes the sequential generator obtained by composing $g_0, \dots, g_{T-1}$ recursively over time, where $\epsilon_c, \epsilon_c' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_c})$; $\epsilon, \epsilon' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{(d_z+d_x)\times T})$; $\widetilde{\eta} = h(C, Z, X, \xi)$, $\widetilde{\eta}' = h(C, Z, X, \xi')$, $\bar{\eta} = h(C, \bar{Z}, \bar{X}, \xi'')$, and $\bar{\eta}' = h(C, \bar{Z}', \bar{X}, \xi''')$ with $\eta \sim \mathcal{N}(0, I_{d_y \times T})$, $\xi, \xi', \xi'', \xi''' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_y \times T})$, $(Z, X) \sim P_{ZX|C}$, $\bar{X} \sim P_X$ and $\bar{Z}, \bar{Z}' \overset{\text{i.i.d.}}{\sim} P_{Z|C\overline{X}_{T-2}}$. Similar to that in the static setting, we learn auxiliary models $e := (e_0, e_1, \dots e_{T-1})$ to match $P_{Z|C\overline{X}_{T-2}}$ via standard engression models at each time step t:

$$e_0^* := \operatorname*{argmin}_{e_0} \mathbb{E}\Big[\|Z_0 - e_0(C, \zeta_0)\| - \frac{1}{2}\|e_0(C, \zeta_0) - e_0(C, \zeta_0')\|\Big]$$
$$\text{and} \quad e_t^* := \operatorname*{argmin}_{e_t} \mathbb{E}\Big[\|Z_t - e_t(C, \overline{X}_{t-1}, \zeta_t)\| - \frac{1}{2}\|e_t(C, \overline{X}_{t-1}, \zeta_t) - e_t(C, \overline{X}_{t-1}, \zeta_t')\|\Big],$$

where $\zeta_0, \zeta_0', \zeta_t, \zeta_t' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_z})$, for $t \in \{1, \dots, T-1\}$.

Because treatments vary over time, we impose the standard *sequential conditional exchangeability* assumption to ensure identification. For detailed discussions, see Hernán and Robins (2025) and Evans and Didelez (2024, Remark 1.5; Sections 6–7).

Corollary 5.5.1 (Sequential model preserves correctness)**.** *If at each time point the static frengression model satisfies the correctness property (Proposition 5.2.7), then the joint sequential model also satisfies correctness.*

Corollary 5.5.2 (Sequential model preserves consistency)**.** *If at each time point the static frengression model enjoys distributional consistency (Theorem 5.3.3), then the joint sequential model is likewise consistent.*

Corollary 5.5.3 (Sequential model preserves extrapolation)**.** *If at each time point the static frengression model allows extrapolation on continuous treatments (Proposition 5.4.2), then the joint sequential model also admits extrapolation on continuous treatments at every time point.*

Proof of Corollaries 5.5.1–5.5.3. The proof proceeds by the induction on t. For the base case $t = 0$, the claims reduce to the static frengression result. Now suppose the statement is true

for all time steps up to $t-1$. Conditional on the history $(C, \overline{Z}_{t-1}, \overline{X}_{t-1})$, the component φ_t is itself a static frengression model with covariates $(C, \overline{Z}_{t-1}, \overline{X}_{t-1})$ and outcome block (Z_t, X_t, Y_t), so the corresponding static result applies (Proposition 5.2.7 for correctness, Theorem 5.3.3 for consistency, and Proposition 5.4.2 for extrapolation); together with the inductive hypothesis for the history, this establishes the property at time t. □

5.5.3 Sampling

The sampling works similarly to the static frengression with sequential chaining. Similar to Section 5.2.5, we provide sampling from φ_s^* for the following scenarios.

1. Simulate $\widehat{C}$: Draw $\varsigma \sim \mathcal{N}(0, I_{d_c})$ and apply the transformation $g_c^*(\varsigma)$.

2. Simulate $(\widehat{Z}, \widehat{X}) \in \mathbb{R}^{T\times(d_z+d_x)}$: Specify c (can be simulated from $g_c^*(\varsigma)$) and sample $\epsilon_0 \sim \mathcal{N}(0, I_{d_z+d_x})$; apply $g_0^*(\cdot)$ on (c, ϵ_0) to get $(\widehat{Z}_0, \widehat{X}_0)$. Then, for $t = 1, \ldots, T-1$: simulate $\epsilon_t \sim \mathcal{N}(0, I_{d_z+d_x})$, then $(\widehat{Z}_t, \widehat{X}_t) = g_t^*(c, \widehat{\overline{Z}}_{t-1}, \widehat{\overline{X}}_{t-1}, \epsilon_t)$. Putting them together we obtain $(\widehat{Z}, \widehat{X}) \in \mathbb{R}^{T\times(d_z+d_x)}$.

3. Simulate $\widehat{Y}(\overline{x}_{T-1})$: specify c (which can be simulated from g_c^*). For $t = 0, \ldots, T-1$, sample $\eta_t \sim \mathcal{N}(0, I_{d_y})$, then apply the transformation f_t^* with sampled η_t, specified c, and specified $\overline{x}_t$ to get $\widehat{Y}_t = f_t^*(c, \overline{x}_t, \eta_t)$. Putting them together, we obtain $\widehat{Y} = (\widehat{Y}_0, \ldots, \widehat{Y}_{T-1})$.

4. Simulate from φ_s to obtain $(\widehat{C}, \widehat{\overline{Z}}_t, \widehat{\overline{X}}_t, \widehat{\overline{Y}}_t) \in \mathbb{R}^{d_c+T\times(d_z+d_x+d_y)}$: follow the instructions above to obtain $\widehat{C}$, $\widehat{Z}$, $\widehat{X}$. For $t = 0, \ldots, T-1$, sample $\xi_t \sim \mathcal{N}(0, I_{d_y})$, and apply the transformation h_t^* on sampled $(\widehat{C}, \widehat{\overline{Z}}_t, \widehat{\overline{X}}_t, \xi_t)$ to get $\widetilde{\eta}_t = h_t^*(\widehat{C}, \widehat{\overline{Z}}_t, \widehat{\overline{X}}_t, \xi_t)$. We obtain $\widehat{Y}_t$ via $\widehat{Y}_t = f_t^*(\widehat{C}, \widehat{\overline{X}}_t, \widetilde{\eta}_t)$.

Proposition 5.5.4 (Simulate data from joint distribution)**.** *We take sequential conditional exchangeability and that Assumption 5.2.4 holds. The simulation procedure stated above generates a draw P_φ of (C, Z, X, Y) whose distribution equals the observed joint distribution of (C, Z, X, Y).*

Proof. We proceed by induction on the time index t. For $t = 0$, first sample $\widehat{C} \sim P_{g_c}$. Conditional on $\widehat{C}$, draw $(\widehat{Z}_0, \widehat{X}_0) \sim P_{Z_0X_0 \,|\, C=\widehat{C}}$ by applying $g_0^*(\cdot)$ on $(\widehat{C}, \epsilon_0)$, where $\epsilon_0 \sim \mathcal{N}(0, I_{d_z+d_x})$. Draw $\xi_0 \sim \mathcal{N}(0, I_{d_y})$, then apply h_0^* on sampled $\widehat{C}, \widehat{Z}_0, \widehat{X}_0, \xi_0$ to get $\widetilde{\eta}_0$. Obtain $\widehat{Y}_0$ by applying $f_0^*(\cdot)$ on $(\widehat{C}, \widehat{X}_0, \widetilde{\eta}_0)$. Guaranteed by Proposition 5.2.7, $(\widehat{Z}_0, \widehat{X}_0) \sim P_{Z_0X_0|C=\widehat{C}}$, $\widehat{Y}_0 \sim P_{Y_0 \,|\, C=\widehat{C}, Z_0=\widehat{Z}_0, X_0=\widehat{X}_0}$. This takes care of the base case.

Suppose that for $t \geq 1$, $(\widehat{C}, \widehat{\overline{Z}}_{t-1}, \widehat{\overline{X}}_{t-1}, \widehat{\overline{Y}}_{t-1}) \sim (C, \overline{Z}_{t-1}, \overline{X}_{t-1}, \overline{Y}_{t-1})$. We now extend to time t by drawing $(\widehat{Z}_t, \widehat{X}_t)$ from the distribution of $(Z_t, X_t \,|\, C = \widehat{C}, \overline{Z}_{t-1} = \widehat{\overline{Z}}_{t-1}, \overline{X}_{t-1} = \widehat{\overline{X}}_{t-1})$ and then $\widehat{Y}_t$ is sampled from $Y_t \,|\, C = \widehat{C}, \overline{Z}_t = \widehat{\overline{Z}}_t, \overline{X}_t = \widehat{\overline{X}}_t$, using the procedure stated in the above simulation settings. Applying the chain rule yields $(\widehat{C}, \widehat{\overline{Z}}_t, \widehat{\overline{X}}_t, \widehat{\overline{Y}}_t) \sim (C, \overline{Z}_t, \overline{X}_t, \overline{Y}_t)$, which completes the inductive step. □

5.6 Numerical examples on supported settings

In this section, we evaluate frengression on synthetic benchmarks under both static and time-varying treatment regimes. Our evaluation focuses on two key aspects:

1. *Estimation and inference.* We evaluate how precisely frengression recovers the underlying data-generating process, especially the marginal interventional quantities. To establish ground truth, we either use existing datasets or simulate new ones with known interventional distributions, employing two packages built on frugal parameterisation: the `causl` package (Evans, 2021a) for static treatment settings, and the `survivl` package (Evans, 2021b) for longitudinal and survival data. Simulating data this way gives us exact control over the marginal quantity targets.

2. *Simulation performance.* We assess frengression's ability to sample new data that both preserves the original data's properties and matches the specified marginal interventional distributions. By comparing the generated samples with the original targets, we verify that our model can capture and replicate complex distributional patterns.

In the experiments that follow, we compare frengression against leading, specialised methods in each domain. While these competitors excel in their specific settings, none combine the estimation and simulation on static treatments, longitudinal dynamics, survival outcomes, and continuous treatments in a single framework. The code for reproducibility and package can be found at https://github.com/linyingyang/frengression_pkg.

5.6.1 Binary intervention

Estimating a treatment effect under a binary intervention via frengression does not require propensity score estimation or adjustment; instead, it directly models the marginal interventional quantity of interest. This contrasts with methods that average outcome-regression pre-

dictions over the covariate distribution, whose performance can suffer when certain covariate combinations are rare or poorly represented, as discussed in Chapter 3. By focusing on the marginal causal parameter, frengression avoids the instability caused by extreme propensity score values. We empirically assess its robustness under weak overlap scenarios, especially in finite samples, to demonstrate its ability to recover causal effects even when classical overlap conditions are violated.

We simulate $d_z = 10$ baseline covariates $Z = (I, C)$, where the instrumental variables $I \in \mathbb{R}^5$ and the confounders $C \in \mathbb{R}^5$ are generated independently as standard normals. Treatment assignment follows $X \mid Z \sim \text{Bernoulli}(\pi(Z))$ with $\pi(Z) = \text{expit}(\beta_I^\top I + \mathbf{1}^\top C)$, $\beta_I \in \mathbb{R}^5$ with all coordinates equal, and the outcome is $Y(x) \sim \mathcal{N}(2x, 1)$, with Gaussian copulae between each confounder C_j, $j = 1, \ldots, 5$, and Y with correlation $\rho = 2\,\text{expit}(1) - 1 \approx 0.46$.[13] By increasing the magnitude of the instrumental coefficients β_I, we push the propensity scores toward 0 or 1, creating a scenario of weak overlap.

For each replication, we simulate $N = 5000$ observations and repeat the experiment for $K = 30$ replications. We report the bias, MAE, and RMSE of the ATE estimates. We compare frengression against the AIPW estimator (with logistic regression to estimate propensity score and Random Forests as the outcome regression model), Dragonnet, and CausalEGM. TARNet, CFRNet and CEVAE demonstrate sensitivity to hyperparameters and perform poorly, hence we exclude these results. Furthermore, these methods were primarily proposed for conditional treatment effect quantities, rather than marginal causal effects.

Results in Table 5.2 show that the AIPW estimator's variance increases as overlap worsens, while frengression remains stable. Even without per-dataset hyperparameter tuning, frengression matches or outperforms Dragonnet and CausalEGM in bias, MAE, and RMSE across all regimes. We present the experiment details in Appendix C.4.1. Weak overlap is mainly discussed in Chapter 3; our aim here is not to add variance theory but to demonstrate the flexibility of frengression, whose marginal modelling handles binary treatment and stays stable under weak overlap. The strong empirical performance suggests frengression is useful in this regime, with the dedicated variance analysis left to Chapter 3.

[13] Within this chapter ρ denotes a correlation, distinct from the equivalence radius of Chapter 4.

β_I	metric	Frengression	AIPW	CausalEGM	Dragonnet
0.0	RMSE	**0.056**	0.290	0.151	0.070
	Bias	**−0.000**	0.225	−0.117	0.031
	MAE	**0.048**	0.253	0.130	0.056
0.5	RMSE	**0.076**	0.290	0.149	0.114
	Bias	**−0.019**	0.258	−0.124	0.042
	MAE	**0.059**	0.258	0.132	0.100
1.0	RMSE	0.071	0.849	0.193	**0.055**
	Bias	**−0.015**	0.480	−0.173	**0.015**
	MAE	0.056	0.480	0.173	**0.039**
1.5	RMSE	**0.086**	0.690	0.262	0.155
	Bias	**−0.028**	0.530	−0.245	0.080
	MAE	**0.069**	0.543	0.245	0.117
2.0	RMSE	**0.084**	0.524	0.287	0.113
	Bias	**−0.026**	0.437	−0.274	0.028
	MAE	**0.071**	0.466	0.274	0.095

Table 5.2: Estimation performance across different degrees of overlap.

5.6.2 Continuous treatment

In the context of continuous treatment, obtaining the average dose-response function (ADRF), defined as $\mu(x) = \mathbb{E}_Z[\mathbb{E}(Y(x) \mid Z)]$, is typically of interest. To evaluate the models' performance, we use RMSE and mean absolute percentage error (MAPE).

We follow Liu et al. (2024) in the synthetic experiments for continuous treatment, using the same data: two synthetic datasets, Hirano and Imbens (2004) and Sun et al. (2015); and one semi-synthetic dataset, Twins, containing weights, mortality and other covariates of all 71,345 pairs of twins born in the USA between 1989–1991. For the two synthetic datasets, we train the models on 1000 samples. The detailed information can be found in the supplementary appendix of Liu et al. (2024). We present the key information in Table 5.3.

Figure 5.4 compares ADRF estimates from frengression and CausalEGM on the three benchmark datasets over $K = 30$ replications. We train both models for 1000 epochs in all experiments. For CausalEGM specifically, we use the best model configuration provided by their package.

For each replication k and treatment level x, we compute the average of 1000 draws from the learned frengression model as the estimated treatment effect on level x. We then aggregate

	Twins	**Hirano & Imbens**	**Sun**
d_z	50	200	200
Z	Real data derived from all births in the USA between 1989–1991.	$(Z_1, \ldots, Z_{d_z}) \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$	$(Z_1, \ldots, Z_{d_z}) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
$Y(x) \mid Z$	$Y(x) = -\frac{2}{1+e^{-3x}} + z^\top\gamma + \epsilon$, $\gamma = (\gamma_1, \ldots, \gamma_{d_z})$, $\gamma_i \sim \mathcal{N}(0, 0.025^2)$, $\epsilon \sim \mathcal{N}(0, 0.25^2)$	$\mathcal{N}(x + (Z_0 + Z_2)e^{-x(Z_0+Z_2)}, 1)$	$\mathcal{N}(x + Z_1 + \cos(Z_2) + Z_5^2 + Z_6 - \frac{1}{2}, 1)$
$X \mid Z$	Observed birth weights (taken from the real data, not generated)	$\exp(Z_0 + Z_1)$	$\mathcal{N}(-2\sin(2Z_1) + Z_2^2 + Z_3 + \cos(Z_4) - \frac{5}{6}, 1)$
$\mu(x)$	$-\frac{2}{1+e^{-3x}} + \mathbb{E}[Z^\top\gamma]$	$x + \frac{2}{(1+x)^3}$	$x + 0.5 + e^{-0.5}$

Table 5.3: Key information of the three datasets in continuous treatments, including the dimension of covariates d_z, the data generation process of Z, X, Y and the ground truth ADRF $\mu(x)$.

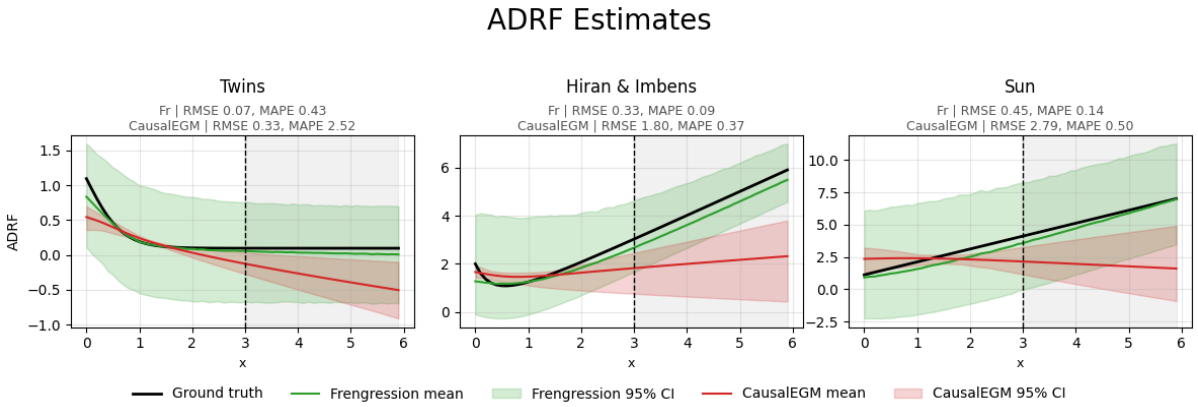


Figure 5.4: ADRF estimation, with RMSE and MAPE, frengression (Fr) and CausalEGM, across 30 replications. We sample 1000 datapoints for training for each dataset. The grey area is not seen during training. We train both models for 1000 epochs in all simulations.

across replications to obtain the ensemble estimate $\hat{\mu}(x) = \frac{1}{K}\sum_{k=1}^{K}\hat{\mu}^k(x)$, plotted as a solid green line for frengression and a red line for CausalEGM.

To illustrate uncertainty, we add shaded bands corresponding to the 2.5%–97.5% interval. For the kth replication of frengression, let $\hat{q}_\alpha^k(x)$ be the empirical α-quantiles of the 1000 samples from $\hat{P}^k_{Y(x)}$. The band at x is then

$$\left[\frac{1}{K}\sum_{k=1}^{K}\hat{q}^k_{2.5\%}(x), \frac{1}{K}\sum_{k=1}^{K}\hat{q}^k_{97.5\%}(x)\right].$$

Since CausalEGM produces point estimates only, we take the 2.5% and 97.5% sample quantiles of $\{\hat{\mu}^k(x)\}_{k=1}^K$ to form its shaded band.

We assess each model's ability to recover the dose-response function on a treatment range

unseen during training by withholding treatments in the interval $(3, 6]$ from the fitting process. After training, we estimate the ADRF over the full treatment range, including the held-out region $x \in (3, 6]$, and compare the estimated ADRF with the ground truth. As shown in Figure 5.4, frengression consistently outperforms CausalEGM across all three benchmark datasets, showing superior accuracy both within the training range and in the extrapolation interval, and does so without any additional hyperparameter tuning. The estimated 2.5% and 97.5% quantile bands accurately bound the true ADRF across all treatment levels.

5.6.3 Longitudinal setting and survival data

We demonstrate the modelling on longitudinal data. To the best of our knowledge, only a few deep learning frameworks have been proposed regarding inference and simulation on marginal causal effects. The recent work of Shirakawa et al. (2024), which combines the Transformer architecture and LTMLE, handles only estimation of causal effects on binary treatment in the longitudinal context.

We design four synthetic settings to showcase the flexibility of frengression. Settings 1–3 are designed for survival analysis, and setting 4 is longitudinal. In settings 1–3, we denote $\widetilde{Y}_t$ as the residual lifetime after surviving to time t. If $\widetilde{Y}_t < 1$, then $Y_t = 1$, otherwise $Y_t = 0$. The details of each setting are given below in Appendix C.4.2.1. We provide the summary of key information as below:

- Setting 1: $T = 10$. C is drawn from a Bernoulli distribution. We set the treatment X_t to be binary and dependent on (Z_t, C); the time-varying covariate Z_t is dependent on previous treatment X_{t-1} and C; the outcome $\widetilde{Y}(\overline{x}_t) \,|\, C$ follows an exponential distribution.
- Setting 2: $T = 10$. C is drawn from an exponential distribution with rate 1 and Z_0 from a standard normal; $X_0 \,|\, Z_0 \sim \text{Bernoulli}\left(\text{expit}(-0.5 + 0.5Z_0)\right)$; the treatment X_t is still binary, and is dependent on both Z_t and the previous treatment X_{t-1}; the time-varying covariate Z_t depends on X_{t-1} and Z_{t-1}. The correlation between Y_t and Z_{t-1} is 0.2, and that for Y_t and Z_t given Z_{t-1} is 0.3. Compared to setting 1, setting 2 has more complicated dependency.
- Setting 3: $T = 10$. C is drawn from a Bernoulli distribution. We set X_t to be continuous from a normal distribution, dependent on Z_t and C. We use a Gaussian copula with correlation 0.4 between Z_t and Y_t. In this setting the treatment is continuous.

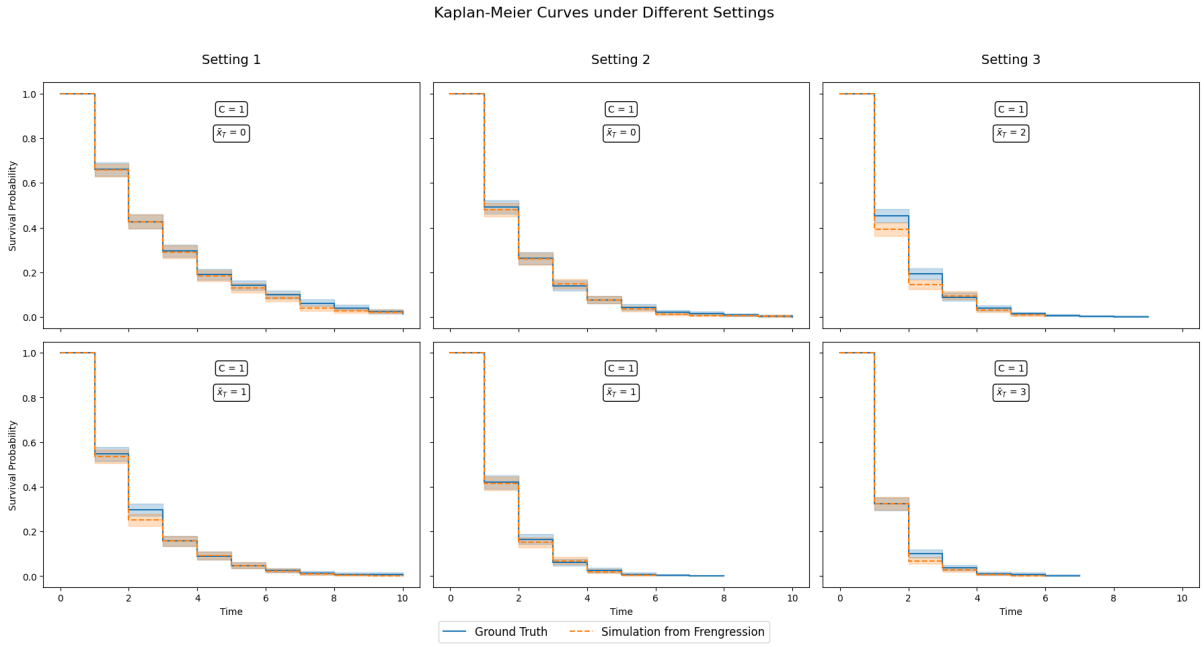


Figure 5.5: Kaplan-Meier curves of data generated from true data generating process and simulated from fitted frengression.

- Setting 4: $T = 5$. C is drawn from a standard normal distribution. We set X_t to be continuous from normal distribution, and the outcome $Y_t(\overline{x}_t) \,|\, C \sim \mathcal{N}(2x_t + x_{t-1} + 0.5x_{t-2} + C, 1)$. This models the longitudinal scenario: at time step t we are able to observe a continuous outcome rather than just the indicator of event occurrence.

In each setting, we simulate $n = 6000$ samples for training. We then sample 1000 datapoints from each fitted model. In Figure 5.5, we compare Kaplan-Meier curves fitted on the true data-generating process with those simulated by frengression with intervention x_t fixed the same across time. Note that frengression is trained on time-varying covariates and interventions as described above. This illustrates the flexibility of frengression in the time-varying setting.

To create a performance baseline, we also fit Deep LTMLE using the `dltmle` package (Shirakawa, 2024) to setting 1, with the horizon set to $T = 5$. The causal estimand of interest here is the risk difference between always-treated and always-placebo regimens over the entire follow-up horizon $\theta = \mathbb{E}\, Y_{T-1}(\mathbf{1}) - \mathbb{E}\, Y_{T-1}(\mathbf{0})$, where, as before, $Y_t = 1$ if the event has occurred by time t and 0 otherwise. The true value of θ is then 0.11. Over 10 replications, Deep LTMLE's mean estimate is 0.13 (SD 0.03), and frengression's is 0.09 (SD 0.04). These results are similar, and frengression achieves them without a separate targeting step or an explicit influence-function calculation, thus it can be applied to a broader class of causal estimands. We would also like to highlight the computational efficiency of frengression compared to `dltmle`. Hyperparameter

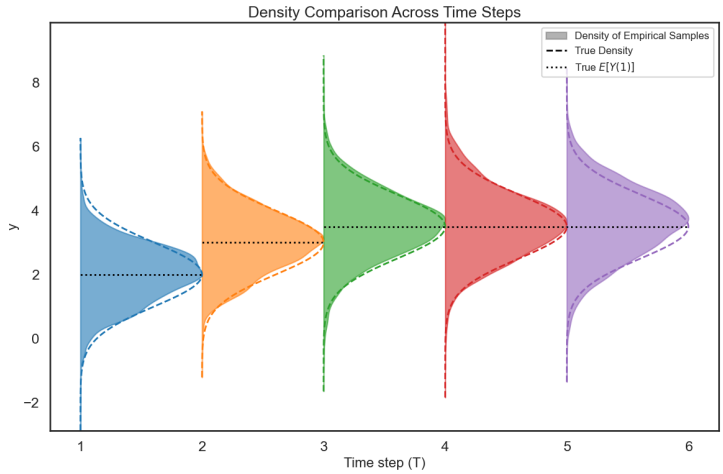


Figure 5.6: Density comparison of data simulated from frengression vs. true density, Setting 4.

tuning with `dltmle`'s default `Optuna` routine takes about 30 minutes for 10 trials, plus 4 minutes to run 100 training epochs, whereas frengression completes 2000 epochs in 1.4 minutes on the same hardware.

Figure 5.6 shows the results for Setting 4, where we plot the density of 5000 points simulated from the interventional distribution $\hat{P}_{Y(x)|C}$ estimated by frengression and set $x = \mathbf{1}$, $C = 0$. As expected, the frengression fit closely matches the true interventional distribution, although the performance worsens slightly at the last time step due to growing complexity.

5.6.4 Distributional regression

In practical applications, it is often necessary to simulate data under a pre-specified interventional distribution while preserving the empirical dependencies observed in the original sample. As a distributional-regression framework, frengression can estimate the target marginal interventional distribution of the observed data; it also allows one to generate synthetic datasets that retain the observed correlation structure between the covariates and treatments, but target an arbitrarily specified marginal causal distribution.

In Figure 5.7, we illustrate our method in the RCT (no confounding variables) context by comparing the empirical histogram from frengression's simulations with the ground truth, and showcase its ability to simulate data with user-specified marginal causal distributions while preserving the joint dependency structure of the covariates and treatment.

First, we generate $n = 5000$ samples from the true data-generating process: the six covariates $Z \in \mathbb{R}^6$ are independently drawn from $\mathcal{N}(0, 1)$, the treatment $X \sim \text{Bernoulli}(0.5)$ as commonly

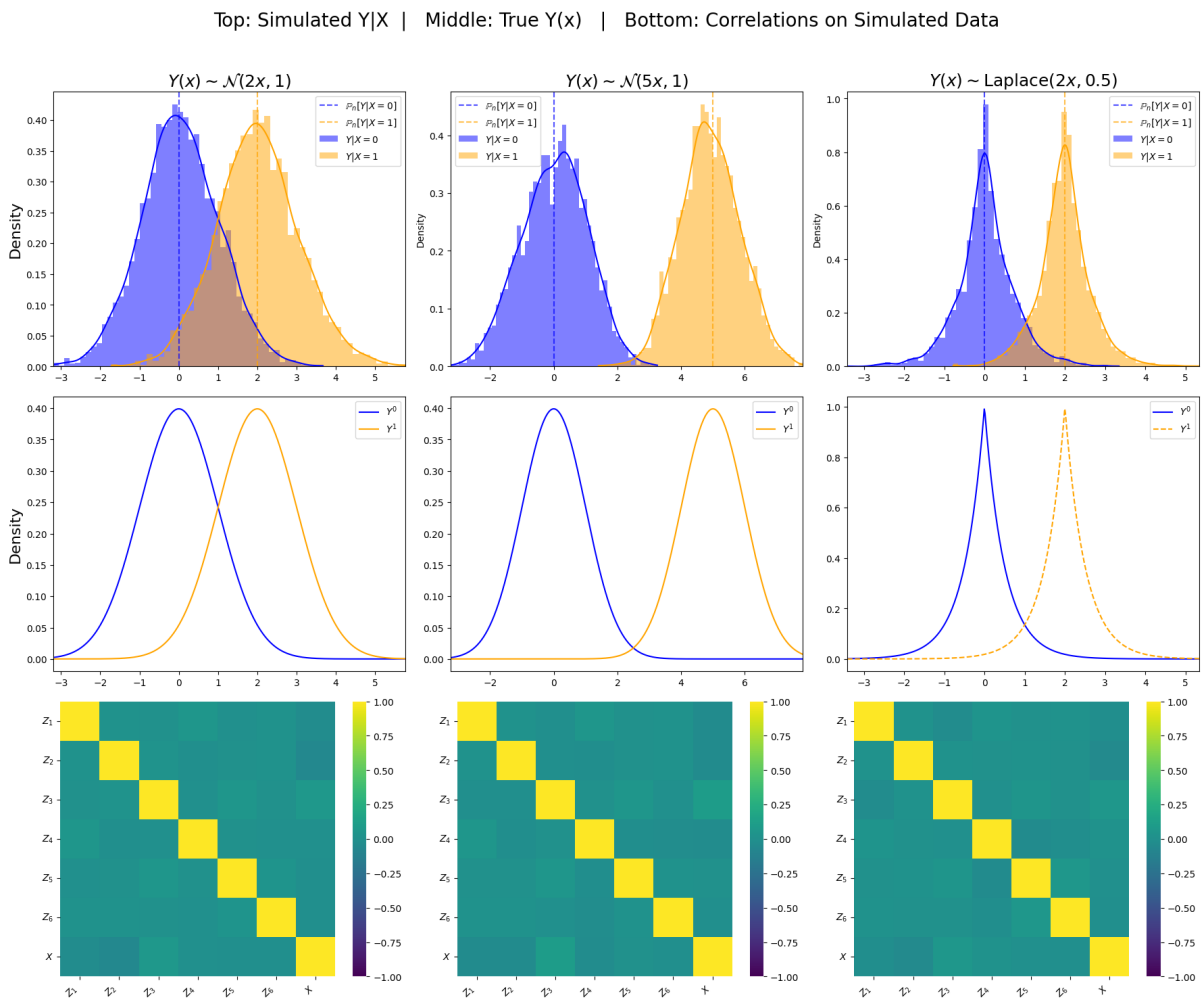


Figure 5.7: Marginal densities of simulations from frengression (top row), target (middle row) and correlation of covariates and treatment on simulations from frengression (bottom row).

seen in an RCT, and the intervened outcome $Y(x) \sim N(2x, 1)$, with a Gaussian copula between (Z_5, Z_6) and Y with correlation $\rho = 2\,\text{expit}(1) - 1 \approx 0.46$. Fitting frengression to these samples recovers the true interventional distribution and the correlation between X and Z (left column).

Next, we replace the fitted $Y(x)$ marginal with two new targets: $\mathcal{N}(5x, 1)$ and Laplace$(2x, 0.5)$, and draw $n = 5000$ simulated triples $(\widehat{Z}, \widehat{X}, \widehat{Y})$ from the modified model (middle and right columns). In Figure 5.7, the top row shows the estimated marginal densities of $Y \,|\, X = 0$ and $Y \,|\, X = 1$ from the simulated data. Because there is no confounding, these curves should match the true interventional marginals. The middle row displays the true target densities (the Gaussians or Laplace) for each intervention; we see that the simulations from the frengression fits align with these densities. The bottom row confirms that the original joint distribution of (Z, X) remains unchanged under the modified interventional model. The results validate that in all cases, the simulated marginals match the newly specified objectives while preserving the

baseline dependencies between X and Z.

5.7 Application on LEADER

We illustrate frengression on the LEADER trial (Novo Nordisk A/S, 2015) to showcase its ability to simulate realistic, time-to-event data with complex covariate patterns. LEADER is a randomised, double-blind, placebo-controlled cardiovascular outcomes trial that enrolled 9340 patients with type 2 diabetes at high cardiovascular risk, randomising patients into liraglutide (up to 1.8 mg daily) or placebo, both on top of standard care, resulting in similarly sized treatment arms. Patients were recruited from August 2010 through April 2012 and followed for 3.5–5 years, with a median follow-up of 3.8 years. The primary endpoint is time to first major adverse cardiovascular event (MACE), including cardiovascular death, non-fatal myocardial infarction and non-fatal stroke. The original analysis found that MACE occurred in significantly fewer patients in the liraglutide group than in the placebo group (Marso et al., 2016).

To build our generative model, we select baseline covariates known to predict MACE. The baseline covariates include four binary indicators: gender, smoker status, carotid stenosis $> 50\%$ on angiography, and diabetic nephropathy at screening. We have seven continuous variables at baseline: age, duration of diabetes (years), high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, total cholesterol, triglycerides, and serum creatinine. We also incorporate three key time-varying biomarkers to capture longitudinal risk dynamics: HbA1c, body-mass index (BMI), and estimated glomerular filtration rate (eGFR). The description of the variables can be found in Table C.6.

We partition the trial timeline into 6-month intervals over a maximum of 60 months, yielding 11 discrete timepoints (i.e. $t = 0, 1, 2, \dots 10$). We define each subject's event time as $k \in (0, \infty)$, where $k \in (0, 10]$ indicates the time MACE occurred in units of 6-month intervals, and $k > 10$ indicates no event within the trial. With this notation, we define the event indicators Y_t, $t \in \{0, 1, \dots 10\}$ as

$$Y_t = \begin{cases} 1, & k \leq t, \\ 0, & \text{otherwise.} \end{cases}$$

Here, we have $Y_t = 1$ if and only if the event occurred by time t, otherwise $Y_t = 0$.

Note that in the time-varying setting, covariates were recorded at different intervals: HbA1c was measured every 6 months, whereas BMI and eGFR were collected only at months 6, 12,

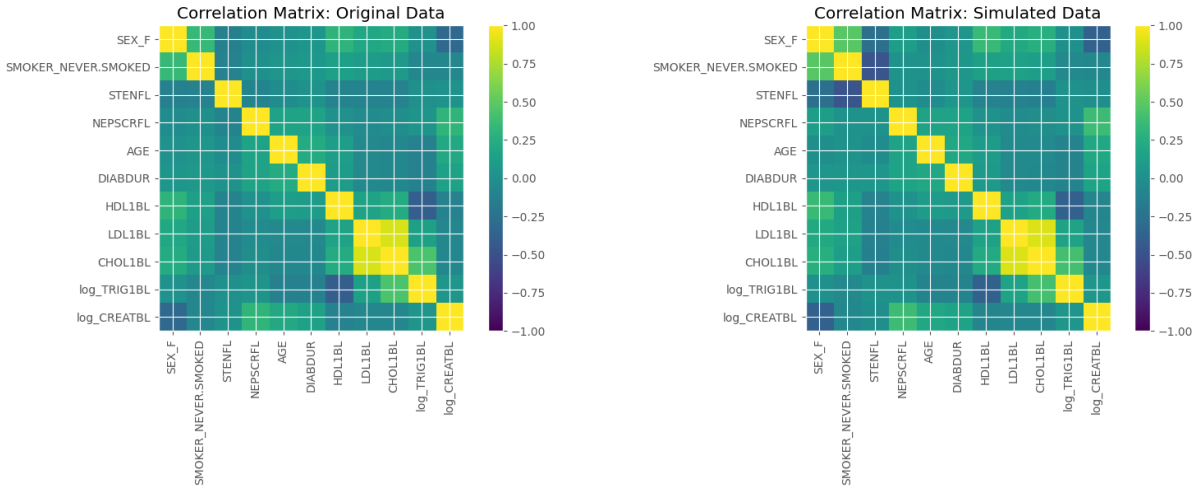


Figure 5.8: Correlation matrix of baseline covariates, real data vs. data simulated from fitted frengression. The heatmap illustrates the similarity of simulated data with real data regarding correlation coefficient.

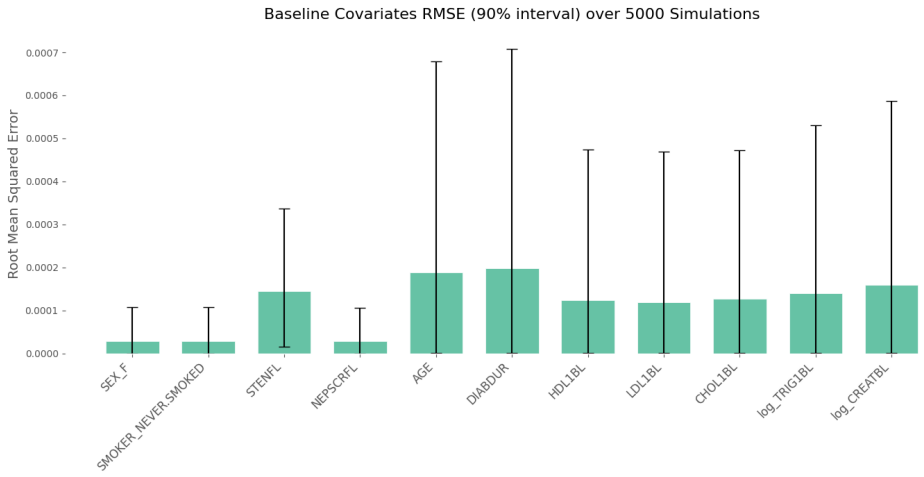


Figure 5.9: RMSE of each baseline covariate (continuous variables are normalised), with the maximum RMSE less than 0.001 over 5000 draws.

18, 30, 48, and 60 from the start of the trial. To retain as much information as possible, we impute missing BMI and eGFR values at intermediate timepoints using the most recent available measurement (last observation carried forward).

We exclude units missing any baseline covariates, leaving 9161 subjects. We take the logarithm of triglycerides and serum creatinine. We then normalise all continuous covariates by subtracting their mean values and dividing them by their standard deviations. After fitting frengression to the processed dataset, we simulate 9161 synthetic records from the model 5000 times.

Figures 5.8–5.10 visualise frengression's empirical performance on data simulation. For

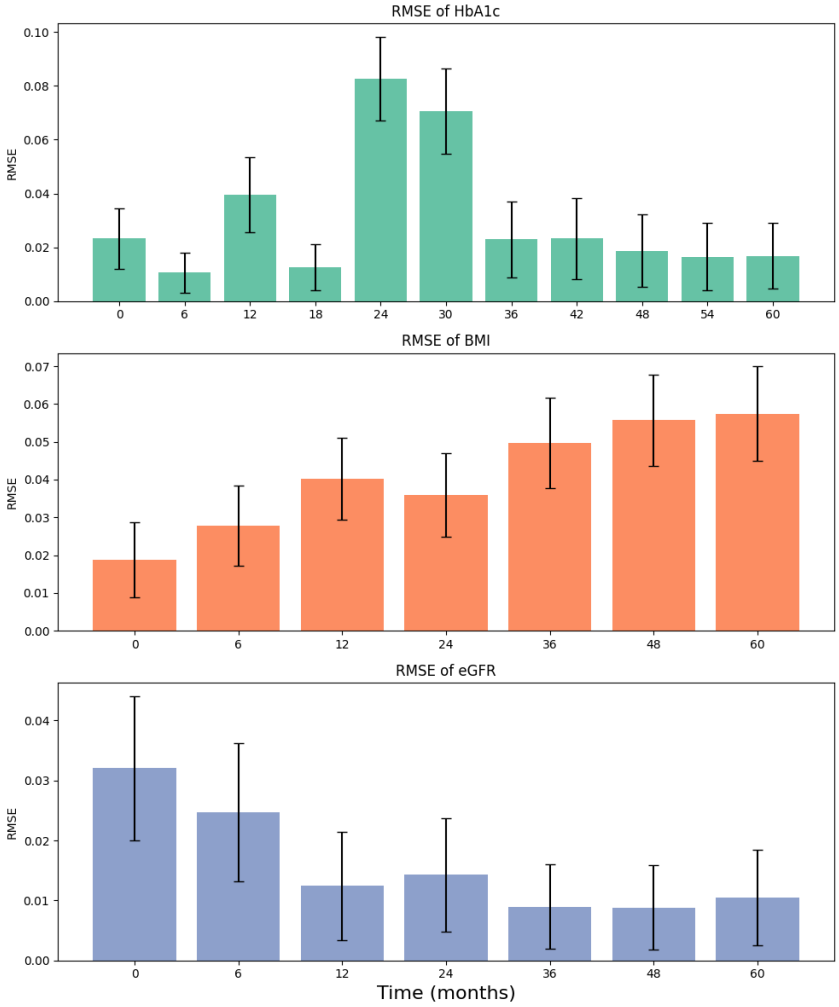


Figure 5.10: RMSE of each time varying covariate (normalised) across different time-points over 5000 draws. All of them are of small scales (maximum 0.1).

each variable at each time point, we compute the RMSE between the mean of the simulated samples and the true mean. Overall, frengression accurately reproduces the linear correlations and achieves low RMSE across both baseline and time-varying covariates. However, as shown in Figure 5.8, the simulated correlations of the STENFL variable (carotid $>$ 50% stenosis on angiography) with SEX_F (gender) and SMOKER_NEVER.SMOKED (indicating smoking status) deviate from those of the true data, reflecting the greater challenge of modelling binary variables using deep generative models compared to continuous ones.

To quantify how well the simulated data match the originals, we train logistic classifiers to distinguish real observations from simulated ones. On baseline covariates, the mean AUC is 0.50 (range 0.47–0.52), indicating classification that is nearly random. For the joint datasets (including baseline covariates, time-varying covariates and MACE occurrence), the mean AUC is 0.55

(range 0.52–0.57), still reflecting substantial similarity between simulated and real distributions.

We also sample event occurrences from the fitted interventional distribution $\hat{P}_{Y(x)|C}$. Over 5000 draws, the mean MACE occurrence rate in the placebo arm is 14.7% with 90% interval $[13.8\%, 15.6\%]$ (5th and 95th percentiles of the 5000 estimates), and in the liraglutide arm it is 13.0% with 90% interval $[12.2\%, 13.9\%]$. These values match, to the reported precision, the 14.7% in the placebo arm and 13.0% in the liraglutide arm from real data, confirming that frengression faithfully captures both covariate structures and event frequencies.

We apply DLTMLE to the LEADER data for comparison; implementation details and hyperparameters are given in Appendix C.4.3. DLTMLE yields an estimated event rate of 15.4% (95% CI: $[14.0\%, 16.5\%]$) under placebo and 12.9% (95% CI: $[11.9\%, 14.0\%]$) under treatment. These deviate more from the observed trial rates than the corresponding frengression estimates. As DLTMLE is designed for estimation rather than generation, it cannot be evaluated in terms of simulation fidelity (e.g., RMSE or correlation structure in Figures 5.8–5.10).

Estimation of $\mathbb{E}[Y(a)]$ and faithful data generation are complementary objectives. DLTMLE addresses the former. Frengression, in contrast, provides a unified approach: it produces consistent estimation while simultaneously learning a generative model capable of producing synthetic data.

5.8 Discussion

The rapid growth of AI makes reliable causal inference more important than ever. Contributing to this, we introduce a deep generative approach that generates synthetic data with a specified marginal interventional distribution, while maintaining a realistic level of complexity. Such a data generation tool is helpful in different applications, including pre-experiment study design for randomised controlled trials, selecting an inference procedure, and stress testing AI methods (Giuffrè and Shung, 2023; Pezoulas et al., 2024).

There are several promising extensions. As noted in Remark 5.2.3, the variation independent decomposition in frengression can be directly adapted for differentially private simulation. Although differential privacy remains the gold standard for protecting individual records, its strict requirements often come at the expense of simulation accuracy in practice. This means that suitable measures require a better understanding of the needs of end users (Yoon et al., 2020; Kenny et al., 2021; McCartan et al., 2023).

Methodologically, although frengression can generate data of trajectories at the subject level under time-varying treatment, or any quantity conditional on a subset of covariates, its robustness and efficiency for individual treatment effects, such as those needed for dynamic treatment regimes, remain open questions. Even so, we find that centring on the target causal distribution explicitly brings clarity and consistency in both estimation and simulation.

At present, frengression's extrapolation works only with continuous treatments; how to extrapolate when treatments are discrete is to be explored. In addition, we provided consistency results for the generative model. Deriving convergence rates would help us build valid confidence intervals for the estimates that frengression produces.

Chapter 6

Stable prediction via causal invariance

In the preceding chapters, stability is secured by the construction of the statistical procedure. This chapter takes up the other direction between causality and stability, set out in Section 2.3, where predictive performance instead inherits its stability from an underlying, invariant causal mechanism: when the relevant predictive relationship is invariant across environments, a prediction function built on it performs stably as the environment shifts. Most recent work on this route relies on observed heterogeneity. We instead ask what can be recovered from a single observational dataset. This chapter presents an exploratory framework, developed jointly with Professor Peter Bühlmann, that remains under development.

6.1 Background

Prediction models are commonly trained on data from one distribution and deployed under another. A prediction model trained with empirical risk minimisation (ERM) is optimised for the training distribution and carries no guarantee when the test distribution differs. The stability sought in this chapter is of this kind: predictive performance that is maintained under such shifts. Causal structure provides one route to it (Section 2.3): when the perturbations leave the structural assignment of Y unchanged, the conditional distribution of Y given its direct causes is the same in every environment. Among prediction functions of those causes, the one built on this invariant conditional minimises the risk in every environment simultaneously, and hence the worst-case risk over the perturbation class. Through invariance, causal structure yields stable prediction.

Methods for invariant prediction differ in the form of invariance they impose and in how they use variation across environments to identify causal structure. Invariant causal prediction (Peters et al., 2016) treats invariance as a testable hypothesis: the parents of Y in a structural equation model are a covariate subset S for which the conditional distribution $P(Y \mid X_S)$ is

the same across observed environments. The procedure scans subsets, tests invariance in each, and returns the intersection of accepted subsets as a confidence set for the causal parents, a structural object rather than a prediction function. It requires environments rich enough to distinguish the causal subset from spurious alternatives, and tests powerful enough to do so across them.

A related line of work pursues robustness more explicitly, founded on the same invariance. Rather than test invariance and return a covariate subset, it minimises the worst-case risk over a class of perturbations, in the sense of Section 2.3; a robustness parameter sets the size of this class, with the invariant prediction function recovered in the most robust limit. Meinshausen (2018) formulate the causal parameter in this distributionally robust sense. Anchor regression (Rothenhäusler et al., 2021) takes the robustness class to be shifts in the span of an exogenous anchor, yielding a continuum between ordinary least squares and a prediction function stable under such shifts. Distributional Robustness via Invariant Gradients (DRIG; Shen et al., 2025) extends this perspective from a single anchor to intervention classes indexed by observed environments. The methods differ in the form of heterogeneity they require (anchors or environment labels) but all require extra information beyond a single observational distribution.

Since anchor regression and DRIG serve as benchmarks in Section 6.8, their details are reviewed here. In each, a parameter γ in the training objective trades model fit on the observed distribution against robustness to a class of shifts; we present both with a linear prediction function $X^\top\beta$.

Anchor regression assumes a linear structural equation model $Z = \mathrm{B}Z + \mathrm{M}A + \varepsilon$, where Z collects the observed and hidden variables and the anchor A acts exogenously through $\mathrm{M}A$. Let Γ_A denote the L_2 projection onto $\mathrm{span}(A)$ and Id the identity, so that $\mathrm{Id} - \Gamma_A$ projects onto the orthogonal complement of $\mathrm{span}(A)$. Anchor regression returns $\hat{\beta}_\gamma = \mathrm{argmin}_\beta \mathcal{L}_{\mathrm{anchor},\gamma}(\beta)$, where

$$\mathcal{L}_{\mathrm{anchor},\gamma}(\beta) = \mathbb{E}\big[\big((\mathrm{Id} - \Gamma_A)(Y - X^\top\beta)\big)^2\big] + \gamma\mathbb{E}\big[\big(\Gamma_A(Y - X^\top\beta)\big)^2\big].$$

The objective weights the residual's component in $\mathrm{span}(A)$ by γ; as $\gamma \to \infty$ the penalty drives the residual orthogonal to the anchor, recovering the prediction function that is invariant to anchor-induced shifts. A shift replaces the anchor term $\mathrm{M}A$ by a vector δ in the column space of M, giving the perturbed system $Z = \mathrm{B}Z + \delta + \varepsilon$. The minimiser $\hat{\beta}_\gamma$ minimises the worst-case prediction risk over such shifts whose second moment satisfies $\mathbb{E}[\delta\delta^\top] \preceq \gamma\,\mathbb{E}\big[(\mathrm{M}A)(\mathrm{M}A)^\top\big]$, so

larger γ protects against larger shifts.

DRIG works with several external environments rather than an anchor variable. Their distributions $\{P_e\}_{e\in\mathcal{E}}$, one observational and the rest interventional, share a linear structural equation model $Z = \mathrm{B}Z + \varepsilon_e$ on $Z = (X, Y)$, with B invariant across environments and only the noise ε_e changing under intervention. Let $e_0 \in \mathcal{E}$ be the observational environment and let weights $\omega_e \geq 0$ satisfy $\sum_{e\in\mathcal{E}} \omega_e = 1$. DRIG returns $\hat{\beta}_\gamma = \operatorname{argmin}_\beta \mathcal{L}_{\mathrm{DRIG},\gamma}(\beta)$, where

$$\mathcal{L}_{\mathrm{DRIG},\gamma}(\beta) = \mathbb{E}_{P_{e_0}}\left[(Y - X^\top\beta)^2\right] + \gamma \sum_{e\in\mathcal{E}} \omega_e \Big(\mathbb{E}_{P_e}\left[(Y - X^\top\beta)^2\right] - \mathbb{E}_{P_{e_0}}\left[(Y - X^\top\beta)^2\right]\Big).$$

The objective is the reference risk on P_{e_0} plus γ times the environments' average excess risk relative to it. An intervention replaces the noise by a vector δ, giving the perturbed system $Z = \mathrm{B}Z + \delta$; environment e is recovered at $\delta = \varepsilon_e$, with second moment $S_e = \mathbb{E}_{P_e}[\varepsilon_e(\varepsilon_e)^\top]$. The minimiser $\hat{\beta}_\gamma$ minimises the worst-case MSE over interventions whose second moment satisfies $\mathbb{E}[\delta\delta^\top] \preceq S_{e_0} + \gamma \sum_{e\in\mathcal{E}} \omega_e (S_e - S_{e_0})$, so larger γ widens this class along the directions $S_e - S_{e_0}$, beyond the observed environments. The same objective applies with a general prediction function $g(X)$ in place of $X^\top\beta$; as our experiments are linear, we keep the linear form and refer to Shen et al. (2025) for the general case.

A different perspective is to seek a representation rather than a subset of the covariates. The idea is to learn a representation such that the optimal prediction function built on it is invariant across environments, and to predict from that representation. Invariant risk minimisation (Arjovsky et al., 2019) is the canonical example. The resulting invariance is only as informative as the observed environments and the representation class permit.

Each of the methods above depends on observed heterogeneity, supplied as environments or anchors, to expose the invariance. In this chapter, we consider a different setting: what robustness can be recovered when only a single observational distribution is available? We assume an additive noise model for Y; its generating function g_0 then has constant conditional MSE along its own prediction, and this is the invariance we work with. The environments needed to detect it are not observed but generated, by reweighting the observational distribution itself. The violations across these environments collect into a single Cramér–von Mises (CvM) penalty, and the penalised risk admits an exact distributionally robust interpretation, as anchor regression and DRIG do, with the perturbation class generated internally rather than supplied externally. The remainder develops this construction and its identification, with the open ends gathered in Section 6.9.

6.2 Invariance and the causal prediction function

Let $X \in \mathbb{R}^d$ collect the predictors[14] and $Y \in \mathbb{R}$ be the response with $\mathbb{E}[Y^2] < \infty$. For a measurable $g : \mathbb{R}^d \to \mathbb{R}$, write

$$R_g \coloneqq (Y - g(X))^2, \qquad F_g \coloneqq g(X), \qquad \sigma_g^2(t) \coloneqq \mathbb{E}[R_g \mid F_g = t], \qquad \bar{\sigma}_g^2 \coloneqq \mathbb{E}[R_g],$$

for the squared residual, the prediction $F_g = g(X)$, its conditional MSE at level t, and the marginal MSE $\bar{\sigma}_g^2$, which we call the *risk* of g; let $H_g(s) = \Pr(g(X) \leq s)$ be the distribution function of the prediction and P_{F_g} its distribution. Throughout the chapter, we consider prediction functions in a class $\mathcal{G}$ defined as follows.

Definition 6.2.1 (Prediction function class)**.** *$\mathcal{G}$ is a class of measurable functions $g : \mathbb{R}^d \to \mathbb{R}$ such that, for every $g \in \mathcal{G}$, $g(X) \in L_2(P_X)$, $\bar{\sigma}_g^2 > 0$, and H_g is continuous.*

We posit the following causal mechanism for Y.

Assumption 6.2.2 (Additive noise model)**.** *We assume $Y \leftarrow g_0(X) + \varepsilon$, where $g_0 \in \mathcal{G}$, $\varepsilon \perp\!\!\!\perp g_0(X)$, $\mathbb{E}[\varepsilon] = 0$, and $\mathrm{Var}(\varepsilon) = \sigma^2 > 0$.*

The independence is imposed against the prediction $g_0(X)$, not against the covariate vector X, and is strictly weaker than $\varepsilon \perp\!\!\!\perp X$: under hidden confounding affecting both X and Y, ε may correlate with the confounded components of X while remaining independent of $g_0(X)$ for an appropriate g_0. Reweighting through $g_0(X)$ preserves $P(Y \mid g_0(X))$ (Proposition 6.3.1). The same reweighting preserves $P(Y \mid X)$ as well, even with the confounded components of X. Such confounding can, however, prevent the identification of g_0, as discussed after Assumption 6.6.1.

Under Assumption 6.2.2, $\mathbb{E}[Y \mid g_0(X)] = g_0(X)$, and the conditional MSE $\sigma_{g_0}^2(t) = \mathbb{E}[(Y - g_0(X))^2 \mid g_0(X) = t] = \mathrm{Var}(\varepsilon) = \sigma^2$ for $P_{F_{g_0}}$-almost every t: the residual of the causal prediction function g_0 is the noise ε, independent of its own prediction, so its conditional MSE is constant. This property involves only the observed distribution, and in this chapter we take it as our definition of an invariant prediction function.

Definition 6.2.3 (Invariant prediction functions)**.** *The class of invariant prediction functions is*

$$\mathcal{G}_I \coloneqq \{g \in \mathcal{G} : \sigma_g^2(t) \equiv c_g \text{ for } P_{F_g}\text{-almost every } t, \text{ for some constant } c_g\}.$$

[14] Within this chapter X denotes the covariates, rather than the treatment in Chapters 3 and 5.

Since $\sigma^2_{g_0}(t)$ is constant, $g_0 \in \mathcal{G}_I$. Section 6.3 justifies the name: g lies in $\mathcal{G}_I$ exactly when its prediction risk is invariant across a family of environments generated from the observational distribution by reweighting.

Constancy of σ^2_g is a condition on the second moment of the residual. Note that it is necessary but not sufficient for the residual's conditional distribution to be constant along the prediction. It also does not by itself determine g_0: a prediction function other than g_0 can also have a residual whose squared error is constant along its own fitted values. Identification therefore requires a further assumption, given in Section 6.6.

6.3 Synthetic environments via reweighting

Definition 6.2.3 defines $\mathcal{G}_I$ from a single observational distribution P, requiring no external environments. Yet the name *invariant* points to behaviour across environments; we justify it by generating such environments from P through reweighting.

Assume P admits a density p, and reweight through the covariates: $p_w(x, y) = w(x)\, p(x, y)$, where $w \geq 0$ is a function of x and not of y, normalised so that $\mathbb{E}[w(X)] = 1$, and P_w is the corresponding distribution. We begin with the following proposition.

Proposition 6.3.1. *Suppose the weight factors as $w(x) = \tilde{w}(s(x))$ through a summary $V = s(X)$ of the covariates. Then reweighting leaves the conditional distribution of Y invariant given V and given the full covariate vector X:*

$$P_w(Y \mid V) = P(Y \mid V) \qquad \text{and} \qquad P_w(Y \mid X) = P(Y \mid X).$$

Proposition 6.3.1 shows that a weight built from a coarse summary V of the covariates still preserves the full conditional $P(Y \mid X)$: reweighting therefore yields synthetic environments of covariate shift, altering only the covariate marginal. The same invariance holds for any statistic of which V is a function; a weight depending only on the covariates X_{S_w}, $S_w \subseteq \{1, \ldots, d\}$, preserves $P(Y \mid X_S)$ for every $S \supseteq S_w$. The proof is in Appendix D.1.

By Definition 6.2.3, g lies in $\mathcal{G}_I$ precisely when its conditional MSE $\sigma^2_g(t)$ is constant along the prediction. For any $g \in \mathcal{G}$, we reweight through its own prediction, taking $w(x) = \tilde{w}(g(x))$ with $\tilde{w} \geq 0$ and $\mathbb{E}[\tilde{w}(F_g)] = 1$. Since w is a function of $g(X)$, Proposition 6.3.1 with $V = g(X)$ gives $P_w(Y \mid g(X)) = P(Y \mid g(X))$, so $\sigma^2_g(t) = \mathbb{E}\big[(Y - t)^2 \mid F_g = t\big]$ is unchanged, while the

covariate distribution changes only through the marginal distribution of F_g, tilted by $\tilde{w}$. The risk under P_w therefore averages the unchanged $\sigma_g^2(t)$ against the new marginal:

$$\mathbb{E}_{P_w}[R_g] = \mathbb{E}\big[\tilde{w}(F_g)\, R_g\big] = \mathbb{E}\big[\tilde{w}(F_g)\, \sigma_g^2(F_g)\big].$$

This equals $\bar{\sigma}_g^2$ for every such $\tilde{w}$ if and only if $\sigma_g^2(F_g)$ is almost surely constant: $g \in \mathcal{G}_I$ exactly when its prediction risk is invariant across all synthetic environments generated by reweighting its own predictions.

To check directly whether a given g lies in $\mathcal{G}_I$, one would compute its largest violation over normalised weights,

$$\sup_{\tilde{w}\geq 0,\, \mathbb{E}[\tilde{w}(F_g)]=1} \Big|\mathbb{E}[\tilde{w}(F_g)\, R_g] - \bar{\sigma}_g^2\Big|, \tag{6.1}$$

an adversarial search over a class of admissible weights. While this characterisation is exact, it depends on the choice of weight class, and requires optimisation over infinitely many admissible functions $\tilde{w}$.

A single one-parameter family of weights, however, already characterises $\mathcal{G}_I$, so the adversarial search in (6.1) is unnecessary. For thresholds s with $\Pr(F_g \leq s) > 0$, consider the indicator weights on the prediction scale,

$$\tilde{w}_s(t) = \frac{\mathbb{1}\{t \leq s\}}{\Pr(F_g \leq s)}.$$

Recall that $g \in \mathcal{G}_I$ exactly when its reweighted risk equals $\bar{\sigma}_g^2$ under every normalised weight. Under $\tilde{w}_s$ this risk is

$$\mathbb{E}\big[\tilde{w}_s(F_g)\, R_g\big] = \frac{\mathbb{E}\big[R_g\, \mathbb{1}\{F_g \leq s\}\big]}{\Pr(F_g \leq s)} = \mathbb{E}\big[R_g \,\big|\, g(X) \leq s\big],$$

the risk conditional on $g(X) \leq s$, and setting this conditional risk equal to $\bar{\sigma}_g^2$ at every s gives $\mathbb{E}[R_g\, \mathbb{1}\{F_g \leq s\}] = \bar{\sigma}_g^2 \Pr(F_g \leq s)$. The equality holds under every indicator weight $\tilde{w}_s$ exactly when $g \in \mathcal{G}_I$, so the adversarial search over all weights reduces to this one-parameter family. The proposition below states this, with proof in Appendix D.2.

Proposition 6.3.2 (Completeness)**.** *$g \in \mathcal{G}_I$ if and only if $\mathbb{E}\big[R_g \mathbb{1}\{F_g \leq s\}\big] = \bar{\sigma}_g^2 \Pr(F_g \leq s)$ for all $s \in \mathbb{R}$.*

The next section turns this invariance criterion into a single penalty.

6.4 Cramér–von Mises penalty

The criterion of Proposition 6.3.2 holds threshold by threshold. Squaring the violation at each s and integrating against the distribution of the prediction collects these conditions into a single non-negative quantity,

$$\mathcal{C}(g) \coloneqq \int \Big(\mathbb{E}[R_g \mathbb{1}\{F_g \le s\}] - \bar{\sigma}_g^2 \Pr(F_g \le s)\Big)^2 \mathrm{d}P_{F_g}(s). \tag{6.2}$$

The two terms inside the parentheses are each $\bar{\sigma}_g^2$ times a distribution function of the prediction. One is the ordinary $H_g(s) = \Pr(F_g \le s)$; the other is the residual-weighted $\widetilde{H}_g(s) \coloneqq \mathbb{E}[R_g \mathbb{1}\{F_g \le s\}]/\bar{\sigma}_g^2$, which reweights the distribution of the prediction by the squared residual. Thus

$$\mathcal{C}(g) = \bar{\sigma}_g^4 \int \big(\widetilde{H}_g - H_g\big)^2 \,\mathrm{d}P_{F_g}.$$

The integral here is the CvM distance between $\widetilde{H}_g$ and H_g, a classical measure of the discrepancy between two distribution functions, and $\mathcal{C}(g)$ is this distance scaled by $\bar{\sigma}_g^4$. The two distribution functions coincide exactly when σ_g^2 is constant, so $\mathcal{C}(g)$ measures how far g departs from $\mathcal{G}_I$.

Proposition 6.4.1 (CvM penalty and invariance)**.** *$\mathcal{C}(g) = 0$ if and only if $g \in \mathcal{G}_I$.*

The proof is given in Appendix D.3. Since $\mathcal{C}(g)$ is zero exactly on $\mathcal{G}_I$ and positive otherwise, using it as a penalty on the prediction risk yields a prediction function that fits the data while staying close to $\mathcal{G}_I$:

$$g_\lambda = \operatorname*{argmin}_{g \in \mathcal{G}} \mathbb{E}[(Y - g(X))^2] + \lambda \mathcal{C}(g), \qquad \lambda \ge 0, \tag{6.3}$$

which reduces to least-squares risk minimisation over $\mathcal{G}$ at $\lambda = 0$ and approaches optimisation over $\mathcal{G}_I$ in the limit $\lambda \to \infty$. The empirical version of $\mathcal{C}(g)$ on a sample $\{(X_i, Y_i)\}_{i=1}^n$ is

$$\hat{\mathcal{C}}_n(g) = \frac{1}{n}\sum_{j=1}^n \left(\frac{1}{n}\sum_{i=1}^n R_{g,i}\, \mathbb{1}\{F_i \le F_j\} - \bar{R}_g \cdot \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{F_i \le F_j\}\right)^2, \tag{6.4}$$

with $F_i = g(X_i)$, $R_{g,i} = (Y_i - g(X_i))^2$, $\bar{R}_g = n^{-1}\sum_{i=1}^n R_{g,i}$. The corresponding finite-sample estimator is

$$\hat{g}_{n,\lambda_n} = \operatorname*{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - g(X_i))^2 + \lambda_n \hat{\mathcal{C}}_n(g).$$

After sorting the fitted values, (6.4) can be evaluated in $O(n \log n)$ time. The indicator may be replaced by a smooth approximation when gradient-based optimisation is required.

6.5 Distributional-robustness interpretation

Recall from Section 6.1 that the anchor and DRIG penalties are not mere regularisation: at any finite γ, the penalised objective equals a worst-case prediction risk over a robustness class. The construction below is analogous: for every $\lambda > 0$, the penalised objective is a regularised worst-case risk over perturbations of the marginal distribution of $g(X)$.

We make this precise. The perturbation class is a family $\{Q_h\}$ of reweightings of the marginal of $g(X)$, indexed by the elements h of a Hilbert space; the penalised objective will be a regularised worst-case risk over this family. Let $U = H_g(F_g) \sim \text{Uniform}(0,1)$, and consider the Hilbert space

$$\mathcal{H} = \left\{ h : [0,1] \to \mathbb{R} \,\middle|\, h(1) = 0,\ h \text{ absolutely continuous},\ \int_0^1 (h'(u))^2 \mathrm{d}u < \infty \right\},$$

with inner product $\langle h_1, h_2 \rangle_{\mathcal{H}} = \int_0^1 h_1'(u) h_2'(u)\, \mathrm{d}u$. To each $h \in \mathcal{H}$, we associate a normalised perturbation Q_h[15] on $\mathbb{R}^d \times \mathbb{R}$ with density ratio

$$\frac{\mathrm{d}Q_h}{\mathrm{d}P}(x,y) = 1 + h(H_g(g(x))) - \int_0^1 h(u)\, \mathrm{d}u.$$

Since U is uniformly distributed on $[0,1]$, the density ratio integrates to one, and hence Q_h has total mass one. The density ratio is a function of $g(X)$, so Q_h preserves both $P(Y \mid X)$ and $P(Y \mid g(X))$ and acts only on the marginal distribution of the prediction, as in Section 6.3. When the density ratio is non-negative, Q_h is a probability measure and corresponds to a covariate shift. Otherwise, Q_h should be interpreted as a normalised signed perturbation.

The worst case over this family comes from a property of $\mathcal{C}(g)$: it is a squared norm in $\mathcal{H}$, that of the kernel mean embedding of the centred conditional MSE. Let $\psi_g(u) \coloneqq \sigma_g^2(H_g^{-1}(u)) - \bar{\sigma}_g^2$ denote this centred conditional MSE on the quantile scale, where H_g^{-1} is the quantile function

[15] Within this chapter Q_h denotes a perturbation of P indexed by $h \in \mathcal{H}$; it is unrelated to the historical-control distribution written Q_h in Chapter 4.

of the prediction; by iterated expectation, $\int_0^1 \psi_g(u)\,\mathrm{d}u = 0$.

Lemma 6.5.1 (The CvM penalty as a squared RKHS norm)**.** *The space $\mathcal{H}$ has reproducing kernel $k_0(v,w) = 1 - \max(v,w)$. The kernel mean embedding $\xi_g \coloneqq \int_0^1 \psi_g(v)\, k_0(v,\,\cdot\,)\,\mathrm{d}v$ belongs to $\mathcal{H}$ and satisfies $\langle h, \xi_g\rangle_{\mathcal{H}} = \int_0^1 \psi_g(u)\, h(u)\,\mathrm{d}u$ for every $h \in \mathcal{H}$, and*

$$\mathcal{C}(g) = \|\xi_g\|_{\mathcal{H}}^2 = \int_0^1 \Psi_g(u)^2\,\mathrm{d}u, \qquad \Psi_g(u) \coloneqq \int_0^u \psi_g(v)\,\mathrm{d}v.$$

The proof is given in Appendix D.4. Because $\mathcal{C}(g)$ is the squared RKHS norm $\|\xi_g\|_{\mathcal{H}}^2$, with $\mathcal{H}$ indexing the perturbations Q_h, the penalised objective coincides with a worst-case risk over that family, as shown below.

Theorem 6.5.2 (Distributional robustness equivalence)**.** *For every $\lambda > 0$ and $g \in \mathcal{G}$,*

$$\mathbb{E}_P\big[(Y - g(X))^2\big] + \lambda\,\mathcal{C}(g) = \sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_{Q_h}\big[(Y - g(X))^2\big] - \frac{\|h\|_{\mathcal{H}}^2}{4\lambda} \right\},$$

with the supremum attained at the unique maximiser $h^* = 2\lambda\,\xi_g$. Consequently, the penalised estimator g_λ of (6.3) *also minimises this worst-case risk:*

$$g_\lambda = \operatorname*{argmin}_{g \in \mathcal{G}} \sup_{h \in \mathcal{H}} \left\{ \mathbb{E}_{Q_h}\big[(Y - g(X))^2\big] - \frac{\|h\|_{\mathcal{H}}^2}{4\lambda} \right\}.$$

Theorem 6.5.2 shows that the penalised objective is the worst-case risk over the perturbations Q_h, traded off against a cost $\|h\|_{\mathcal{H}}^2/(4\lambda)$. These reshape the rank distribution of the fitted values while preserving $P(Y \mid X)$, and the cost keeps the worst case finite and sets its scale: larger λ allows larger perturbations and stronger robustness, as γ does in anchor regression and DRIG. As $\mathcal{C}(g)$ is zero exactly when the conditional MSE is flat, the causal g_0 incurs no penalty: no perturbation inflates its risk. The proof is in Appendix D.5.

6.6 Identification

As $\lambda \to \infty$, the estimator (6.3) approaches a minimiser of the risk $\mathbb{E}[(Y - g(X))^2]$ over $\mathcal{G}_I$, so identification rests on this minimiser being unique and equal to g_0. Neither is automatic: $\mathcal{G}_I$ can be non-convex, and a prediction function whose fitted values correlate with ε may attain a smaller risk than g_0. The orthogonality condition below removes both, leaving g_0 as the unique minimiser.

Assumption 6.6.1 (Orthogonality)**.** *For every $g \in \mathcal{G}_I$, $\mathbb{E}[g(X)\,\varepsilon] = 0$.*

Assumption 6.6.1 rules out prediction functions whose residuals remain homoscedastic but whose prediction still correlates with the structural noise. Hidden confounding can create one: a function that uses the confounded covariates may keep a homoscedastic residual, so it lies in $\mathcal{G}_I$, while its prediction and ε both depend on the hidden confounder and are correlated, so $\mathbb{E}[g(X)\varepsilon] \neq 0$. Then, without Assumption 6.6.1, g_0 need not minimise risk over $\mathcal{G}_I$.

Proposition 6.6.2 (Population identification)**.** *Under Assumptions 6.6.1 and 6.2.2, g_0 is the unique minimiser of $\mathbb{E}[(Y - g(X))^2]$ over $\mathcal{G}_I$, and any $L_2(P_X)$-limit point of g_λ as $\lambda \to \infty$ that lies in $\mathcal{G}$ coincides with g_0 in $L_2(P_X)$.*

For $g \in \mathcal{G}_I$, orthogonality cancels the cross term in $\mathbb{E}[(Y - g(X))^2]$, which then reduces to $\mathbb{E}[(g_0(X) - g(X))^2] + \sigma^2$ and is minimised only at g_0. The limiting statement requires, in addition, that $\mathcal{C}$ is continuous along $L_2(P_X)$-convergent sequences in $\mathcal{G}$, which we establish as Lemma D.6.1. The proof is given in Appendix D.6.

6.7 Using observed environments

So far the environments have been synthetic, generated from the single distribution P by reweighting. When several environments are instead observed directly, the same invariance criterion is imposed in each, and we look for prediction functions that are jointly invariant. Suppose K environments are observed, with distributions $\{P_e\}_{e=1}^K$, sample sizes n_e, and environment weights $\omega_e = n_e / \sum_e n_e$; throughout this section the conditions of Definition 6.2.1 are imposed under every P_e. We keep the same causal prediction function across environments and let only the residual distribution change.

Assumption 6.7.1 (Multi-environment structural model)**.** *We assume $Y^e \leftarrow g_0(X^e) + \varepsilon_e$, where $g_0 \in \mathcal{G}$, $\varepsilon_e \perp\!\!\!\perp g_0(X^e)$, and $0 < \mathbb{E}[\varepsilon_e^2] < \infty$ for every $e \in \{1, \dots, K\}$.*

As in Assumption 6.2.2, the independence is imposed against the prediction $g_0(X^e)$ rather than the covariates, so the residual distribution, and with it the confounding, may differ across environments. Write $\sigma^2_{g,e}(t) = \mathbb{E}_{P_e}[R_g \mid F_g = t]$ for the conditional MSE of g on environment e and $P^e_{F_g}$ for the distribution of F_g under P_e, let

$$\mathcal{G}_e = \{g \in \mathcal{G} : \sigma^2_{g,e}(t) \equiv c_{g,e} \text{ for } P^e_{F_g}\text{-almost every } t\}$$

be the corresponding invariant class, and write $\mathcal{G}_\cap = \bigcap_{e=1}^{K} \mathcal{G}_e$ for their intersection.

Lemma 6.7.2. *Under Assumption 6.7.1,* $g_0 \in \mathcal{G}_\cap$.

In each environment the residual of g_0 is ε_e, independent of $g_0(X^e)$, so its conditional MSE is constant; the proof is given in Appendix D.7.

The environment-weighted CvM (eCvM) penalty

$$\mathcal{C}^{\mathrm{ew}}(g) = \sum_{e=1}^{K} \omega_e \, \mathcal{C}^e(g) \tag{6.5}$$

sums the single-environment penalties $\mathcal{C}^e$, each computed from (6.2) under P_e. Because the $\mathcal{C}^e$ are non-negative and the environment weights positive, and by Proposition 6.4.1 applied under each P_e, $\mathcal{C}^{\mathrm{ew}}(g) = 0$ if and only if $g \in \mathcal{G}_\cap$; in particular, $\mathcal{C}^{\mathrm{ew}}(g_0) = 0$ by Lemma 6.7.2.

The intersection can still be larger than $\{g_0\}$, since a prediction function may be invariant in every environment without being causal. We therefore impose, as an identification condition, that no such prediction function exists. Equivalently, the environments are heterogeneous enough to leave g_0 as the only jointly invariant prediction function.

Assumption 6.7.3 (Environment heterogeneity)**.** *For every* $g \in \mathcal{G} \setminus \{g_0\}$*, there is an environment* e *with* $\mathcal{C}^e(g) > 0$.

Proposition 6.7.4 (Identification across environments)**.** *Under Assumptions 6.7.1 and 6.7.3,* $\mathcal{G}_\cap = \{g_0\}$*; equivalently,* g_0 *is the only* $g \in \mathcal{G}$ *with* $\mathcal{C}^{\mathrm{ew}}(g) = 0$.

The proof is given in Appendix D.8. The structural model guarantees that g_0 belongs to $\mathcal{G}_\cap$, and heterogeneity makes it the unique element.

6.8 Synthetic experiment

Consider the linear non-Gaussian data-generating process (DGP):

$$X_1 \sim \mathcal{N}(0,1), \quad Y \leftarrow 2X_1 + \varepsilon_Y, \quad X_2 \leftarrow \tfrac{1}{2}Y + \varepsilon_{\mathrm{ch}},$$

in which X_2 is a child of Y. The noises ε_Y and $\varepsilon_{\mathrm{ch}}$ are independent of X_1 and of each other, each an equal mixture of two Gaussians,

Estimator	Non-shift	Sign flip	Tripled child noise	Strong sign flip
ERM ($\lambda = 0$)	**2.86**	9.17	8.93	13.9
CvM ($\lambda = 10^6$)	4.01	**4.03**	**4.02**	**4.01**
Oracle g_0	4.01	4.00	4.00	4.01

Table 6.1: Test MSE on the linear non-Gaussian DGP, under the unshifted distribution and under the three child-mechanism shifts described above. Training and test sizes are 2000 and 1000, respectively.

$$\varepsilon_Y \sim \tfrac{1}{2}\mathcal{N}(-2,\, 0.1^2) + \tfrac{1}{2}\mathcal{N}(2,\, 0.1^2), \qquad \varepsilon_{\text{ch}} \sim \tfrac{1}{2}\mathcal{N}(-1.5,\, 0.1^2) + \tfrac{1}{2}\mathcal{N}(1.5,\, 0.1^2),$$

so each is bimodal, centred at zero, and non-Gaussian. The causal target is $g_0(X) = 2X_1$. To probe robustness, the test distribution is modified through the child mechanism only, while the structural equation $Y \leftarrow 2X_1 + \varepsilon_Y$ is left unchanged. We consider three perturbations: a sign flip $X_2 \leftarrow -\frac{1}{2}Y + \varepsilon_{\text{ch}}$, a tripled child noise $X_2 \leftarrow Y + 3\varepsilon_{\text{ch}}$, and a strong sign flip $X_2 \leftarrow -Y + \varepsilon_{\text{ch}}$.

Within the linear class $g_\beta(X) = \beta_1 X_1 + \beta_2 X_2$, the empirical objective is minimised by exhaustive grid search on a 101×101 grid over $[-1, 4]^2$. Table 6.1 reports the test MSE of the unpenalised estimator and of the proposed estimator at $\lambda = 10^6$, averaged over 20 replicates.

To compare with methods that require an anchor or environment labels, we augment the child mechanism with an exogenous anchor $A \in \{-1, 0, +1\}$, uniformly distributed and independent of $(X_1, \varepsilon_Y, \varepsilon_{\text{ch}})$, through $X_2 \leftarrow \frac{1}{2}Y + \varepsilon_{\text{ch}} + 0.3\,A$. The structural equation for Y is unchanged, so $g_0(X) = 2X_1$ remains the target. The three values of A partition the 2000 training points into three environments; every method is fitted on this same pooled sample and differs only in whether the anchor variable or the environment labels are used. Test data are drawn under the three shifts above with $A = 0$. Anchor regression observes A as a variable, DRIG (with the $A = 0$ environment as its observational reference e_0) and the eCvM observe the three anchor-induced environments through their labels, and the pooled CvM (ignoring environments) observes (X, Y) alone. The objectives and worst-case classes are those recorded in Section 6.1, and the eCvM applies the single-environment construction within each anchor-induced environment, averaging via $\mathcal{C}^{\text{ew}}$ (Section 6.7). Table 6.2 reports the worst-case test MSE across the three child-mechanism shifts, together with the ℓ_2 distance of each estimated coefficient to $\beta_0 = (2, 0)$.

The comparison illustrates one regime rather than a general ranking; the shift acts only through the child X_2. DRIG and anchor regression reduce the prediction function's sensitivity to how the training environments differ, but the anchor only shifts the mean of X_2 by $0.3A$,

Method	Best hyperparameter	Worst-case MSE	$\|\hat{\beta} - \beta_0\|_2$
ERM	—	13.9 ± 0.08	0.868 ± 0.01
DRIG	$\gamma = 10$	11.7 ± 0.74	0.862 ± 0.12
Anchor regression	$\gamma = 100$	6.02 ± 0.25	0.320 ± 0.03
CvM (pooled)	$\lambda = 10^5$	4.29 ± 0.08	0.0713 ± 0.02
eCvM	$\lambda = 10^5$	$\mathbf{4.05 \pm 0.03}$	$\mathbf{0.040 \pm 0.01}$
Oracle g_0	—	4.01	0

Table 6.2: Worst-case test MSE across the three child-mechanism shifts and ℓ_2 distance to $\beta_0 = (2, 0)$, for the linear non-Gaussian DGP with three-level anchor variable A. Means $\pm$ standard errors over 20 replications. Training and test sizes are 2000 and 1000, respectively. The robustness parameter is tuned over $\gamma \in \{1, 5, 10, 50, 100\}$ (anchor regression, DRIG) and the penalty weight over $\lambda \in \{0, 10, 10^2, 10^3, 10^4, 10^5, 10^6\}$ (CvM, eCvM); the reported configuration minimises the mean worst-case shifted test MSE.

too small a difference to reveal that X_2 is non-causal. DRIG therefore stays close to ERM, and anchor regression, though it improves more, still keeps weight on X_2. The pooled CvM estimator, from the single observational distribution alone, already lands close to β_0, because a prediction function that uses X_2 has a non-flat conditional MSE and is penalised directly, without heterogeneous environments. The eCvM, given the three anchor environments, sharpens this further, bringing both the worst-case risk and the coefficient closest to the oracle.

6.9 Ongoing work

Several properties of the framework are still under study. Identification currently relies on an additive-noise model together with an orthogonality condition, and how far either can be weakened is under exploration. A second open direction is to clarify the relationship to invariant risk minimisation, which seeks a comparable invariance from observed environments rather than from a single distribution.

The underlying manuscript develops further results. They are omitted here so that the present chapter can stay focused on the main construction. These include convergence rates for the penalised estimator towards the causal prediction function g_0, a goodness-of-fit test based on the CvM statistic, and a combined objective that incorporates both eCvM and DRIG when multiple environments are available. Beyond the linear illustration presented here, the method has also been explored on nonlinear simulations and real-world data.

Chapter 7

Discussion and future work

Whether a conclusion is causal depends on its invariance to changes in the data-generating process. An observed association may be specific to a particular population, sampling scheme, or set of background conditions. A causal relationship, by contrast, is expected to persist across a range of such changes, reflecting an underlying feature of the system rather than a contingent regularity in the data. This thesis designs for that persistence directly: in each problem we decide in advance which perturbations a conclusion must withstand, and build a method that controls a chosen criterion across them. The result is an estimator, a test, and a prediction function that each come with such a guarantee, and a generative model that supplies the ground truth against which stability is judged. This concluding chapter returns to that common perspective, to set out what it offers, where the methods reach their limits, and what it leaves open.

7.1 Securing and evaluating stability

Each chapter carries the framework of Chapter 2 into a different regime. Chapter 3 controls the variance of an effect estimator under weak overlap, and Chapter 4 the Type-I error of a test-then-pool procedure under source heterogeneity, both by construction; Chapter 5 secures the fidelity of simulated data under the complexity of generative models, giving a ground truth against which stability can be evaluated; and Chapter 6 attains prediction stability under distribution shift by inheriting it from an invariant causal mechanism.

Beneath them lies a single question: against which perturbations is a procedure stable, and how well? Answering it means fixing a class of perturbations Π and a criterion J and reporting a level δ that the worst-case criterion $\sup_{Q \in \Pi} J(\hat{\theta}, Q)$ does not exceed (Definition 2.2.1). The two terms pull against each other: a wider Π can raise δ. A guarantee therefore carries weight especially when the class is wide and the level it secures is still low. Most of the thesis fixes the

class the problem dictates and drives δ down within it; Chapter 6 instead makes the tension itself the object, tracing a family of perturbations whose limit is causal invariance.

7.2 Limitations

The methods developed in this thesis target specific sources of instability. Their guarantees are correspondingly focused, and each leaves aspects of the broader problem unresolved. The proposed estimator of Chapter 3 reduces variance at a given data-generating distribution, but its efficiency gain holds only within a restricted submodel and not in the full nonparametric model, where the estimator is irregular. The test-then-pool procedure of Chapter 4 either pools an external source in full or discards it, rather than using it only partially, according to how closely it agrees with the primary study. The generative model of Chapter 5 is consistent but lacks convergence rates, so its point estimates and simulated samples come without quantified uncertainty. The prediction function of Chapter 6 derives its robustness from an additive noise model, requires a further orthogonality condition for identification, and remains stable only under perturbations that preserve the conditional distribution of the response.

Taken together, these limitations emphasise that stability is never an absolute property. Every guarantee is tied to assumptions on the data, the perturbations admitted, or the causal structure behind them. Understanding where those assumptions fail is therefore part of understanding the guarantee itself.

7.3 Future directions

Extensions specific to each method are discussed within the corresponding chapters. We mention here only directions that arise across the thesis as a whole.

Many causal questions involve richer structure than the settings studied here. Some already enter the thesis, such as longitudinal treatments in Chapter 5, while others, such as mediation, interference, and identification through instrumental variables, lie outside its scope. Studying stability for these structures more fully would mean choosing perturbation classes and notions of invariance adapted to each, since a notion adequate for an average treatment effect need not preserve the quantities relevant to mediation, interference, or a dynamic treatment regime. These are also among the settings practitioners most often face, which makes a stability theory

aimed explicitly at them the more pressing.

Causal quantities are often used to support decisions rather than as ends in themselves. It remains to study how stability carries into decision-making: how an effect estimate and its uncertainty should shape a rule for assigning treatment, and whether a rule built on a stable quantity inherits that stability under the same perturbations. This would connect the guarantees developed here to policy learning and individualised treatment rules, where a distribution shift is paid for directly in the quality of the decision.

7.4 Concluding remarks

A causal conclusion earns its name by holding beyond the data from which it was obtained. Stability gives this statement meaning: to call a quantity causal is to specify a class of perturbations under which the conclusion is expected to remain valid. The strength of the conclusion is therefore determined by the class for which stability can be established.

The methods developed in this thesis follow this principle. They identify the perturbation classes relevant to the problem, control performance across them, and establish the corresponding guarantees. The objective is not stability as an abstract ideal, but stability for a well-specified, scientifically relevant class.

Appendix A

Supplementary material for Chapter 3

A.1 Main proofs

Throughout the appendix, expressions involving estimated nuisance functions are read in their cross-fitted form from Algorithm 3.1: for $i \in D_k$, $\hat{\pi}, \hat{\mu}^x, \hat{w}^*$ and w^{*0} denote the fold-specific functions trained on D_{-k}. The fold index is suppressed unless needed.

A.1.1 Cross-fitting lemma

Lemma A.1.1 (Cross-fitting)**.** *Let $W_1, \dots, W_n$ be i.i.d. from P, let $K \in \mathbb{N}$ be fixed, and let $\{D_k\}_{k=1}^K$ be a partition of $\{1, \dots, n\}$, either deterministic or generated independently of $W_1, \dots, W_n$, satisfying $|D_k|/n \to 1/K$. The lemma is read conditionally on the fold assignment. For each k, define*

$$\mathcal{F}_{-k} \coloneqq \sigma\big(\{W_i : i \notin D_k\}\big),$$

and let $\hat{\xi}^{(-k)}$ be $\mathcal{F}_{-k}$-measurable. Let $\psi(z, \xi)$ be real-valued and jointly measurable, and assume that for any $i \in D_k$, $\mathbb{E}\Big[\psi(W_i; \hat{\xi}^{(-k)})^2 \,\Big|\, \mathcal{F}_{-k}\Big] = O_P(1)$. Define

$$m_k \coloneqq \mathbb{E}\Big[\psi(W_i; \hat{\xi}^{(-k)}) \,\Big|\, \mathcal{F}_{-k}\Big], \quad \forall i \in D_k, \qquad P_n^{\mathrm{cf}}\psi \coloneqq \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in D_k} \psi(W_i; \hat{\xi}^{(-k)}).$$

If $m_k \xrightarrow{p} m$ for each k, then

$$P_n^{\mathrm{cf}}\psi \xrightarrow{p} m.$$

Proof. It suffices to show that $T_k \xrightarrow{p} m$ for each fixed k, where

$$T_k \coloneqq \frac{1}{|D_k|} \sum_{i \in D_k} \psi(W_i; \hat{\xi}^{(-k)}),$$

since

$$P_n^{\mathrm{cf}} \psi = \sum_{k=1}^{K} \frac{|D_k|}{n} T_k \xrightarrow{p} \sum_{k=1}^{K} \frac{1}{K} m = m.$$

Fix k and $\epsilon > 0$. Conditional on $\mathcal{F}_{-k}$, the variables $\psi(W_i; \hat{\xi}^{(-k)})$, $i \in D_k$, are i.i.d. with mean m_k and variance

$$v_k \coloneqq \mathrm{Var}\left(\psi(W_i; \hat{\xi}^{(-k)}) \,|\, \mathcal{F}_{-k}\right) \leq \mathbb{E}\Big[\psi(W_i; \hat{\xi}^{(-k)})^2 \,|\, \mathcal{F}_{-k}\Big] = O_P(1).$$

By conditional Chebyshev,

$$\Pr\Big(|T_k - m_k| > \epsilon \,|\, \mathcal{F}_{-k}\Big) \leq \min\left\{ \frac{v_k}{|D_k|\,\epsilon^2},\, 1 \right\}.$$

Taking expectations,

$$\Pr(|T_k - m_k| > \epsilon) \leq \mathbb{E}\left[\min\left\{ \frac{v_k}{|D_k|\,\epsilon^2},\, 1 \right\} \right].$$

Since $v_k = O_P(1)$ and $|D_k| \to \infty$, we have $v_k/(|D_k|\epsilon^2) = o_P(1)$, and bounded convergence yields

$$\Pr(|T_k - m_k| > \epsilon) \to 0.$$

Thus $T_k - m_k \xrightarrow{p} 0$, and since $m_k \xrightarrow{p} m$, it follows that $T_k \xrightarrow{p} m$. □

Lemma A.1.2 (Cross-fitting rate)**.** *Under the conditions of Lemma A.1.1, suppose that for any $i \in D_k$,*

$$\mathbb{E}\Big[\psi(W_i; \hat{\xi}^{(-k)})^2 \,|\, \mathcal{F}_{-k}\Big] = o_P(1).$$

Then

$$\sqrt{n}\,(T_k - m_k) = o_P(1) \quad \textit{for each } k,$$

and

$$\sqrt{n}\left(P_n^{\mathrm{cf}}\psi - \sum_{k=1}^{K}\frac{|D_k|}{n}\,m_k\right) = o_P(1).$$

Proof. Conditional on $\mathcal{F}_{-k}$, the variables $\psi(W_i;\hat{\xi}^{(-k)})$, $i \in D_k$, are i.i.d. with variance $v_k \leq \mathbb{E}[\psi(W_i;\hat{\xi}^{(-k)})^2 \,|\, \mathcal{F}_{-k}] = o_P(1)$. Hence, for any $\eta > 0$,

$$\Pr\Big(\sqrt{n}\,|T_k - m_k| > \eta \,\big|\, \mathcal{F}_{-k}\Big) \leq \min\left\{\frac{n\,v_k}{|D_k|\,\eta^2},\, 1\right\}.$$

Taking expectations and using $|D_k|/n \to 1/K$ and $v_k = o_P(1)$ yields

$$\Pr(\sqrt{n}\,|T_k - m_k| > \eta) \to 0.$$

Thus $\sqrt{n}(T_k - m_k) = o_P(1)$. Since $P_n^{\mathrm{cf}}\psi - \sum_{k=1}^{K}(|D_k|/n)\,m_k = \sum_{k=1}^{K}(|D_k|/n)(T_k - m_k)$, with K fixed and $|D_k|/n \leq 1$,

$$\sqrt{n}\left(P_n^{\mathrm{cf}}\psi - \sum_{k}\frac{|D_k|}{n}m_k\right) = \sum_{k=1}^{K}\frac{|D_k|}{n}\,\sqrt{n}\,(T_k - m_k) = o_P(1),$$

since each $\sqrt{n}\,(T_k - m_k) = o_P(1)$.

□

A.1.2 Proof of Proposition 3.3.3

Proof. This proof follows Taufiq et al. (2024). Since $w(Y) = \mathbb{E}[h(X,Z)\,|\,Y]$,

$$\mathbb{E}[h(X,Z)Y\,|\,Y] = w(Y)Y, \qquad \mathbb{E}[h(X,Z)Y] = \mathbb{E}[w(Y)Y] = \theta,$$

and the conditional-variance identity gives

$$\mathrm{Var}\,(h(X,Z)Y) - \mathrm{Var}\,(w(Y)Y) = \mathbb{E}[\mathrm{Var}(h(X,Z)Y\,|\,Y)] = \mathbb{E}[Y^2\,\mathrm{Var}(h(X,Z)\,|\,Y)] \geq 0.$$

When the nuisances are known, so that $w^0 = w$, $\hat{\theta}^0_{IPW}$ and $\hat{\theta}^0_{MR}$ are sample averages of i.i.d. copies of $h(X,Z)Y$ and $w(Y)Y$, and $\mathrm{Var}(\hat{\theta}^0_{IPW}) - \mathrm{Var}(\hat{\theta}^0_{MR}) = n^{-1}\,\mathbb{E}[Y^2\,\mathrm{Var}(h(X,Z)\,|\,Y)] \geq 0$. The conclusion for AIPW and AMR follows by replacing Y with Y^*. □

A.1.3 Proof of Proposition 3.3.5

Proof. Part (a). Condition on X inside the conditional expectation:

$$\begin{aligned}\mathbb{E}\Big[\frac{X-\pi(I)}{\pi(I)\{1-\pi(I)\}}\,\Big|\,Y\Big] &= \sum_{x\in\{0,1\}} \Pr(X=x\,|\,Y)\,\mathbb{E}\Big[\frac{x-\pi(I)}{\pi(I)\{1-\pi(I)\}}\,\Big|\,Y,X=x\Big]\\ &= \Pr(X=1\,|\,Y)\,\mathbb{E}[\pi(I)^{-1}\,|\,Y,X=1]\\ &\quad - \Pr(X=0\,|\,Y)\,\mathbb{E}[\{1-\pi(I)\}^{-1}\,|\,Y,X=0].\end{aligned}$$

Under $Y \perp\!\!\!\perp I\,|\,X$, $\mathbb{E}[\pi(I)^{-1}\,|\,Y,X=1] = \mathbb{E}[\pi(I)^{-1}\,|\,X=1]$. Since the propensity score depends on Z only through I, $\mathbb{E}[\mathbb{1}\{X=1\}\,|\,I] = \pi(I)$, so

$$\mathbb{E}[\pi(I)^{-1}\,|\,X=1] = \frac{\mathbb{E}[\pi(I)^{-1}\,\mathbb{1}\{X=1\}]}{\Pr(X=1)} = \frac{\mathbb{E}[\pi(I)^{-1}\,\pi(I)]}{\Pr(X=1)} = \frac{1}{\Pr(X=1)}.$$

For $X=0$ the same steps apply. Under $Y \perp\!\!\!\perp I\,|\,X$, $\mathbb{E}[\{1-\pi(I)\}^{-1}\,|\,Y,X=0] = \mathbb{E}[\{1-\pi(I)\}^{-1}\,|\,X=0]$, and since $\mathbb{E}[\mathbb{1}\{X=0\}\,|\,I] = 1-\pi(I)$,

$$\begin{aligned}\mathbb{E}[\{1-\pi(I)\}^{-1}\,|\,X=0] &= \frac{\mathbb{E}[\{1-\pi(I)\}^{-1}\,\mathbb{1}\{X=0\}]}{\Pr(X=0)}\\ &= \frac{\mathbb{E}[\{1-\pi(I)\}^{-1}\{1-\pi(I)\}]}{\Pr(X=0)} = \frac{1}{\Pr(X=0)}.\end{aligned}$$

Substituting $\mathbb{E}[\pi(I)^{-1}\,|\,Y,X=1] = \Pr(X=1)^{-1}$ and $\mathbb{E}[\{1-\pi(I)\}^{-1}\,|\,Y,X=0] = \Pr(X=0)^{-1}$ into the conditional expectation above gives the stated form.

Part (b). Since the propensity score depends only on (I,C), write $\pi(I,C) = \Pr(X=1\,|\,I,C)$. Conditioning on X and then on C within each treatment arm,

$$\begin{aligned}\mathbb{E}\Big[\frac{X-\pi(I,C)}{\pi(I,C)\{1-\pi(I,C)\}}\,\Big|\,Y\Big] &= \Pr(X=1\,|\,Y)\,\mathbb{E}[\pi(I,C)^{-1}\,|\,Y,X=1]\\ &\quad - \Pr(X=0\,|\,Y)\,\mathbb{E}[\{1-\pi(I,C)\}^{-1}\,|\,Y,X=0].\end{aligned}$$

Under $I \perp\!\!\!\perp Y\,|\,X,C$, the inner expectations factor through C: for $x=1$,

$$\mathbb{E}[\pi(I,C)^{-1}\,|\,Y,X=1] = \mathbb{E}[\mathbb{E}\{\pi(I,C)^{-1}\,|\,X=1,C\}\,|\,Y,X=1].$$

The same argument conditional on C, using $\mathbb{E}[\mathbb{1}\{X=1\}\,|\,I,C] = \pi(I,C)$, yields $\mathbb{E}[\pi(I,C)^{-1}\,|\,X=1,C] = \Pr(X=1\,|\,C)^{-1}$, and analogously $\mathbb{E}[\{1-\pi(I,C)\}^{-1}\,|\,X=0,C] = \Pr(X=0\,|\,C)^{-1}$.

Combining each inner identity with the factorisation above (and its $x = 0$ analogue),

$$\mathbb{E}[\pi(I,C)^{-1} \,|\, Y, X = 1] = \mathbb{E}[\Pr(X = 1 \,|\, C)^{-1} \,|\, Y, X = 1],$$
$$\mathbb{E}[\{1 - \pi(I,C)\}^{-1} \,|\, Y, X = 0] = \mathbb{E}[\Pr(X = 0 \,|\, C)^{-1} \,|\, Y, X = 0].$$

Substituting these into the conditional expectation at the start of part (b) yields

$$\mathbb{E}\left[\frac{X - \pi(I,C)}{\pi(I,C)\{1 - \pi(I,C)\}} \,\middle|\, Y\right] = \Pr(X = 1 \,|\, Y)\, \mathbb{E}[\Pr(X = 1 \,|\, C)^{-1} \,|\, Y, X = 1] - \Pr(X = 0 \,|\, Y)\, \mathbb{E}[\Pr(X = 0 \,|\, C)^{-1} \,|\, Y, X = 0],$$

the stated form. □

A.1.4 Proof of Proposition 3.3.6

Proof. Set $\Phi_0 \coloneqq w^{*0}(\hat{Y}^*)\hat{Y}^*$, so that $\Delta = \hat{w}^*(\hat{Y}^*)\hat{Y}^* - \Phi_0$. By definition of w^{*0}, $\Phi_0 = \mathbb{E}[\hat{h}(X,Z)\hat{Y}^* \,|\, \hat{Y}^*]$, hence $\hat{h}(X,Z)\hat{Y}^* - \Phi_0$ is orthogonal in $L_2(P)$ to every $\sigma(\hat{Y}^*)$-measurable function; in particular,

$$\mathrm{Cov}\,(\Phi_0,\, \hat{h}(X,Z)\hat{Y}^* - \Phi_0) = 0, \qquad \mathrm{Cov}\,(\hat{h}(X,Z)\hat{Y}^* - \Phi_0,\, \Delta) = 0.$$

By (3.10), $\mathbb{E}\Phi_0 = \mathbb{E}[\hat{h}(X,Z)\hat{Y}^*]$, so

$$\mathbb{E}[\hat{h}(X,Z)\hat{Y}^*] - \theta = \mathbb{E}\Phi_0 - \theta = b, \qquad \mathbb{E}[\hat{w}^*(\hat{Y}^*)\hat{Y}^*] - \theta = b + B.$$

The orthogonality of $\hat{h}(X,Z)\hat{Y}^* - \Phi_0$ to $\sigma(\hat{Y}^*)$ gives the variance decomposition $\mathrm{Var}(\hat{h}(X,Z)\hat{Y}^*) = \mathrm{Var}(\Phi_0) + V_h$, and $\Phi_0 = w^{*0}(\hat{Y}^*)\hat{Y}^*$. Since $W_1, \ldots, W_n$ are i.i.d. from P,

$$\mathbb{E}[(\hat{\theta}_{AIPW} - \theta)^2] = b^2 + n^{-1}\big[\,\mathrm{Var}\,(w^{*0}(\hat{Y}^*)\hat{Y}^*) + V_h\big],$$
$$\mathbb{E}[(\hat{\theta}_{AMR} - \theta)^2] = (b + B)^2 + n^{-1}\,\mathrm{Var}\,(\hat{w}^*(\hat{Y}^*)\hat{Y}^*).$$

Taking the differences between the two terms, we obtain (3.9). If $\hat{w}^* = w^{*0}$, then $\Delta = 0$, so $B = 0$ and $\mathrm{Var}(\hat{w}^*(\hat{Y}^*)\hat{Y}^*) = \mathrm{Var}(w^{*0}(\hat{Y}^*)\hat{Y}^*)$, and the right-hand side equals $-V_h/n$. □

A.1.5 Proof of Theorem 3.4.3

Proof. We use the cross-fitting convention stated at the beginning of the appendix. For observations in fold D_k, all nuisance functions are trained on D_{-k} and are therefore fixed conditional on $\mathcal{F}_{-k}$. Under Assumption 3.4.1, both Y^* and $\hat{Y}^*$ have controlled fourth moments:

$$\mathbb{E}[(Y^*)^4] < \infty, \qquad \mathbb{E}[(\hat{Y}^*)^4 \,|\, \mathcal{F}_{-k}] = O_P(1),$$

where $\mathcal{F}_{-k} = \sigma(\{W_i : i \notin D_k\})$. This follows because μ^* and $\hat{\mu}^*$ are uniformly bounded.

We first prove consistency of $\hat{\theta}^0_{AMR}$. Write

$$\hat{\theta}^0_{AMR} - \theta = \underbrace{P_n[\{w^{*0}(\hat{Y}^*) - \hat{h}(X,Z)\}\hat{Y}^*]}_{\text{Ⓐ}} + \underbrace{P_n[\hat{h}(X,Z)\hat{Y}^*] - \theta}_{\text{Ⓑ}}.$$

For term Ⓐ, the bridge identity holds conditionally on the training fold:

$$\begin{aligned}\mathbb{E}[w^{*0}(\hat{Y}^*)\hat{Y}^* \,|\, \mathcal{F}_{-k}] &= \mathbb{E}[\mathbb{E}[\hat{h}(X,Z) \,|\, \hat{Y}^*, \mathcal{F}_{-k}]\hat{Y}^* \,|\, \mathcal{F}_{-k}] \\ &= \mathbb{E}[\hat{h}(X,Z)\hat{Y}^* \,|\, \mathcal{F}_{-k}].\end{aligned}$$

Thus the summand in Ⓐ has conditional mean zero on each validation fold. By Assumption 3.4.1, $\hat{h}$ and w^{*0} are bounded with probability tending to one, and

$$\mathbb{E}[(\hat{Y}^*)^2 \,|\, \mathcal{F}_{-k}] = O_P(1).$$

Hence the conditional second moment of the summand in Ⓐ is $O_P(1)$. Applying Lemma A.1.1 gives Ⓐ $\xrightarrow{p}$ 0.

It remains to control Ⓑ. Since

$$\hat{Y}^* = Y - \hat{\pi}(Z)\hat{\mu}^0(Z) - \{1 - \hat{\pi}(Z)\}\hat{\mu}^1(Z),$$

a direct calculation gives

$$\mathbb{E}[\hat{h}(X,Z)\hat{Y}^* \,|\, \mathcal{F}_{-k}] - \theta = \mathbb{E}\Big[\frac{\hat{\pi}(Z) - \pi(Z)}{\hat{\pi}(Z)}\{\hat{\mu}^1(Z) - \mu^1(Z)\} + \frac{\hat{\pi}(Z) - \pi(Z)}{1 - \hat{\pi}(Z)}\{\hat{\mu}^0(Z) - \mu^0(Z)\} \,\Big|\, \mathcal{F}_{-k}\Big].$$

Under strict overlap, this is bounded in absolute value by a constant times

$$\|\hat{\pi} - \pi\|_2(\|\hat{\mu}^1 - \mu^1\|_2 + \|\hat{\mu}^0 - \mu^0\|_2).$$

Therefore the conditional mean of Ⓑ is $o_P(1)$ if either $\hat{\pi}$ is consistent, with bounded $\hat{\mu}^x$, or both $\hat{\mu}^0$ and $\hat{\mu}^1$ are consistent, with bounded $\hat{\pi}$ away from 0 and 1. The conditional second moment is $O_P(1)$ by Assumption 3.4.1. Applying Lemma A.1.1 gives $\text{Ⓑ} \xrightarrow{p} 0$.

Combining the two displays gives

$$\hat{\theta}^0_{AMR} \xrightarrow{p} \theta.$$

Now consider the feasible estimator. We have

$$\hat{\theta}_{AMR} - \hat{\theta}^0_{AMR} = P_n[\{\hat{w}^* - w^{*0}\}(\hat{Y}^*)\hat{Y}^*].$$

By Cauchy–Schwarz,

$$|\hat{\theta}_{AMR} - \hat{\theta}^0_{AMR}| \leq \|\{\hat{w}^* - w^{*0}\}(\hat{Y}^*)\|_{n,2}\, \|\hat{Y}^*\|_{n,2}.$$

The second factor is $O_P(1)$ by Assumption 3.4.1. For the first factor, apply Lemma A.1.1 to

$$\psi_1(W) = \big((\hat{w}^* - w^{*0})(\hat{Y}^*)\big)^2.$$

By Assumption 3.4.2, its conditional mean is

$$\|(\hat{w}^* - w^{*0})(\hat{Y}^*)\|_2^2 = o_P(1),$$

and its conditional second moment is $O_P(1)$, since $\hat{w}^*$ is uniformly bounded with probability tending to one by Assumption 3.4.2, and $w^{*0} = \mathbb{E}[\hat{h} \,|\, \hat{Y}^*]$ is bounded by $\|\hat{h}\|_\infty$ under strict overlap. Hence

$$\|(\hat{w}^* - w^{*0})(\hat{Y}^*)\|_{n,2} = o_P(1).$$

Therefore

$$\hat{\theta}_{AMR} - \hat{\theta}^0_{AMR} = o_P(1),$$

and since $\hat{\theta}^0_{AMR} \xrightarrow{p} \theta$, we conclude that

$$\hat{\theta}_{AMR} \xrightarrow{p} \theta.$$

□

A.1.6 Proof of Theorem 3.4.4

Throughout this proof, when $\hat{\xi}^0$ appears we read P_n as the cross-fitted average $\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in D_k}(\cdot)$ evaluated at $\hat{\xi}^{0,(-k)}$ trained on D_{-k}, and $\mathbb{E}[\cdot]$ as the matching fold-weighted conditional expectation $\sum_{k=1}^{K}(|D_k|/n)\,\mathbb{E}[\cdot \mid \mathcal{F}_{-k}]$; thus $(P_n - \mathbb{E})$ is the centred cross-fitted empirical process.

In Theorem 3.4.4, condition (a) controls the bias from estimating ξ that appears in term Ⓐ below. Condition (c) is a projection-stability condition that controls the cross-fitted empirical-process remainder in term Ⓒ below.

Proof. Write

$$\begin{aligned}\hat{\theta}^0_{AMR} - \theta = &\underbrace{\mathbb{E}\varphi_{AMR}(W;\hat{\xi}^0)}_{\text{Ⓐ}} \\ &+ \underbrace{(P_n - \mathbb{E})\varphi_{AMR}(W;\xi)}_{\text{Ⓑ}} \\ &+ \underbrace{(P_n - \mathbb{E})\{\varphi_{AMR}(W;\hat{\xi}^0) - \varphi_{AMR}(W;\xi)\}}_{\text{Ⓒ}}.\end{aligned}$$

Here Ⓐ is the plug-in bias from the estimated nuisances, Ⓑ the leading stochastic term, and Ⓒ a cross-fitted remainder. We show Ⓐ and Ⓒ are $o_P(n^{-1/2})$, so that Ⓑ alone determines the limit.

For Ⓐ: By (3.10), $\mathbb{E}\varphi_{AMR}(W;\hat{\xi}^0) = \mathbb{E}\varphi_{AIPW}(W;\hat{\xi}^0)$.

A direct expansion gives

$$\mathbb{E}\varphi_{AIPW}(W;\hat{\xi}^0) = \mathbb{E}\Big[\frac{\hat{\pi}(Z)-\pi(Z)}{\hat{\pi}(Z)}\{\hat{\mu}^1(Z)-\mu^1(Z)\} + \frac{\hat{\pi}(Z)-\pi(Z)}{1-\hat{\pi}(Z)}\{\hat{\mu}^0(Z)-\mu^0(Z)\}\Big],$$

so under strict overlap,

$$|\text{Ⓐ}| \lesssim \|\hat{\pi}-\pi\|_2(\|\hat{\mu}^1-\mu^1\|_2 + \|\hat{\mu}^0-\mu^0\|_2) = o_P(n^{-1/2})$$

by condition (a). See also Chernozhukov et al. (2018, Thm. 3.1).

For Ⓑ: by the CLT,

$$\sqrt{n}\,Ⓑ \rightsquigarrow \mathcal{N}(0, \sigma^2_{AMR,0}), \qquad \sigma^2_{AMR,0} = \mathrm{Var}\left(\varphi_{AMR}(W;\xi)\right).$$

For Ⓒ: For $i \in D_k$, set

$$\psi_i = \varphi_{AMR}(W_i; \hat{\xi}^{0,(-k)}) - \varphi_{AMR}(W_i;\xi).$$

Conditional on $\mathcal{F}_{-k}$, the terms ψ_i are i.i.d. on fold D_k. To bound Ⓒ it suffices to show $\mathbb{E}[\psi_i^2 \,|\, \mathcal{F}_{-k}] = o_P(1)$, which we obtain by splitting ψ_i into three pieces,

$$\psi_i = \{w^{*0}(\hat{Y}_i^*) - w^*(\hat{Y}_i^*)\}\hat{Y}_i^* + \{w^*(\hat{Y}_i^*) - w^*(Y_i^*)\}\hat{Y}_i^* + w^*(Y_i^*)(\hat{Y}_i^* - Y_i^*).$$

For $\{w^{*0}(\hat{Y}_i^*) - w^*(\hat{Y}_i^*)\}\hat{Y}_i^*$, condition (c) in Theorem 3.4.4 gives

$$\|w^{*0}(\hat{Y}^*) - w^*(\hat{Y}^*)\|_2 = o_P(1).$$

For $\{w^*(\hat{Y}_i^*) - w^*(Y_i^*)\}\hat{Y}_i^*$, condition (b) in Theorem 3.4.4 gives $\hat{Y}^* \to Y^*$ in probability. Since w^* is bounded and continuous at Y^* almost surely,

$$w^*(\hat{Y}^*) \to w^*(Y^*) \quad \text{in probability.}$$

Because w^* is bounded, this convergence is also in L_2, so

$$\|w^*(\hat{Y}^*) - w^*(Y^*)\|_2 = o_P(1).$$

For $w^*(Y_i^*)(\hat{Y}_i^* - Y_i^*)$, boundedness of w^* and condition (b) in Theorem 3.4.4 give

$$\|w^*(Y^*)(\hat{Y}^* - Y^*)\|_2 \le \|w^*\|_\infty \|\hat{Y}^* - Y^*\|_2 = o_P(1).$$

For the first two terms, the weight differences are bounded, w^{*0} by strict overlap and w^* by condition (c), so Cauchy–Schwarz together with $\mathbb{E}[(\hat{Y}^*)^4 \,|\, \mathcal{F}_{-k}] = O_P(1)$ (Assumption 3.4.1) upgrades the L_2 bounds above to

$$\left\|\{w^{*0}(\hat{Y}^*) - w^*(\hat{Y}^*)\}\hat{Y}^*\right\|_2 = o_P(1), \qquad \left\|\{w^*(\hat{Y}^*) - w^*(Y^*)\}\hat{Y}^*\right\|_2 = o_P(1).$$

Hence $\mathbb{E}[\psi_i^2 \mid \mathcal{F}_{-k}] = o_P(1)$, and therefore by Lemma A.1.2,

$$\sqrt{n}\,\mathbb{C} = o_P(1).$$

Combining Ⓐ, Ⓑ and Ⓒ,

$$\sqrt{n}(\hat{\theta}^0_{AMR} - \theta) = \sqrt{n}\,(P_n - \mathbb{E})\,\varphi_{AMR}(W;\xi) + o_P(1) \rightsquigarrow \mathcal{N}(0, \sigma^2_{AMR,0}).$$

□

A.1.7 Proof of Proposition 3.4.6

Proof. Set $\epsilon^* := \hat{w}^* - w^{*0}$. Then

$$\hat{\theta}_{AMR} - \theta = P_n(\epsilon^*(\hat{Y}^*)\,\hat{Y}^*) + (\hat{\theta}^0_{AMR} - \theta). \tag{A.1}$$

Apply Lemma A.1.2 to $\psi(z;\hat{\xi}) = \epsilon^*(\hat{y}^*)\,\hat{y}^*$. Under strict overlap, $w^{*0} = \mathbb{E}[\hat{h} \mid \hat{Y}^*]$ is bounded by $\|\hat{h}\|_\infty$, and $\hat{w}^*$ is bounded by Assumption 3.4.2; hence $|\epsilon^*| \le 2H$ for some constant $H < \infty$ with probability tending to one, so $(\epsilon^*)^2 \le 2H\,|\epsilon^*|$ pointwise on this event. Combining this with Cauchy–Schwarz,

$$\mathbb{E}\big[\psi(W;\hat{\xi}^{(-k)})^2 \mid \mathcal{F}_{-k}\big] \le 2H\,\mathbb{E}\big[|\epsilon^*|(\hat{Y}^*)^2 \mid \mathcal{F}_{-k}\big] \le 2H\,\left\|\epsilon^{*,(-k)}\right\|_2 \sqrt{\mathbb{E}\big[(\hat{Y}^*)^4 \mid \mathcal{F}_{-k}\big]} = o_P(1),$$

using $\|\epsilon^{*,(-k)}\|_2 = o_P(1)$ from Assumption 3.4.2 and $\mathbb{E}\big[(\hat{Y}^*)^4 \mid \mathcal{F}_{-k}\big] = O_P(1)$ from Assumption 3.4.1. Hence $P_n(\epsilon^*(\hat{Y}^*)\hat{Y}^*) = B_n + o_P(n^{-1/2})$, and combining with (A.1) and Theorem 3.4.4 yields the stated convergence. □

A.1.8 Discussion of Proposition 3.4.7

Proof. By Proposition 3.4.6 and the asymptotic linearity $\sqrt{n}\,(\hat{\theta}^0_{AMR} - \theta) = \sqrt{n}\,P_n[\varphi_{AMR}(W;\xi)] + o_P(1)$ from Appendix A.1.6,

$$\sqrt{n}\,(\hat{\theta}_{AMR} - \theta) = \sqrt{n}\,P_n[\varphi_{AMR}(W;\xi)] + \sqrt{n}\,B_n + o_P(1).$$

Thus we require $\sqrt{n}\, B_n = o_P(1)$. The limiting distribution follows from the central limit theorem applied to $\sqrt{n}\, P_n[\varphi_{AMR}(W;\xi)]$. □

By Cauchy–Schwarz, we obtain $|B_n| \le \|\hat{w}^* - w^{*0}\|_2 \, \|\hat{Y}^*\|_2$, with $\|\hat{Y}^*\|_2$ bounded under Assumption 3.4.1; hence the rate $\|\hat{w}^* - w^{*0}\|_2 = o_P(n^{-1/2})$ is sufficient for $\sqrt{n}\, B_n = o_P(1)$. This is too strong for generic nonparametric estimation of w^{*0} and is therefore not a useful general sufficient condition.

A.1.9 Proof of Proposition 3.4.8

Proof. By Proposition 3.4.6,

$$\sqrt{n}\,(\hat{\theta}_{AMR} - \theta) = \sqrt{n}\, B_n + \sigma_0 G_n + o_P(1), \qquad G_n \rightsquigarrow N(0,1),$$

and $\hat{\sigma}'_{AMR} \xrightarrow{p} \sigma_X$ by the law of large numbers together with the consistency of the cross-fitted nuisances. Set $E_n := \{|G_n| \le q_{1-\alpha/2}\} \cap \{\sqrt{n}\,|B_n| \le q_{1-\alpha/2}\nu - \delta\}$. On E_n,

$$\begin{aligned}|\sqrt{n}\,(\hat{\theta}_{AMR} - \theta)| &\le \sigma_0|G_n| + \sqrt{n}\,|B_n| + o_P(1) \\ &\le q_{1-\alpha/2}(\sigma_0 + \nu) - \delta + o_P(1) = q_{1-\alpha/2}\,\hat{\sigma}'_{AMR} - \delta + o_P(1),\end{aligned}$$

using $\sigma_0 + \nu = \sigma_X$ and $\hat{\sigma}'_{AMR} = \sigma_X + o_P(1)$. Since $\delta > 0$ is fixed, $E_n \subseteq \{|\hat{\theta}_{AMR} - \theta| \le q_{1-\alpha/2}\,\hat{\sigma}'_{AMR}/\sqrt{n}\}$ with probability tending to one. By assumption (3.12) and $G_n \rightsquigarrow N(0,1)$,

$$\liminf_{n\to\infty} \Pr(E_n) \ge \Pr(|G| \le q_{1-\alpha/2}) = 1 - \alpha,$$

and the same lower bound applies to the coverage probability. □

A.1.10 Proofs of Theorem 3.4.9 and Corollary 3.4.10

We use the following convention. A mean-zero function $\varphi \in L_2(P)$ is an influence function for θ relative to a regular submodel $\mathcal{M}'$ if, for every regular path $P_\varepsilon \subset \mathcal{M}'$ through P with score s,

$$\frac{\partial}{\partial \varepsilon}\theta(P_\varepsilon)\Big|_{\varepsilon=0} = \mathbb{E}[\varphi(W)s(W)].$$

Proof of Theorem 3.4.9. Let $P_\varepsilon \subset \mathcal{M}'$ be a regular path through P with score $s \in \mathcal{T}(P, \mathcal{M}')$. Since φ_{AIPW} is the EIF in the full model $\mathcal{M}$,

$$\frac{\partial}{\partial \varepsilon}\theta(P_\varepsilon)\Big|_{\varepsilon=0} = \mathbb{E}[\varphi_{AIPW}(W;\xi)s(W)].$$

Thus φ_{AMR} is an influence function relative to $\mathcal{M}'$ if and only if, for every such score s,

$$\mathbb{E}[\varphi_{AMR}(W;\xi)s(W)] = \mathbb{E}[\varphi_{AIPW}(W;\xi)s(W)].$$

This is equivalent to

$$\mathbb{E}[r(W)s(W)] = 0, \qquad \forall s \in \mathcal{T}(P, \mathcal{M}'),$$

which proves the result. □

Proof of Corollary 3.4.10. First, $\mathbb{E}[\varphi_{AMR}(W;\xi)] = 0$ by the bridge identity, and strict overlap gives $|w^*(Y^*)| \leq \|h\|_\infty < \infty$. Together with $\mathbb{E}[(Y^*)^2] < \infty$, this implies $\varphi_{AMR} \in L_2^0(P)$.

Let $s(W) = f(Y^*) \in \mathcal{T}^*(P)$. By the definition $w^*(Y^*) = \mathbb{E}[h(X,Z) \,|\, Y^*]$,

$$\begin{aligned}\mathbb{E}[r(W)s(W)] &= \mathbb{E}\Big[Y^* f(Y^*)\mathbb{E}\{h(X,Z) - w^*(Y^*) \,|\, Y^*\}\Big] \\ &= 0.\end{aligned}$$

Therefore φ_{AMR} is an influence function for θ relative to $\mathcal{M}^*$ by Theorem 3.4.9.

It remains to identify the EIF. The EIF of a submodel is the $L_2(P)$ orthogonal projection of the full-model EIF onto its tangent space. The tangent space of $\mathcal{M}^*$ is $\mathcal{T}^*(P)$, the mean-zero functions of Y^*, so this projection is the conditional expectation given Y^*:

$$\begin{aligned}\Pi_{\mathcal{T}^*(P)}\varphi_{AIPW} &= \mathbb{E}[\varphi_{AIPW}(W;\xi) \,|\, Y^*] \\ &= \mathbb{E}[h(X,Z) \,|\, Y^*]Y^* - \theta \\ &= w^*(Y^*)Y^* - \theta \\ &= \varphi_{AMR}(W;\xi).\end{aligned}$$

Thus φ_{AMR} is the EIF in $\mathcal{M}^*$. □

A.2 Weights estimation

We compare three univariate regression methods for estimating $\hat{w}$ and $\hat{w}^*$ by minimising the L_2 losses (3.7) and (3.8).

Nadaraya–Watson regression (Nadaraya, 1964; Watson, 1964). In our experiments, $\hat{\theta}_{AMR}$ and $\hat{\theta}_{MR}$ have small bias but high variance; the variance is sensitive to the bandwidth choice, particularly near the boundary in regions where data are sparse (Wasserman, 2006). Increasing the kernel order reduces bias but inflates variance.

Kernel ridge regression (Saunders et al., 1998). For training data $\{(x_i, y_i)\}_{i=1}^n$, a reproducing-kernel Hilbert space $\mathcal{H}$ with kernel k, and ridge parameter $\lambda > 0$, KRR estimates the regression function by

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

which by the representer theorem admits the closed form $\hat{f}(x) = \mathbf{k}(x)^\top (K + n\lambda I)^{-1} \mathbf{y}$, where $K_{ij} = k(x_i, x_j)$, $\mathbf{k}(x) = \{k(x, x_i)\}_{i=1}^n$, and $\mathbf{y} = (y_1, \ldots, y_n)^\top$; we take k to be the RBF kernel $k(x, x') = \exp(-\|x - x'\|^2 / 2\sigma^2)$. In our experiments, KRR gave the smallest MSE among the three weight-regression methods considered, especially at small sample sizes. The ridge term introduces bias but stabilises the projection estimate; with a regularisation parameter shrinking at an appropriate rate, KRR is asymptotically unbiased and converges to the true function whenever the latter lies in the RKHS of the kernel.

Nonparametric kernel-based local polynomial regression (Calonico et al., 2018). Implemented with the package `nprobust` (Calonico et al., 2019). In our experiments this method has lower bias than the other two but higher variance, and hence a higher MSE of $\hat{\theta}_{AMR}$ than KRR or $\hat{\theta}_{AIPW}$. Each prediction point requires its own weighted least squares fit, which makes the method sensitive to noise and outliers and computationally costly.

A.3 Derivation of weights in Example 3.3.4

In this section we provide the derivation of weights $w(y)$, $w^0(y)$, $w^*(y^*)$ and $w^{*0}(y^*)$ in Example 3.3.4. As demonstrated in Example 3.3.4, we assume $Y(x)|Z \sim \mathcal{N}(\mu^x(Z), \sigma^2)$, $x \in \{0, 1\}$,

and $X|Z \sim \textit{Bernoulli}(\pi(Z))$. Further, let $\mathcal{N}(t; m, s^2)$ denote the density of t under a Gaussian distribution with mean m and variance s^2. We have:

$$\begin{aligned} Y &= XY(1) + (1-X)Y(0), \\ Y(x)|Z &\sim N(\mu^x(Z), \sigma^2). \end{aligned}$$

We also have:

$$\begin{aligned} p_{XZ|Y}(x,z|y) &= \frac{p_{XZY}(x,z,y)}{p_Y(y)} \\ &= \frac{p_{Y|XZ}(y|x,z)\, p_{X|Z}(x|z)\, p_Z(z)}{\sum_{x\in\{0,1\}} \int_{\mathcal{Z}} p_{Y|XZ}(y|x,z)\, p_{X|Z}(x|z)\, p_Z(z)\,\mathrm{d}z}; \\ p_{XZ|Y}(1,z|y) &= \pi(z)\,\mathcal{N}(y;\mu^1(z),\sigma^2)\, p_Z(z)/p_Y(y) \\ p_{XZ|Y}(0,z|y) &= \{1-\pi(z)\}\,\mathcal{N}(y;\mu^0(z),\sigma^2)\, p_Z(z)/p_Y(y) \\ p_Y(y) &= \sum_{x\in\{0,1\}} \int_{\mathcal{Z}} p_{Y|X,Z}(y|x,z)\, p_{X|Z}(x|z) p_Z(z)\,\mathrm{d}z \\ &= \int_{\mathcal{Z}} \Big[\pi(z)\mathcal{N}(y;\mu^1(z),\sigma^2) + \{1-\pi(z)\}\mathcal{N}(y;\mu^0(z),\sigma^2)\Big] p_Z(z)\,\mathrm{d}z \\ &= \mathbb{E}_Z\Big[\pi(Z)\mathcal{N}(y;\mu^1(Z),\sigma^2) + \{1-\pi(Z)\}\mathcal{N}(y;\mu^0(Z),\sigma^2)\Big]. \end{aligned}$$

We then find:

$$\begin{aligned} w^0(y) &= \mathbb{E}\left[\frac{X-\hat\pi(Z)}{\hat\pi(Z)\{1-\hat\pi(Z)\}}\,\middle|\, Y=y\right] \\ &= \sum_{x\in\{0,1\}} \int_{\mathcal{Z}} \frac{x-\hat\pi(z)}{\hat\pi(z)\{1-\hat\pi(z)\}} p_{XZ|Y}(x,z|y)\,\mathrm{d}z \\ &= \int_{\mathcal{Z}} \left\{\frac{1}{\hat\pi(z)} p_{XZ|Y}(1,z|y) - \frac{1}{1-\hat\pi(z)} p_{XZ|Y}(0,z|y)\right\}\mathrm{d}z \\ &= \frac{1}{p_Y(y)} \int_{\mathcal{Z}} \left\{\frac{\pi(z)}{\hat\pi(z)}\mathcal{N}(y;\mu^1(z),\sigma^2) - \frac{1-\pi(z)}{1-\hat\pi(z)}\mathcal{N}(y;\mu^0(z),\sigma^2)\right\} p_Z(z)\,\mathrm{d}z \\ &= \frac{\mathbb{E}_Z\left[\frac{\pi(Z)}{\hat\pi(Z)}\mathcal{N}(y;\mu^1(Z),\sigma^2) - \frac{1-\pi(Z)}{1-\hat\pi(Z)}\mathcal{N}(y;\mu^0(Z),\sigma^2)\right]}{\mathbb{E}_Z\Big[\pi(Z)\mathcal{N}(y;\mu^1(Z),\sigma^2) + \{1-\pi(Z)\}\mathcal{N}(y;\mu^0(Z),\sigma^2)\Big]}. \end{aligned}$$

When $\hat\pi = \pi$, the displayed expression for $w^0(y)$ reduces to

$$w(y) = \frac{\mathbb{E}_Z[\mathcal{N}(y;\mu^1(Z),\sigma^2) - \mathcal{N}(y;\mu^0(Z),\sigma^2)]}{\mathbb{E}_Z\Big[\pi(Z)\mathcal{N}(y;\mu^1(Z),\sigma^2) + \{1-\pi(Z)\}\mathcal{N}(y;\mu^0(Z),\sigma^2)\Big]},$$

that is, $w^0(y)$ coincides with $w(y)$ under oracle nuisance estimation.

Similarly, we provide the derivation of AMR weights. With $\hat{Y}^* = Y - \hat{\mu}^*(Z)$, we have:

$$\begin{aligned}
&\hat{Y}^* = XY(1) + (1-X)Y(0) - \Big[\hat{\pi}(Z)\hat{\mu}^0(Z) + \{1-\hat{\pi}(Z)\}\hat{\mu}^1(Z)\Big] \\
&\hat{Y}^*|X=1, Z \sim \mathcal{N}(\mu^1(Z) - \hat{\mu}^*(Z), \sigma^2) \\
&\hat{Y}^*|X=0, Z \sim \mathcal{N}(\mu^0(Z) - \hat{\mu}^*(Z), \sigma^2).
\end{aligned}$$

We also have:

$$\begin{aligned}
p_{XZ|\hat{Y}^*}(x,z|y^*) &= \frac{p_{XZ\hat{Y}^*}(x,z,y^*)}{p_{\hat{Y}^*}(y^*)} \\
&= \frac{p_{\hat{Y}^*|XZ}(y^*|x,z)\, p_{X|Z}(x|z)\, p_Z(z)}{\sum_{x\in\{0,1\}} \int_{\mathcal{Z}} p_{\hat{Y}^*|XZ}(y^*|x,z)\, p_{X|Z}(x|z)\, p_Z(z)\,\mathrm{d}z} \\
p_{XZ|\hat{Y}^*}(1,z|y^*) &= \pi(z)\mathcal{N}(y^*; \mu^1(z) - \hat{\mu}^*(z), \sigma^2)\, p_Z(z)/p_{\hat{Y}^*}(y^*) \\
p_{XZ|\hat{Y}^*}(0,z|y^*) &= \{1-\pi(z)\}\mathcal{N}(y^*; \mu^0(z) - \hat{\mu}^*(z), \sigma^2)\, p_Z(z)/p_{\hat{Y}^*}(y^*),
\end{aligned}$$

where

$$\begin{aligned}
p_{\hat{Y}^*}(y^*) &= \sum_{x\in\{0,1\}} \int_{\mathcal{Z}} p_{\hat{Y}^*|XZ}(y^*|x,z)\, p_{X|Z}(x|z)\, p_Z(z)\,\mathrm{d}z \\
&= \int_{\mathcal{Z}} \Big[\pi(z)\mathcal{N}(y^*; \mu^1(z) - \hat{\mu}^*(z), \sigma^2) + \{1-\pi(z)\}\mathcal{N}(y^*; \mu^0(z) - \hat{\mu}^*(z), \sigma^2)\Big]\, p_Z(z)\mathrm{d}z \\
&= \mathbb{E}_Z\Big[\pi(Z)\mathcal{N}(y^*; \mu^1(Z) - \hat{\mu}^*(Z), \sigma^2) + \{1-\pi(Z)\}\mathcal{N}(y^*; \mu^0(Z) - \hat{\mu}^*(Z), \sigma^2)\Big].
\end{aligned}$$

We thus obtain:

$$\begin{aligned}
w^{*0}(y^*) &= \mathbb{E}\left[\frac{X - \hat{\pi}(Z)}{\hat{\pi}(Z)\big\{1-\hat{\pi}(Z)\big\}} \,\middle|\, \hat{Y}^* = y^*\right] \\
&= \sum_{x\in\{0,1\}} \int_{\mathcal{Z}} \frac{x - \hat{\pi}(z)}{\hat{\pi}(z)\big\{1-\hat{\pi}(z)\big\}} p_{XZ|\hat{Y}^*}\big(x,z|y^*\big)\,\mathrm{d}z \\
&= \int_{\mathcal{Z}} \left[\frac{1}{\hat{\pi}(z)} p_{XZ|\hat{Y}^*}\big(1,z|y^*\big) - \frac{1}{1-\hat{\pi}(z)} p_{XZ|\hat{Y}^*}\big(0,z|y^*\big)\right]\mathrm{d}z \\
&= \frac{\mathbb{E}_Z\left[\frac{\pi(Z)}{\hat{\pi}(Z)}\mathcal{N}\big(y^*; \mu^1(Z) - \hat{\mu}^*(Z), \sigma^2\big) - \frac{1-\pi(Z)}{1-\hat{\pi}(Z)}\mathcal{N}\big(y^*; \mu^0(Z) - \hat{\mu}^*(Z), \sigma^2\big)\right]}{\mathbb{E}_Z\Big[\pi(Z)\mathcal{N}\big(y^*; \mu^1(Z) - \hat{\mu}^*(Z), \sigma^2\big) + \big\{1-\pi(Z)\big\}\mathcal{N}\big(y^*; \mu^0(Z) - \hat{\mu}^*(Z), \sigma^2\big)\Big]}.
\end{aligned}$$

When we have oracle estimators $\hat{\pi} = \pi, \hat{\mu}^x = \mu^x$ for $x \in \{0,1\}$, we find

$$w^*(y^*) = \frac{\mathbb{E}_Z\Big[\mathcal{N}(y^*; \pi(Z)\tau(Z), \sigma^2) - \mathcal{N}(y^*; -\{1-\pi(Z)\}\tau(Z), \sigma^2)\Big]}{\mathbb{E}_Z\Big[\pi(Z)\mathcal{N}(y^*; \pi(Z)\tau(Z), \sigma^2) + \{1-\pi(Z)\}\mathcal{N}(y^*; -\{1-\pi(Z)\}\tau(Z), \sigma^2)\Big]},$$

where $\tau(z) = \mu^1(z) - \mu^0(z)$.

A.4 Implementation details

For AMR and MR, the synthetic-study nuisances are estimated as follows: the propensity score by logistic regression; the potential-outcome model by a feedforward network with two hidden layers of widths 64 and 32 and ReLU activations, trained for 100 epochs with Adam (learning rate 10^{-3}, batch size 64); and the weight function $\hat{w}^*$ by kernel ridge regression with an RBF kernel, its ridge parameter and bandwidth selected by five-fold cross-validation. The comparators CBPS, DOPE, TMLE and CTMLE are run through their authors' implementations with the default settings and their own built-in tuning. Configurations for the remaining experiments and the full code are available at https://github.com/linyingyang/AMR.

A.5 Detailed synthetic experiment results

The detailed bias, MAE and RMSE values summarised in Figures 3.2 and 3.3 are in Tables A.1–A.3 across $n \in \{400, 600, 1000\}$ and $p_I \in \{1, 5, 10, 15, 20\}$.

A.6 Additional experiment: Amazon review

In this section, we compare the performance of our method with the Treatment Ignorant (TI) estimator proposed by Gui and Veitch (2022) on the publicly available Amazon review dataset. We mainly use reviews for mp3, CD and vinyl, setting the treatment X as whether the review is five stars ($X = 1$) or one/two stars ($X = 0$). We use the same semi-synthetic data generation setting as created in Gui and Veitch (2022) to simulate outcome Y, that is, we additionally set $C = 1$ if the product is a CD and $C = 0$ otherwise, $Y = \beta_x X + \beta_c(\pi(C) - \beta_o) + \gamma\epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$. Here β_x, β_c, β_o stand for the average treatment effect, the strength of confounding and the offset term for $\mathbb{E}\pi(C)$, where $\pi(C) = \Pr(X = 1 \,|\, C)$ is the propensity score. The dataset has 10,658 entries in total.

The TI estimator addresses the challenges of insufficient overlap in causal effect estimation

n	p_I	TMLE	CTMLE	CBPS	IPW	AIPW	DOPE	MR	AMR
400	1	−0.795	−2.01	−1.80	−3.25	−1.99	103	−3.23	−3.32
	5	0.121	1.15	1.33	14.9	10.9	103	−0.380	−3.86
	10	0.823	0.754	0.381	2.23	2.52	103	−1.45	−3.11
	15	−0.862	−0.55	−1.12	14.6	7.14	103	−3.82	−3.91
	20	−4.56	−0.601	−3.21	−156	−165	104	−2.19	−4.61
600	1	−0.305	−0.501	−0.322	0.118	−0.309	103	−0.396	−1.25
	5	−1.40	−2.65	−1.53	−5.50	−1.88	102	−2.56	−1.87
	10	0.435	0.630	−0.0821	6.30	2.83	104	−1.74	−1.52
	15	−0.654	−0.331	−0.295	−1.84	−0.904	103	−0.732	−2.91
	20	−2.01	−1.18	−0.131	4.60	4.44	102	−0.960	−2.70
1000	1	0.263	0.0283	0.242	0.878	0.282	102	0.614	0.675
	5	−0.924	−0.209	−1.33	−3.27	−0.653	103	−1.64	−0.331
	10	0.813	1.17	1.36	13.0	−1.70	103	1.90	−0.795
	15	−0.711	−0.958	0.0696	−2.84	5.37	103	−2.49	0.153
	20	1.98	1.85	1.17	3.72	2.39	103	−2.43	0.289

Table A.1: Bias on synthetic data.

n	p_I	TMLE	CTMLE	CBPS	IPW	AIPW	DOPE	MR	AMR
400	1	14.5	11.0	12.0	19.2	17.1	103	11.3	6.20
	5	16.2	12.3	12.0	43.1	35.0	103	13.2	6.58
	10	17.6	12.3	13.6	29.1	26.3	103	11.1	6.29
	15	19.6	14.9	14.9	58.8	40.5	103	10.8	6.28
	20	18.8	13.2	14.7	207	215	104	9.15	7.39
600	1	11.3	9.06	10.6	14.9	11.9	103	9.56	8.25
	5	14.3	8.68	11.8	24.5	18.1	102	13.1	8.14
	10	14.5	11.7	12.6	35.5	25.8	104	13.4	8.62
	15	15.4	9.30	12.3	54.1	40.5	103	12.5	7.31
	20	15.6	9.64	13.3	33.5	31.5	102	9.43	7.63
1000	1	5.70	9.01	8.42	11.7	6.08	102	8.17	4.68
	5	6.93	6.19	8.93	18.3	10.2	103	11.5	5.99
	10	10.2	9.95	10.4	28.4	19.7	103	14.1	9.96
	15	10.4	7.63	12.1	33.3	24.1	103	13.5	9.36
	20	9.64	10.1	10.3	20.0	17.7	103	11.0	8.54

Table A.2: MAE on synthetic data.

on text data, by estimating propensity scores conditional on the potential outcomes, i.e. $\hat{\pi} = \mathbb{E}[X \mid \hat{\mu}^0, \hat{\mu}^1]$. Specifically, the authors propose a transformer-based neural network architecture similar to the DragonNet, that first learns the potential outcomes using the propensity score estimation information by loss function design, and then finalises the propensity score estimation

n	p_I	TMLE	CTMLE	CBPS	IPW	AIPW	DOPE	MR	AMR
400	1	19.1	14.2	15.4	29.8	26.0	103	14.8	8.06
	5	20.7	16.9	15.2	195	151	103	22.6	10.4
	10	22.4	16.2	17.4	44.9	40.3	103	14.5	8.28
	15	25.4	24.2	18.6	164	113	104	15.7	7.97
	20	25.2	21.7	18.7	2190	1860	104	11.8	9.91
600	1	14.3	11.4	12.9	19.8	16.0	103	12.0	10.5
	5	18.1	11.0	14.7	33.9	23.9	103	19.1	10.4
	10	19.0	16.7	16.4	92.3	53.2	104	18.5	12.5
	15	19.8	12.4	15.5	176	110	103	18.3	9.05
	20	19.7	13.5	17.1	68.1	65.3	103	13.9	9.89
1000	1	7.12	11.5	10.6	17.0	7.90	102	10.4	5.95
	5	8.98	7.90	11.2	31.0	16.1	103	15.4	7.78
	10	13.1	15.7	13.8	87.8	42.2	103	20.7	14.4
	15	13.6	9.87	15.8	63.1	45.6	103	26.1	13.4
	20	12.6	13.0	12.8	27.7	26.5	103	14.3	12.6

Table A.3: RMSE on synthetic data.

	TI	**IPW**	**AIPW**	**MR**	**AMR**
Bias	−0.0300	0.622	−0.0634	−0.231	**0.0201**
MAE	0.192	1.12	0.901	0.252	**0.182**
RMSE	0.252	2.45	1.55	0.312	**0.251**

Table A.4: Estimator performances on $R = 50$ replicates of *Amazon Reviews* dataset, $n = 1000$, default configuration, $\theta = 2$.

using the estimated potential outcomes as features. This approach closely resembles Benkeser et al. (2020) but is tailored for text-based applications in that the TI estimator is a transformer, attention-based architecture.

To ensure a fair comparison with the TI estimator, we used BERT to produce embeddings as covariates for each Amazon review. Specifically, we pooled token embeddings for each review by using the `[CLS]` embedding. These embeddings were then fed into a 3-layer feedforward neural network model to estimate $\hat{\pi}$. For $\hat{\mu}^x, x \in \{0, 1\}$, we keep them the same as those used in the TI estimator.

We first provide results in comparison with TI estimator under its recommended configurations in model building and data simulation ($\beta_x = 2, \beta_c = 50, \pi(0) = 0.8, \pi(1) = 0.6$) in Figure A.1 and Table A.4. We randomly sampled 1000 data points from the dataset.

The main limitation of the TI estimator is apparent: it hinges on accurate estimation of

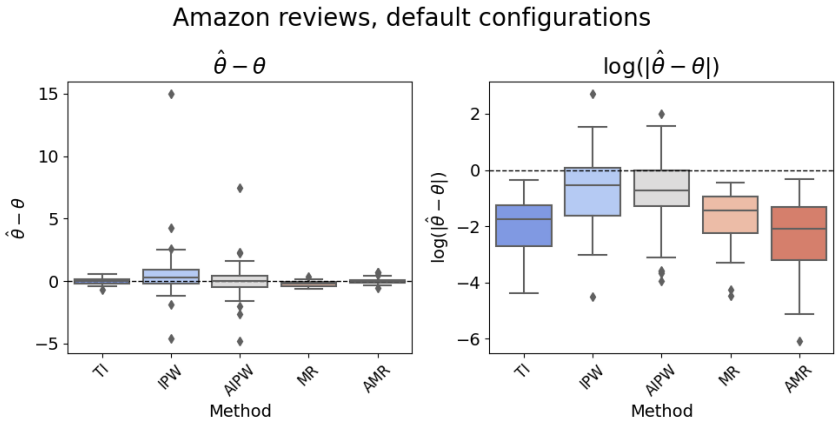


Figure A.1: Estimator performances on $R = 50$ replicates of *Amazon Reviews* dataset, $n = 1000$, default configuration.

	TI	IPW	AIPW	MR	AMR
Bias	0.360	−13.5	−0.150	1.82	**0.000112**
MAE	1.05	16.7	16.3	1.87	**0.723**
RMSE	1.92	30.5	27.3	1.92	**1.06**

Table A.5: Estimator performances on $R = 50$ replicates of *Amazon Reviews* dataset, $n = 1000$, data simulation changed, $\theta = 0.2$.

the potential outcomes. Under the default configuration recommended by the authors, these estimates are very close to the true values, which explains the strong performance of TI. We present results in two modifications of the simulation: an extra interaction term in the true potential outcome (Table A.5), and replacement of the transformer-based architecture in TI with linear regression for potential outcome estimation (Table A.6). TI's performance degrades in both, consistent with its dependence on accurate outcome estimation. MR is more stable than TI in both modifications; AMR is competitive with TI in the interaction-term setting but degrades when the linear potential-outcome model is used. The results are consistent with the robustness motivation but indicate that AMR can be sensitive to the quality of the outcome residual when $\hat{\mu}^x$ is severely misspecified.

	TI	IPW	AIPW	MR	AMR
Bias	-2.58×10^6	0.372	5.23×10^7	**-0.221**	0.803
MAE	2.58×10^6	0.791	6.31×10^7	**0.242**	2.53
RMSE	1.13×10^7	1.37	3.75×10^7	**0.283**	9.43

Table A.6: Estimator performances on $R = 50$ replicates of *Amazon Reviews* dataset, $n = 1000$, linear model for potential outcome, $\theta = 3$.

Appendix B

Supplementary material for Chapter 4

B.1 Auxiliary results

This section collects auxiliary results that support the main theorems.

B.1.1 Consequences of a fixed pooling ratio

The pooled sample $Q_f^{m+\ell}$ is formed by combining two mutually independent samples Q_c^m and Q_h^ℓ, which are i.i.d. samples from Q_c and Q_h, respectively. However, the pooled sample $Q_f^{m+\ell}$ is *not* an i.i.d. sample from the mixture distribution $\frac{m}{m+\ell}Q_c + \frac{\ell}{m+\ell}Q_h$. This is because the sample sizes m and ℓ are *deterministic*, whereas under i.i.d. sampling from the mixture, the number of observations drawn from each component would be random and follow a binomial distribution. In fact, $Q_f^{m+\ell}$ is *exchangeable*, meaning that the joint distribution of $Q_f^{m+\ell}$ remains unchanged under permutations.

Despite this distinction, this technical subtlety does not significantly complicate our analysis. For example, the kernel mean embedding of $Q_f^{m+\ell}$ coincides exactly with that of an i.i.d. sample from the mixture $\frac{m}{m+\ell}Q_c + \frac{\ell}{m+\ell}Q_h$, as shown in the following result.

Lemma B.1.1. *Let $\{X_i\}_{i=1}^\infty$ and $\{Y_i\}_{i=1}^\infty$ be sequences of independent random variables drawn from Q_1, Q_2. For any positive integers n_1, n_2, define $n = n_1 + n_2$, and let Q^n be the empirical measure based on $\{Z_i\}_{i=1}^n$, where $Z_i = X_i$ if $1 \le i \le n_1$ and $Z_i = Y_i$ if $n_1 < i \le n$. Moreover, let $Q_1^{n_1}$ and $Q_2^{n_2}$ be the empirical samples based on $\{X_i\}_{i=1}^{n_1}, \{Y_i\}_{i=1}^{n_2}$, respectively, and define the mixture distribution $Q = \omega Q_1 + (1-\omega)Q_2$. Then the kernel mean embedding of Q^n takes the form*

$$\hat{\xi}_{Q^n} = \frac{n_1}{n}\mathbb{E}_{X\sim Q_1^{n_1}} k(\cdot, X) + \frac{n_2}{n}\mathbb{E}_{Y\sim Q_2^{n_2}} k(\cdot, Y) = \frac{n_1}{n}\hat{\xi}_{n_1} + \frac{n_2}{n}\hat{\xi}_{n_2}.$$

Since $\hat{\xi}_{Q^n}$ can be decomposed into the kernel mean embeddings based on $Q_1^{n_1}$ and $Q_2^{n_2}$, and since $Q_1^{n_1}$ and $Q_2^{n_2}$ are mutually independent, the asymptotic properties of $\hat{\xi}_{Q^n}$ can be readily derived from classic results on kernel mean embeddings of independent samples. We will leverage these observations in most of our proofs, such as Appendix B.2.2.

Proof of Lemma B.1.1. This follows immediately from the definition of kernel mean embeddings of empirical samples:

$$\begin{aligned}\hat{\xi}_{Q^n} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, Z_i) = \frac{1}{n}\sum_{i=1}^{n_1} k(\cdot, X_i) + \frac{1}{n}\sum_{i=n_1+1}^{n} k(\cdot, Y_i) = \frac{n_1}{n}\mathbb{E}_{X\sim Q_1^{n_1}} k(\cdot, X) + \frac{n_2}{n}\mathbb{E}_{Y\sim Q_2^{n_2}} k(\cdot, Y) \\ = \frac{n_1}{n}\hat{\xi}_{n_1} + \frac{n_2}{n}\hat{\xi}_{n_2}.\end{aligned}$$

□

B.1.2 Validity with merged samples

This section contains some auxiliary results that will be used to show the validity of the overall procedure in Corollary 4.3.17; to be more specific, we show that (4.12) is met. These results are extensions of Theorem 4.3.5 and Theorem 4.3.8 to the case where we condition on the event that the fusion test, $\mathcal{H}_0^f$, is rejected, namely that historical and current controls are merged. We present the results for both partial bootstrap (Appendix B.1.2.1) and partial permutation (Appendix B.1.2.2).

B.1.2.1 Partial bootstrap

We first show that the same weak limit in Proposition 4.3.4 holds conditional on the event that the fusion test is rejected.

Proposition B.1.2 (Weak convergence of the partial bootstrap provided merged controls under non-identical control groups)**.** *Assume Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, that Assumption 4.3.2 is satisfied, and that $\sigma_c^2 > 0$. Let $\mathcal{A}_n = \{\phi_f(Q_c^m, Q_h^\ell) = 1\}$ denote the event of rejecting the fusion null, and assume $\lim_{n\to\infty} \Pr(\mathcal{A}_n) = 1$. Under the causal null $\mathcal{H}_0 : Q_c = Q_t$, the following holds conditionally on $\mathcal{A}_n$:*

$$\Delta \rightsquigarrow \mathcal{N}(0, 4(1 + 1/c_1)\sigma_c^2).$$

Moreover,

$$\sup_{z\in\mathbb{R}}\Big|\Pr^*\{\Delta^* \le z\} - \Pr\{\Delta \le z\}\Big| \xrightarrow{a.s.} 0.$$

Proof. The proof of Proposition 4.3.4 (present in Appendix B.2.2) gives, under $\sigma_c^2 > 0$, the unconditional weak convergence $\Delta \rightsquigarrow \mathcal{N}(0, 4(1+1/c_1)\sigma_c^2)$ and the bootstrap consistency. Its argument does not transfer directly under conditioning, since $A_n = (1-\gamma)(\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m})$ is no longer a weighted average of independent samples once we condition on the event $\mathcal{A}_n = \{\phi_f(Q_c^m, Q_h^\ell) = 1\}$. We instead use that $\Pr(\mathcal{A}_n) \to 1$: conditioning on an event of probability tending to one leaves a weak limit unchanged. For any continuity point z,

$$\Big|\Pr\{\Delta \le z \,|\, \mathcal{A}_n\} - \Pr\{\Delta \le z\}\Big| \le \frac{2\Pr(\mathcal{A}_n^c)}{\Pr(\mathcal{A}_n)} \to 0,$$

so the conditional law of Δ has the same weak limit, and the bootstrap consistency carries over likewise. □

Using this result, we can then show a conditional version of the partial bootstrap validity.

Theorem B.1.3 (Partial bootstrap validity provided merged controls under non-identical control groups)**.** *Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, and Assumption 4.3.2 is satisfied. Let $\hat{q}_{1-\alpha,B}^{m+\ell,n}$ be the partial bootstrap critical value defined in* (4.9)*. Moreover, let $\mathcal{A}_n = \{\phi_f(Q_c^m, Q_h^\ell) = 1\}$ denote the event of rejecting the fusion null, and assume $\sigma_c^2 > 0$ and $D(Q_c, Q_h) < \rho$, so that $\tau := \lim_{n\to\infty}\Pr(\mathcal{A}_n) = 1$. Under the causal null hypothesis $\mathcal{H}_0 : Q_c = Q_t$, the test, conditionally on merging, has asymptotic level α:*

$$\limsup_{n\to\infty}\lim_{B\to\infty}\Pr\Big\{\Delta > \hat{q}_{1-\alpha,B}^{m+\ell,n}\,\Big|\,\mathcal{H}_0, \mathcal{A}_n\Big\} \le \alpha.$$

Proof. The claim holds by following the same proof in Theorem 4.3.5 and using Proposition B.1.2 instead of Proposition 4.3.4. □

B.1.2.2 Partial permutation

The following result shows that the partial permutation also remains valid conditionally on the event that the fusion test is rejected.

Theorem B.1.4 (Partial permutation validity provided merged controls)**.** *Suppose Assumption 4.2.2 holds under $R = Q_c, Q_h, Q_t$, and Assumption 4.3.2 is satisfied. Let $\mathcal{A}_n = \{\phi_f(Q_c^m, Q_h^\ell) =$*

1} *denote the event of rejecting the fusion null, and assume* $D(Q_c, Q_h) < \rho$, *so that* $\tau \coloneqq \lim_{n\to\infty} \Pr(\mathcal{A}_n) = 1$. *Under the null hypothesis* $H_0 : Q_c = Q_t$,

$$\limsup_{n\to\infty} \Pr\Big\{T_{n,m,\ell} \geq \hat{c}^{m+\ell,n}_{1-\alpha,B} \,|\, \mathcal{A}_n\Big\} \leq \alpha.$$

Proof. Since $D(Q_c, Q_h) < \rho$, the MMD equivalence fusion test rejects the fusion null with probability tending to one (Theorem 4.3.1), so $\tau \coloneqq \lim_{n\to\infty} \Pr(\mathcal{A}_n) = 1$ and $\Pr(\cdot \,|\, \mathcal{A}_n)$ is well defined for all large n. By the unconditional validity of the partial permutation test (Theorem 4.3.8), $\Pr\{T_{n,m,\ell} \geq \hat{c}^{m+\ell,n}_{1-\alpha,B}\} \leq \alpha$ for every n, so

$$\begin{aligned}
\Pr\Big\{T_{n,m,\ell} \geq \hat{c}^{m+\ell,n}_{1-\alpha,B} \,|\, \mathcal{A}_n\Big\} &= \frac{\Pr\Big(\{T_{n,m,\ell} \geq \hat{c}^{m+\ell,n}_{1-\alpha,B}\} \cap \mathcal{A}_n\Big)}{\Pr(\mathcal{A}_n)} \\
&\leq \frac{\Pr\{T_{n,m,\ell} \geq \hat{c}^{m+\ell,n}_{1-\alpha,B}\}}{\Pr(\mathcal{A}_n)} \\
&\to \frac{\limsup_{n\to\infty} \Pr\{T_{n,m,\ell} \geq \hat{c}^{m+\ell,n}_{1-\alpha,B}\}}{\tau} \leq \alpha,
\end{aligned}$$

where the last step uses $\tau = 1$.

□

B.1.3 Consistency with merged samples

Similar to Appendix B.1.2, we show that Theorem 4.3.6 and Theorem 4.3.9 can be extended to the case where we condition on the event $\mathcal{A}_n = \{\phi_f(Q_c^m, Q_h^\ell) = 1\}$. Since both the partial bootstrap and partial permutation satisfy condition (B.1), as shown in Theorems 4.3.6 and 4.3.9, the following result applies to both methods.

Proposition B.1.5 (Consistency provided merged controls)**.** *Let* $\mathcal{A}_n = \{\phi_f(Q_c^m, Q_h^\ell) = 1\}$ *denote the event of rejecting the fusion null, and assume* $\tau \coloneqq \lim_{n\to\infty} \Pr(\mathcal{A}_n) > 0$. *Suppose*

$$\lim_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1\Big) = 1 \tag{B.1}$$

under $\mathcal{H}_1$. *Then*

$$\lim_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,|\, \mathcal{A}_n\Big) = 1.$$

Proof. There exists $N \in \mathbb{N}_+$ such that $\Pr(\mathcal{A}_n) > 0$ for all $n \geq N$. In what follows, we restrict

attention to $n \geq N$, so that $\Pr(\cdot \,|\, \mathcal{A}_n)$ is well defined. Using the definition of conditional probability, we have

$$\begin{aligned}
\Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,|\, \mathcal{A}_n\Big) &= \frac{\Pr\Big(\{\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1\} \cap \mathcal{A}_n\Big)}{\Pr(\mathcal{A}_n)} \\
&= 1 - \frac{\Pr\Big(\{\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 0\} \cap \mathcal{A}_n\Big)}{\Pr(\mathcal{A}_n)} \\
&\geq 1 - \frac{\Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 0\Big)}{\Pr(\mathcal{A}_n)} \\
&\to 1 - \frac{1 - \lim_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1\Big)}{\tau} = 1,
\end{aligned}$$

where the last line results from $\tau > 0$ and (B.1), which is not conditional on $\mathcal{A}_n$. We thus show the consistency provided the controls are merged. □

B.2 Main proofs

B.2.1 Proof of Theorem 4.3.1

Proof. Theorem 4.3.1 differs slightly from Liu and Gandy (2026, Theorem 8) in that we use the quantile $q_{1-\alpha_f}^{m,\ell}$ of $S_{m,\ell}$ as the critical rule, whereas Liu and Gandy (2026, Theorem 8) is stated using a bootstrapped estimate of this quantity. Specifically, their critical value is the corresponding quantile of the conditional of $D_W^2(Q_c^m)$ given Q_c^m, where $D_W^2(Q_c^m)$ is defined in (4.7). Nevertheless, a closer inspection of their proof in Liu and Gandy (2026, Appendix B.2.3) reveals that they first prove the result using $q_{1-\alpha_f}^{m,\ell}$, and then show that the bootstrapping estimation errors arising from using its bootstrapped estimate instead of $q_{1-\alpha_f}^{m,\ell}$ converge to 0 as $m \to \infty$ (Liu and Gandy, 2026, Lemma 7). Consequently, Theorem 4.3.1 follows by applying the same proof strategy for Liu and Gandy (2026, Theorem 8). □

B.2.2 Proof of Proposition 4.3.4

Proof. We define $A \coloneqq \xi_f - \xi_c = (1-\gamma)(\xi_h - \xi_c)$, $B \coloneqq \xi_t - \xi_c$. Then $B = 0$, since we assume $Q_c = Q_t$. Furthermore, define $A_n \coloneqq \hat{\xi}_{f,m+\ell} - \hat{\xi}_{c,m} = (1-\gamma)(\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m})$, $B_n \coloneqq \hat{\xi}_{t,n} - \hat{\xi}_{c,m}$.

With these notations, we have $T = \|A\|_{\mathcal{H}_k}^2$, $T_{n,m,\ell} = \|A_n - B_n\|_{\mathcal{H}_k}^2$, $\widetilde{T}_{n,m,\ell} = \|A_n\|_{\mathcal{H}_k}^2$. Thus

$$T_{n,m,\ell} - \widetilde{T}_{n,m,\ell} = \|A_n - B_n\|_{\mathcal{H}_k}^2 - \|A_n\|_{\mathcal{H}_k}^2 = -2\langle A_n, B_n\rangle_{\mathcal{H}_k} + \|B_n\|_{\mathcal{H}_k}^2.$$

By the central limit theorem (CLT) in Hilbert space (Berlinet and Thomas-Agnan, 2011, Section 9.1; Wolfer and Alquier, 2025, Section 1.2), which applies as $\mathbb{E}k(X,X) < \infty$ under Assumption 4.2.2:

$$B_n = B_n - B = O_P(n^{-1/2}) \implies \sqrt{n}\|B_n\|_{\mathcal{H}_k}^2 = o_P(1).$$

Similarly, $A_n - A = O_P(n^{-1/2})$. Thus

$$\begin{aligned}\Delta = \sqrt{n}(T_{n,m,\ell} - \widetilde{T}_{n,m,\ell}) &= -2\sqrt{n}\langle A_n, B_n\rangle_{\mathcal{H}_k} + \sqrt{n}\|B_n\|_{\mathcal{H}_k}^2 \\ &= -2\sqrt{n}\langle A_n - A, B_n\rangle_{\mathcal{H}_k} - 2\langle A, \sqrt{n}B_n\rangle_{\mathcal{H}_k} + o_P(1) \\ &= -2\sqrt{n}\langle A, B_n\rangle_{\mathcal{H}_k} + o_P(1). \end{aligned} \tag{B.2}$$

Note that $B_n = \frac{1}{n}\sum_{i=1}^n k(\cdot, y_i^t) - \frac{1}{m}\sum_{j=1}^m k(\cdot, y_j^c)$, $A = \mathbb{E}_{Y\sim Q_f} k(\cdot, Y) - \mathbb{E}_{Z\sim Q_c} k(\cdot, Z)$. Thus

$$\langle A, B_n\rangle_{\mathcal{H}_k} = \underbrace{\frac{1}{n}\sum_{i=1}^n \left[\mathbb{E}_{Y\sim Q_f} k(y_i^t, Y) - \mathbb{E}_{Z\sim Q_c} k(y_i^t, Z)\right]}_{\hat{\psi}_t^n} - \underbrace{\frac{1}{m}\sum_{j=1}^m \left[\mathbb{E}_{Y\sim Q_f} k(y_j^c, Y) - \mathbb{E}_{Z\sim Q_c} k(y_j^c, Z)\right]}_{\hat{\psi}_c^m}.$$

Denote

$$\begin{aligned}\psi_t &\coloneqq \mathbb{E}_{X\sim Q_t}\left[\mathbb{E}_{Y\sim Q_f} k(X,Y) - \mathbb{E}_{Z\sim Q_c} k(X,Z)\right], \\ \psi_c &\coloneqq \mathbb{E}_{X\sim Q_c}\left[\mathbb{E}_{Y\sim Q_f} k(X,Y) - \mathbb{E}_{Z\sim Q_c} k(X,Z)\right]; \\ \sigma_t^2 &\coloneqq \mathrm{Var}_{X\sim Q_t}\left[\mathbb{E}_{Y\sim Q_f} k(X,Y) - \mathbb{E}_{Z\sim Q_c} k(X,Z)\right], \\ \sigma_c^2 &\coloneqq \mathrm{Var}_{X\sim Q_c}\left[\mathbb{E}_{Y\sim Q_f} k(X,Y) - \mathbb{E}_{Z\sim Q_c} k(X,Z)\right].\end{aligned}$$

We have $\psi_t = \mathbb{E}_{X\sim Q_t}\hat{\psi}_t^n$, $\psi_c = \mathbb{E}_{X\sim Q_c}\hat{\psi}_c^m$. Under the assumption $Q_t = Q_c$, we also have $\psi_t = \psi_c$, $\sigma_t^2 = \sigma_c^2$.

By the classical central limit theorem(CLT) and independence of samples,

$$\left(\sqrt{n}(\hat{\psi}_t^n - \psi_t),\ \sqrt{n}(\hat{\psi}_c^m - \psi_c)\right) \rightsquigarrow \mathcal{N}\Big(0, \mathrm{diag}\left(\sigma_c^2, \sigma_c^2/c_1\right)\Big).$$

Note that $\sqrt{n}\langle A, B_n\rangle_{\mathcal{H}_k} = \sqrt{n}(\hat{\psi}_t^n - \psi_t) - \sqrt{n}(\hat{\psi}_c^m - \psi_c)$ is a linear continuous map of $\left(\sqrt{n}(\hat{\psi}_t^n - \psi_t),\ \sqrt{n}(\hat{\psi}_c^m - \psi_c)\right)$. By the continuous mapping theorem, we obtain

$$\sqrt{n}\langle A, B_n\rangle_{\mathcal{H}_k} \rightsquigarrow \mathcal{N}(0, \sigma_t^2 + \sigma_c^2/c_1). \tag{B.3}$$

With $\sigma_c^2 = \sigma_t^2$ we get

$$\Delta \rightsquigarrow \mathcal{N}\Big(0, 4(1 + 1/c_1)\sigma_c^2\Big).$$

For $\Delta^* \coloneqq \sqrt{n}(T^*_{n,m,\ell} - \widetilde{T}^*_{n,m,\ell})$, we follow a similar procedure. Denote $A_n^* \coloneqq \hat{\xi}^*_{f,m+\ell} - \hat{\xi}^*_{c,m} = (1-\gamma)(\hat{\xi}^*_{h,\ell} - \hat{\xi}^*_{c,m})$, $B_n^* \coloneqq \hat{\xi}^*_{t,n} - \hat{\xi}^*_{c,m}$. Thus

$$T^*_{n,m,\ell} - \widetilde{T}^*_{n,m,\ell} = \|A_n^* - B_n^*\|^2_{\mathcal{H}_k} - \|A_n^*\|^2_{\mathcal{H}_k} = -2\langle A_n^*, B_n^*\rangle_{\mathcal{H}_k} + \|B_n^*\|^2_{\mathcal{H}_k}.$$

Throughout this argument, the subscript P^* on stochastic-order symbols refers to the bootstrap law $\Pr^*$. Conditional on $\mathcal{D}_{n,m,\ell}$, we get

$$B_n^* - \mathbb{E}^* B_n^* = B_n^* - (\hat{\xi}_{c,m} - \hat{\xi}_{c,m}) = B_n^* = O_{P^*}(n^{-1/2}).$$

This is because both $\hat{\xi}^*_{t,n}$ and $\hat{\xi}^*_{c,m}$ are embeddings of bootstrap samples around $\hat{\xi}^m_c$. Again we have $\sqrt{n}\|B_n^*\|^2_{\mathcal{H}_k} = o_{P^*}(1)$. Similarly, $A_n^* - A_n = O_{P^*}(n^{-1/2})$. So

$$\Delta^* = -2\sqrt{n}\langle A_n, B_n^*\rangle_{\mathcal{H}_k} + o_{P^*}(1).$$

Note that $B_n^* = \frac{1}{n}\sum_{i=1}^n k(\cdot, y_i^{t,*}) - \frac{1}{m}\sum_{j=1}^m k(\cdot, y_j^{c,*})$, $A_n = \frac{1}{m+\ell}\Big[\sum_{i=1}^m k(\cdot, y_i^c) + \sum_{j=1}^\ell k(\cdot, y_j^h)\Big] - \frac{1}{m}\sum_{i=1}^m k(\cdot, y_i^c)$. Thus

$$\langle A_n, B_n^*\rangle_{\mathcal{H}_k} = \hat{\psi}_t^{n,*} - \hat{\psi}_c^{m,*}$$

where

$$\hat{\psi}_t^{n,*} \coloneqq \frac{1}{n}\sum_{i'=1}^n \left\{\frac{1}{m+\ell}\Big[\sum_{i=1}^m k(y_{i'}^{t,*}, y_i^c) + \sum_{j=1}^\ell k(y_{i'}^{t,*}, y_j^h)\Big] - \frac{1}{m}\sum_{i=1}^m k(y_{i'}^{t,*}, y_i^c)\right\},$$
$$\hat{\psi}_c^{m,*} \coloneqq \frac{1}{m}\sum_{i'=1}^m \left\{\frac{1}{m+\ell}\Big[\sum_{i=1}^m k(y_{i'}^{c,*}, y_i^c) + \sum_{j=1}^\ell k(y_{i'}^{c,*}, y_j^h)\Big] - \frac{1}{m}\sum_{i=1}^m k(y_{i'}^{c,*}, y_i^c)\right\}.$$

Denote

$$\hat{\psi} \coloneqq \mathbb{E}^*_{X\sim Q_c^m}\Big\{\frac{1}{m+\ell}\Big[\sum_{i=1}^{m} k(X, y_i^c) + \sum_{j=1}^{\ell} k(X, y_j^h)\Big] - \frac{1}{m}\sum_{i=1}^{m} k(X, y_i^c)\Big\}.$$

Note that $\hat{\psi}_t^{n,*}$ and $\hat{\psi}_c^{m,*}$ are the means of $g_n(y) \coloneqq \langle A_n, k(\cdot, y)\rangle_{\mathcal{H}_k}$ over the treatment and control bootstrap samples, and that $\hat{\psi} = \mathbb{E}^* g_n(X)$ with $X \sim Q_c^m$. Define the fixed function $g(y) \coloneqq \langle A, k(\cdot, y)\rangle_{\mathcal{H}_k}$. By the strong law of large numbers in $\mathcal{H}_k$, $\|A_n - A\|_{\mathcal{H}_k} \to 0$ almost surely, and therefore

$$v_m \coloneqq \frac{1}{m}\sum_{i=1}^{m}(g_n - g)^2(y_i^c) \le \|A_n - A\|^2_{\mathcal{H}_k}\,\frac{1}{m}\sum_{i=1}^{m} k(y_i^c, y_i^c) \to 0 \quad \text{almost surely.}$$

Hence, after centring and scaling by $\sqrt{n}$, the bootstrap means of g_n differ from those of g by conditionally centred terms with conditional variances at most v_m and $\frac{n}{m}\,v_m$, respectively, which are $o_{P^*}(1)$ by Chebyshev's inequality. Moreover, $\mathbb{E}_{X\sim Q_c} g(X)^2 \le \|A\|^2_{\mathcal{H}_k}\,\mathbb{E}_{X\sim Q_c} k(X, X) < \infty$ under Assumption 4.2.2, and the two bootstrap samples are independent given $\mathcal{D}_{n,m,\ell}$. With $\rightsquigarrow_{P^*}$ denoting conditional weak convergence given the data, Bickel and Freedman (1981, Theorem 2.2) applied to g and Slutsky's theorem give

$$\Big(\sqrt{n}(\hat{\psi}_t^{n,*} - \hat{\psi}), \sqrt{n}(\hat{\psi}_c^{m,*} - \hat{\psi})\Big) \rightsquigarrow_{P^*} \mathcal{N}\Big(0, \operatorname{diag}\left(\sigma_c^2, \sigma_c^2/c_1\right)\Big),$$

where $\sigma_c^2 \coloneqq \operatorname{Var}_{X\sim Q_c} g(X)$. With the continuous mapping theorem and Slutsky's theorem we obtain

$$\Delta^* \rightsquigarrow_{P^*} \mathcal{N}\Big(0, 4(1 + 1/c_1)\sigma_c^2\Big).$$

Since Δ and Δ^* converge to the same continuous limit, Pólya's theorem and the triangle inequality give us

$$\sup_{z\in\mathbb{R}}\left|\Pr^*\Big\{\sqrt{n}(T^*_{n,m,\ell} - \widetilde{T}^*_{n,m,\ell}) \le z\Big\} - \Pr\Big\{\sqrt{n}(T_{n,m,\ell} - \widetilde{T}_{n,m,\ell}) \le z\Big\}\right| \xrightarrow{a.s.} 0.$$

□

B.2.3 Proof of Theorem 4.3.5

Proof. We write $F_n^*(z) = \Pr^*\Big\{\sqrt{n}(T^*_{n,m,\ell} - \widetilde{T}^*_{n,m,\ell}) \le z\Big\}$ as the conditional bootstrap CDF, and $F_n(z) = \Pr\Big\{\sqrt{n}(T_{n,m,\ell} - \widetilde{T}_{n,m,\ell}) \le z\Big\}$ as the unconditional CDF. Also write $F(\cdot)$ for the CDF

of $\mathcal{N}(0, 4(1+1/c_1)\sigma_c^2)$; we assume $\sigma_c^2 > 0$, so that F is continuous and strictly increasing. By Proposition 4.3.4 and Portmanteau theorem, we have

$$F_n(z) \to F(z)$$

at every continuity point z of F. Moreover, we have the uniform consistency

$$\sup_{z\in\mathbb{R}} \left| F_n^*(z) - F_n(z) \right| \xrightarrow{a.s.} 0. \tag{B.4}$$

Define $q^*_{n,1-\alpha} \coloneqq \inf_{z\in\mathbb{R}}\{z : F_n^*(z) \geq 1-\alpha\}$, $q_{n,1-\alpha} \coloneqq \inf_{z\in\mathbb{R}}\{z : F_n(z) \geq 1-\alpha\}$ and $q_{1-\alpha} \coloneqq \inf_{z\in\mathbb{R}}\{z : F(z) \geq 1-\alpha\}$.

Since F is continuous and strictly increasing, (B.4) and $F_n \to F$ give $q^*_{n,1-\alpha} \xrightarrow{a.s.} q_{1-\alpha}$.

We denote $F_{n,B}$ as the empirical CDF of the bootstrap draws Δ^* and define $\delta'_B = \sqrt{\frac{1}{2B}\log(2B)}$. By the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956), the event $\mathcal{A} = \left\{ \sup_u \left| F_{n,B}(u) - F_n^*(u) \right| \leq \delta'_B \right\}$ occurs with probability $\Pr^*(\mathcal{A}) \geq 1 - B^{-1}$. On $\mathcal{A}$,

$$\begin{aligned} q^{m+\ell,n}_{1-\alpha,B} &= \inf\left\{u \in \mathbb{R} : 1-\alpha \leq F_{n,B}(u)\right\} \geq \inf\left\{u \in \mathbb{R} : 1-\alpha \leq F_n^*(u) + \delta'_B\right\} \\ &= \inf\left\{u \in \mathbb{R} : 1-\alpha-\delta'_B \leq F_n^*(u)\right\} = q^*_{n,1-\alpha-\delta'_B}, \end{aligned}$$

so $\{\Delta > q^{m+\ell,n}_{1-\alpha,B}\} \cap \mathcal{A} \subseteq \{\Delta > q^*_{n,1-\alpha-\delta'_B}\}$; taking expectation over the Monte Carlo draws conditional on $\mathcal{D}_{n,m,\ell}$ and then over the data,

$$\Pr(\Delta > q^{m+\ell,n}_{1-\alpha,B}) \leq \Pr(\Delta > q^*_{n,1-\alpha-\delta'_B}) + \Pr(\mathcal{A}^c) \leq \Pr(\Delta > q^*_{n,1-\alpha-\delta'_B}) + B^{-1}.$$

Since F_n^* has finite support conditional on the data, $F_n^*(q^{*-}_{n,1-\alpha}) < 1-\alpha$ by definition of $q^*_{n,1-\alpha}$ as an infimum, so $q^*_{n,1-\alpha-\delta'_B} = q^*_{n,1-\alpha}$ for all large B; letting $B \to \infty$ gives $\lim_{B\to\infty} \Pr(\Delta > q^{m+\ell,n}_{1-\alpha,B}) \leq \Pr(\Delta > q^*_{n,1-\alpha})$. As $q^*_{n,1-\alpha} \xrightarrow{a.s.} q_{1-\alpha}$, a constant, and $\Delta \rightsquigarrow \mathcal{N}(0, 4(1+1/c_1)\sigma_c^2)$ with F continuous at $q_{1-\alpha}$, Slutsky's theorem gives $\limsup_{n\to\infty} \Pr(\Delta > q^*_{n,1-\alpha}) = 1 - F(q_{1-\alpha}) = \alpha$. Therefore

$$\limsup_{n\to\infty} \lim_{B\to\infty} \Pr(\Delta > q^{m+\ell,n}_{1-\alpha,B}) \leq \alpha,$$

which completes the proof. □

B.2.4 Proof of Theorem 4.3.6

Proof. We start by using $q^*_{1-\alpha}$, which is the $(1-\alpha)$-quantile of the distribution of $\sqrt{n}(T^*_{n,m,\ell} - \widetilde{T}^*_{n,m,\ell})$. We write $T \coloneqq D^2(Q_f, Q_t) = \|\xi_f - \xi_t\|^2_{\mathcal{H}_k}$. Using the notations in Appendix B.2.2, we denote $B \coloneqq \xi_t - \xi_c$, $A \coloneqq \xi_f - \xi_c$. Define the population gap

$$\Delta_\infty \coloneqq T - \widetilde{T} = \|\xi_f - \xi_t\|^2_{\mathcal{H}_k} - \|\xi_f - \xi_c\|^2_{\mathcal{H}_k} = \|A - B\|^2_{\mathcal{H}_k} - \|A\|^2_{\mathcal{H}_k} = -2\langle A, B\rangle_{\mathcal{H}_k} + \|B\|^2_{\mathcal{H}_k}.$$

Assume the alternative hypothesis and that $2(1-\gamma)D(Q_h, Q_c)\cos\beta < D(Q_c, Q_t)$. Since $A = \xi_f - \xi_c = (1-\gamma)(\xi_h - \xi_c)$, so that $(1-\gamma)D(Q_h, Q_c) = \|A\|_{\mathcal{H}_k}$, we have $2\|A\|_{\mathcal{H}_k}\cos\beta < \|B\|_{\mathcal{H}_k}$. Consequently,

$$\Delta_\infty = \|B\|^2_{\mathcal{H}_k} - 2\langle A, B\rangle_{\mathcal{H}_k} = \|B\|^2_{\mathcal{H}_k} - 2\|A\|_{\mathcal{H}_k}\|B\|_{\mathcal{H}_k}\cos\beta > 0.$$

With the assumption that all kernel mean embeddings have finite first RKHS moment, the Bochner integral is well defined and the empirical embedding satisfies a strong law of large numbers (Araujo and Giné, 1980, Theorem 1.1; Ledoux and Talagrand, 2013, Corollary 7.10). Together with Lemma B.1.1, we have $T_{n,m,\ell} \to T$, $\widetilde{T}_{n,m,\ell} \to \widetilde{T}$ almost surely. Hence

$$(T_{n,m,\ell} - \widetilde{T}_{n,m,\ell}) \to \Delta_\infty > 0.$$

Thus,

$$\Delta = \sqrt{n}(T_{n,m,\ell} - \widetilde{T}_{n,m,\ell}) \to \infty.$$

The partial bootstrap statistic $\sqrt{n}(T^*_{n,m,\ell} - \widetilde{T}^*_{n,m,\ell})$ is built only from Q^m_c and Q^ℓ_h, its bootstrap treatment arm being resampled from Q^m_c, so its limit is unchanged under $\mathcal{H}_1$: as proved in Appendix B.2.2, it converges to $\mathcal{N}(0, 4(1+1/c_1)\sigma^2_c)$. Thus $q^*_{1-\alpha} = O_P(1)$. By the conditional Glivenko-Cantelli, we have the quantile of bootstrap samples $\hat{q}^{m+\ell,n}_{1-\alpha,B} \to q^*_{1-\alpha}$. We thus obtain

$$\lim_{n\to\infty} \Pr(\Delta > \hat{q}^{m+\ell,n}_{1-\alpha,B} \,|\, \mathcal{H}_1) = 1.$$

Note that $\Delta_\infty > 0$ is the sufficient and necessary condition for consistency to hold. We next show that if $\Delta_\infty = 0$, the test is not consistent.

Using the same notations in Appendix B.2.3, we have

$$A_n - A = O_P(n^{-1/2}), \quad B_n - B = O_P(n^{-1/2}),$$

and

$$\begin{aligned}
\widetilde{V}_n &\coloneqq \Delta - \Delta_\infty \\
&= \sqrt{n}\Big(- 2\langle A_n, B_n\rangle_{\mathcal{H}_k} + \|B_n\|^2_{\mathcal{H}_k} + 2\langle A, B\rangle_{\mathcal{H}_k} - \|B\|^2_{\mathcal{H}_k}\Big) \\
&= -2\sqrt{n}\langle A_n - A, B\rangle_{\mathcal{H}_k} - 2\sqrt{n}\langle A, B_n - B\rangle_{\mathcal{H}_k} + 2\sqrt{n}\langle B, B_n - B\rangle_{\mathcal{H}_k} + o_P(1) \\
&= 2\sqrt{n}\Big(\underbrace{\langle B - A, \hat{\xi}_{t,n} - \xi_t\rangle_{\mathcal{H}_k}}_{\widetilde{\psi}_t^n - \widetilde{\psi}_t} - \underbrace{\langle B, \hat{\xi}_{f,m+\ell} - \xi_f\rangle_{\mathcal{H}_k}}_{\widetilde{\psi}_f^{m+\ell} - \widetilde{\psi}_f} + \underbrace{\langle A, \hat{\xi}_{c,m} - \xi_c\rangle_{\mathcal{H}_k}}_{\widetilde{\psi}_c^m - \widetilde{\psi}_c}\Big) + o_P(1),
\end{aligned}$$

where we have

$$\begin{aligned}
\widetilde{\psi}_t(X) &\coloneqq \mathbb{E}_{Y\sim Q_t} k(X, Y) - \mathbb{E}_{Z\sim Q_f} k(X, Z), \\
\widetilde{\psi}_f(X) &\coloneqq \mathbb{E}_{Y\sim Q_t} k(X, Y) - \mathbb{E}_{Z\sim Q_c} k(X, Z), \\
\widetilde{\psi}_c(X) &\coloneqq \mathbb{E}_{Y\sim Q_f} k(X, Y) - \mathbb{E}_{Z\sim Q_c} k(X, Z).
\end{aligned}$$

Denote the corresponding variances

$$\widetilde{\sigma}_t^2 \coloneqq \mathrm{Var}_{X\sim Q_t}\, \widetilde{\psi}_t(X), \quad \widetilde{\sigma}_f^2 \coloneqq \mathrm{Var}_{X\sim Q_f}\, \widetilde{\psi}_f(X), \quad \widetilde{\sigma}_c^2 = \mathrm{Var}_{X\sim Q_c}\, \widetilde{\psi}_c(X).$$

Let the empirical counterparts be

$$\widetilde{\psi}_t^n \coloneqq \frac{1}{n}\sum_{i=1}^n \widetilde{\psi}_t(y_i^t), \quad \widetilde{\psi}_f^{m+\ell} \coloneqq \frac{1}{m+\ell}\Big[\sum_{i=1}^m \widetilde{\psi}_f(y_i^c) + \sum_{i=1}^\ell \widetilde{\psi}_f(y_i^h)\Big], \quad \widetilde{\psi}_c^m \coloneqq \frac{1}{m}\sum_{i=1}^m \widetilde{\psi}_c(y_i^c).$$

By the CLT in Hilbert space, the linearity and continuous mapping, we get

$$\widetilde{V}_n \rightsquigarrow \mathcal{N}\Big(0, 4(\widetilde{\sigma}_t^2 + \widetilde{\sigma}_c^2/c_1 + \widetilde{\sigma}_f^2/(c_1 + c_2))\Big).$$

Thus when $\Delta_\infty = 0$, we have $\Delta \rightsquigarrow \mathcal{N}(0, \sigma_1^2)$ where $\sigma_1^2 \coloneqq 4(\widetilde{\sigma}_t^2 + \widetilde{\sigma}_c^2/c_1 + \widetilde{\sigma}_f^2/(c_1 + c_2))$ can be different from $\sigma_0^2 \coloneqq 4(1 + 1/c_1)\sigma_c^2$. As a result, we have

$$\Pr\{\Delta > q^*_{1-\alpha}\} = 1 - \Phi\Big(\frac{\sigma_0}{\sigma_1} z_{1-\alpha}\Big)$$

where $\Phi(\cdot)$ is the standard Gaussian CDF and $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of standard Gaussian. Since $\frac{\sigma_0}{\sigma_1} z_{1-\alpha} > 0$, $1 - \Phi(\frac{\sigma_0}{\sigma_1} z_{1-\alpha}) < 0.5$. By replacing $q^*_{1-\alpha}$ with $\hat{q}^{m+\ell,n}_{1-\alpha,B}$ and taking $n \to \infty$, we obtain $\lim_{n\to\infty} \Pr(\Delta > \hat{q}^{m+\ell,n}_{1-\alpha,B} \,|\, \mathcal{H}_1) < 0.5$, thus the test is not consistent. □

B.2.5 Proof of Theorem 4.3.8

Proof. Write $Z := \{Z_1, \ldots, Z_{m+n}\}$ for the pooled sample $Q^m_c \cup Q^n_t$. Let G be the set of all permutations π of $I_Z := \{1, \ldots, m+n\}$. Denote the observed control indices as I_0.

For any $\pi \in G$, write the permuted control index set $I_c(\pi)$, permuted treatment index set $I_t(\pi)$, with $|I_c(\pi)| = m$, $|I_t(\pi)| = n$. Define the permuted empirical measures

$$Q^m_{c,\pi} := \frac{1}{m} \sum_{i \in I_c(\pi)} \delta_{Z_i}, \quad Q^n_{t,\pi} := \frac{1}{n} \sum_{i \in I_t(\pi)} \delta_{Z_i},$$

with the fused control

$$Q^{m+\ell}_{f,\pi} := \frac{m}{m+\ell} Q^m_{c,\pi} + \frac{\ell}{m+\ell} Q^\ell_h,$$

and let $T^\pi_{n,m,\ell} := D^2(Q^{m+\ell}_{f,\pi}, Q^n_{t,\pi})$. The observed statistic is $T_{n,m,\ell} = T^{\pi^0}_{n,m,\ell}$ where π^0 is the identity $\pi^0(i) = i$.

Under $H_0 : Q_c = Q_t$, the pooled sample $Z_1, \ldots, Z_{m+n} \overset{\text{i.i.d.}}{\sim} Q_c$ has a joint distribution that is invariant under permutation. Moreover, since Q^ℓ_h is independent of Z and not permuted, the joint distribution, on the population level, satisfies

$$(Z, Q^\ell_h) \overset{d}{=} (\pi \cdot Z, Q^\ell_h), \quad \forall \pi \in G.$$

This is precisely the randomisation hypothesis: the joint distribution of the data is invariant under the group G. Since $T^\pi_{n,m,\ell}$ is a measurable function of $(\pi \cdot Z, Q^\ell_h)$, the collection $\{T^\pi_{n,m,\ell} : \pi \in G\}$ is invariant under the group so satisfies the assumptions of the randomised test theorem (Lehmann and Romano, 2022, Definition 17.2.1; Theorem 17.2.2).

In our algorithm, we only sample a subset of permutations $\pi^1, \ldots, \pi^B$ of size B uniformly from G, and calculate $\hat{c}^{m+\ell,n}_{1-\alpha,B}$ as the empirical quantile of $\{T^{\pi_b}_{n,m,\ell}\}_{b=0}^B$. Conditional on the data $\mathcal{D}_{n,m,\ell}$, the collection $\{T^{\pi_b}_{n,m,\ell}\}_{b=1}^B$ is exchangeable. By exchangeability, the rank of $T_{n,m,\ell} = T^{\pi_0}_{n,m,\ell}$ among $\{T^{\pi_b}_{n,m,\ell}\}_{b=1}^B$ is uniform on $\{1, \ldots, B+1\}$. We choose $\hat{c}^{m+\ell,n}_{1-\alpha,B}$ so that at most an α-fraction of the sample has value larger than or equal to it. Hence, by the Monte Carlo

randomisation tests (Lehmann and Romano, 2022), we have

$$\Pr\Big\{T_{n,m,\ell} \geq \hat{c}^{m+\ell,n}_{1-\alpha,B} \,|\, Z, Q_h^\ell\Big\} \leq \alpha. \tag{B.5}$$

Finally, take expectations over (Z, Q_h^ℓ) we get the unconditional result under $\mathcal{H}_0$:

$$\Pr\{T_{n,m,\ell} \geq \hat{c}^{m+\ell,n}_{1-\alpha,B}\} \leq \alpha,$$

for all n, m, ℓ, B. □

B.2.6 Proof of Theorem 4.3.9

We first introduce the Hadamard differentiability of the squared MMD.

Lemma B.2.1 (Hadamard differentiability of the squared MMD)**.** *For $\gamma \in (0,1)$, the squared MMD function $D^2(\gamma Q_c + (1-\gamma)Q_h, Q_t)$ is Hadamard differentiable with respect to probability measures Q_c, Q_h and Q_t.*

Proof of Lemma B.2.1. Define $\Psi : \mathcal{H}_k^3 \times [0,1] \to \mathbb{R}$,

$$\Psi(\xi_c, \xi_h, \xi_t; \gamma) := \Big\|\xi_t - (\gamma\xi_c + (1-\gamma)\xi_h)\Big\|^2_{\mathcal{H}_k}.$$

Since Ψ is a quadratic form on a Hilbert product space, it is Fréchet, hence Hadamard, differentiable with respect to (ξ_c, ξ_h, ξ_t). Composing with the linear embedding map $R \mapsto \xi_R$, where R is a probability measure, gives Hadamard differentiability of $(Q_c, Q_h, Q_t) \mapsto T(Q_c, Q_h, Q_t) = \Psi(\xi_c, \xi_h, \xi_t; \gamma)$. □

Now we provide the proof for Theorem 4.3.9.

Proof of Theorem 4.3.9. Denote $Q_d := (1-\lambda)Q_c + \lambda Q_t$, $Q_{f,d} = \gamma Q_d + (1-\gamma)Q_h$. The corresponding population kernel mean embeddings are written as $\xi_d := (1-\lambda)\xi_c + \lambda\xi_t$, $\xi_{f,d} := \gamma\xi_d + (1-\gamma)\xi_h$.

Denote the pooled empirical measure of $Z := (Z_1, \ldots, Z_{m+n})$ (defined in Appendix B.2.5) as $Q_d^{m+n} := (1-\lambda)Q_c^m + \lambda Q_t^n$. In each permutation b, let Q^m_{c,π^b} be the empirical measure of the m observations drawn without replacement from the $m+n$ pooled observations, and similarly Q^n_{t,π^b} for the remaining n. The fused control in permutation b is thus $Q^{m+\ell}_{f,\pi^b} = \gamma Q^m_{c,\pi^b} + (1-\gamma)Q^\ell_{h,b}$;

the empirical measure of the b-th pooled data is denoted $Q^{m+n}_{d,\pi^b} \coloneqq (1-\lambda)Q^m_{c,\pi^b} + \lambda Q^n_{t,\pi^b}$. The statistic $T^b_{n,m,\ell}$ of the b-th permutation is thus $D^2(Q^{m+\ell}_{f,\pi^b}, Q^n_{t,\pi^b})$.

By the extended Glivenko–Cantelli theorem for finite populations (Motoyama, 2024, Corollary 3.2), applied conditionally on the pooled sample, the empirical distribution functions of Q^m_{c,π^b} and Q^n_{t,π^b} converge uniformly to that of Q^{m+n}_d almost surely; since $Q^{m+n}_d \rightsquigarrow Q_d$ almost surely by the classical Glivenko–Cantelli theorem, it follows that

$$Q^m_{c,\pi^b} \rightsquigarrow Q_d, \qquad Q^n_{t,\pi^b} \rightsquigarrow Q_d,$$

almost surely as $m, n \to \infty$.

For Q_h, the empirical kernel mean embedding satisfies $\xi^\ell_h \to \xi_h$ almost surely. By linearity of kernel mean embeddings and the continuous mapping theorem, it follows that

$$Q^{m+\ell}_{f,\pi^b} \rightsquigarrow Q_{f,d}.$$

As $D^2(\cdot,\cdot)$ is Hadamard differentiable according to Lemma B.2.1, together with continuous mapping theorem we get

$$T^b_{n,m,\ell} = D^2(Q^{m+\ell}_{f,\pi^b}, Q^n_{t,\pi^b}) \to D^2(Q_{f,d}, Q_d).$$

So the partial permutation yields $T^b_{n,m,\ell}$ that converges to

$$\begin{aligned} D^2(Q_{f,d}, Q_d) &= \Big\| \frac{m}{m+\ell}\xi_d + \frac{\ell}{m+\ell}\xi_h - \xi_d \Big\|^2_{\mathcal{H}_k} \\ &= \|(1-\gamma)(\xi_h - \xi_d)\|^2_{\mathcal{H}_k} \\ &= \|(1-\gamma)(\xi_h - \xi_c) - \lambda(1-\gamma)(\xi_t - \xi_c)\|^2_{\mathcal{H}_k}. \end{aligned}$$

While $T_{n,m,\ell}$ converges to

$$\|(1-\gamma)(\xi_h - \xi_c) - (\xi_t - \xi_c)\|^2_{\mathcal{H}_k}.$$

We want them to be separable, which means

$$\underbrace{\|(1-\gamma)(\xi_h - \xi_c) - \lambda(1-\gamma)(\xi_t - \xi_c)\|^2_{\mathcal{H}_k}}_{(1)} < \underbrace{\|(1-\gamma)(\xi_h - \xi_c) - (\xi_t - \xi_c)\|^2_{\mathcal{H}_k}}_{(2)}.$$

We have

$$(1)-(2)=\big(1-\lambda(1-\gamma)\big)\Big\{2(1-\gamma)\langle\xi_h-\xi_c,\xi_t-\xi_c\rangle_{\mathcal{H}_k}-\big(1+\lambda(1-\gamma)\big)\|\xi_t-\xi_c\|^2_{\mathcal{H}_k}\Big\}.$$

Since $\lambda\in(0,1)$, $\gamma\in(0,1)$, we have $1-\lambda(1-\gamma)>0$. Thus

$$(1)<(2)\iff 2(1-\gamma)\langle\xi_h-\xi_c,\xi_t-\xi_c\rangle_{\mathcal{H}_k}-\big(1+\lambda(1-\gamma)\big)\|\xi_t-\xi_c\|^2_{\mathcal{H}_k}<0.$$

Since we have $\|\xi_t-\xi_c\|_{\mathcal{H}_k}\neq 0$, the above inequality stands if and only if $\xi_h-\xi_c=0$ or $2\cos\beta\,\|\xi_h-\xi_c\|_{\mathcal{H}_k}<\Big(\frac{1}{1-\gamma}+\lambda\Big)\|\xi_t-\xi_c\|_{\mathcal{H}_k}$. □

B.2.7 Proof of Theorem 4.3.12

Proof. Assuming first that $\mathcal{H}_0^f : D(Q_h,Q_c)\geq\rho$ holds, we can rewrite the Type-I error rate using law of total probability as

$$\begin{aligned}
&\Pr\Big(\phi_*(Q_c^m,Q_h^\ell,Q_t^n)=1\,|\,\mathcal{H}_0,\,\mathcal{H}_0^f\Big)\\
&=\underbrace{\Pr\Big(\phi_*(Q_c^m,Q_h^\ell,Q_t^n)=1\,|\,\mathcal{H}_0,\,\mathcal{H}_0^f,\,\phi_f(Q_c^m,Q_h^\ell)=1\Big)}_{T_{11}}\underbrace{\Pr\Big(\phi_f(Q_c^m,Q_h^\ell)=1\,|\,\mathcal{H}_0,\,\mathcal{H}_0^f\Big)}_{T_{12}}\\
&\quad+\underbrace{\Pr\Big(\phi_*(Q_c^m,Q_h^\ell,Q_t^n)=1\,|\,\mathcal{H}_0,\,\mathcal{H}_0^f,\,\phi_f(Q_c^m,Q_h^\ell)=0\Big)}_{T_{13}}\underbrace{\Pr\Big(\phi_f(Q_c^m,Q_h^\ell)=0\,|\,\mathcal{H}_0,\,\mathcal{H}_0^f\Big)}_{T_{14}}\\
&=T_{11}T_{12}+T_{13}(1-T_{12}),
\end{aligned}$$

where the last line holds since $T_{14}=1-T_{12}$. Intuitively, T_{11} is the probability of rejecting the causality test given that the fusion test is also rejected, i.e., Q_c^m and Q_h^ℓ are merged, and T_{12} is the probability of rejecting the fusion test. The terms T_{13} and T_{14} have similar interpretations. We will bound each term separately. We first observe that the limiting value of T_{12} is

$$T_{12}^\infty\coloneqq\lim_{n\to\infty}T_{12}\overset{(a)}{=}\lim_{n\to\infty}\Pr\Big(\phi_f(Q_c^m,Q_h^\ell)=1\,|\,\mathcal{H}_0^f\Big)\overset{(b)}{=}\begin{cases}0\,, & D(Q_c,Q_h)>\rho\,,\\ a, & D(Q_c,Q_h)=\rho,\end{cases}\tag{B.6}$$

where a is defined in (4.10), step (a) holds since the events $\{\phi_f(Q_c^m,Q_h^\ell)=1\}$ and $\{\mathcal{H}_0:Q_c=Q_t\}$ are independent, and (b) holds by noting that the LHS is the limiting Type-I error rate of the fusion test under $\mathcal{H}_0^f : D(Q_h,Q_c)\geq\rho$, which is either 0 or a by the assumption (4.10). Moreover, we can bound T_{13} as

$$\limsup_{n\to\infty} T_{13} = \limsup_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_0\Big) \le \alpha,$$

where the equality holds since $\phi_f(Q_c^m, Q_h^\ell) = 0$ means no merging, so $\phi_*(Q_c^m, Q_h^\ell, Q_t^n)$ does not depend on Q_h^ℓ, and the last step follows by the assumed Type-I error control of the causality test ϕ_*, namely (4.11). We now bound T_{11}, the Type-I error of the causality test given merging, by treating it separately for $D(Q_h, Q_c) > \rho$ and $D(Q_h, Q_c) = \rho$. When $D(Q_h, Q_c) > \rho$, we bound the probability T_{11} by 1 and combine with the above results to yield

$$\begin{aligned}\limsup_{n\to\infty} \Pr\Big(\phi_*(Q_f^{m+\ell}, Q_t^n) = 1 \,\big|\, \mathcal{H}_0,\, D(Q_c, Q_h) > \rho\Big) &\le 1 \times \lim_{n\to\infty} T_{12} + \limsup_{n\to\infty} T_{13} \times \lim_{n\to\infty} (1 - T_{12}) \\ &\le 0 + \alpha \times 1 \\ &= \alpha. \end{aligned} \tag{B.7}$$

where in the first step we have used the inequality $\limsup_{n\to\infty} f_n g_n \le (\limsup_{n\to\infty} f_n)(\limsup_{n\to\infty} g_n)$ for non-negative sequences f_n, g_n, and the second line holds due to (B.6) and that we have assumed $D(Q_c, Q_h) > \rho$.

When instead $D(Q_c, Q_h) = \rho$, we invoke assumption (4.12) to give

$$\limsup_{n\to\infty} T_{11} = \limsup_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_0,\, D(Q_c, Q_h) = \rho,\, \phi_f(Q_c^m, Q_h^\ell) = 1\Big) \le \alpha.$$

In this case, the overall Type-I error rate can be bounded as

$$\begin{aligned}&\limsup_{n\to\infty} \Pr\Big(\phi_*(Q_f^{m+\ell}, Q_t^n) = 1 \,\big|\, \mathcal{H}_0,\, D(Q_h, Q_c) = \rho\Big) \\ &\le \limsup_{n\to\infty} T_{11} \times \lim_{n\to\infty} T_{12} + \limsup_{n\to\infty} T_{13} \times \lim_{n\to\infty} (1 - T_{12}) \\ &\le \alpha T_{12}^\infty + \alpha(1 - T_{12}^\infty) \\ &= \alpha. \end{aligned} \tag{B.8}$$

This proves the desired result when $\mathcal{H}_0^f : D(Q_h, Q_c) \ge \rho$.

We now prove the claimed result assuming $\mathcal{H}_1^f : D(Q_h, Q_c) < \rho$. Decomposing the Type-I error probability in a similar way as before, we have

$$
\begin{aligned}
&\Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_0,\, \mathcal{H}_1^f\Big) \\
&= \underbrace{\Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_0,\, \mathcal{H}_1^f,\, \phi_f(Q_c^m, Q_h^\ell) = 1\Big)}_{T_{21}} \underbrace{\Pr\Big(\phi_f(Q_c^m, Q_h^\ell) = 1 \,\big|\, \mathcal{H}_0,\, \mathcal{H}_1^f\Big)}_{T_{22}} \\
&\quad + \underbrace{\Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_0,\, \mathcal{H}_1^f,\, \phi_f(Q_c^m, Q_h^\ell) = 0\Big)}_{T_{23}} \underbrace{\Pr\Big(\phi_f(Q_c^m, Q_h^\ell) = 0 \,\big|\, \mathcal{H}_0,\, \mathcal{H}_1^f\Big)}_{T_{24}} \\
&= T_{21} + (T_{23} - T_{21})T_{24},
\end{aligned}
$$

where in the last line we have used $T_{22} = 1 - T_{24}$. Under the assumed condition (4.10) on the fusion test, we have $\lim_{n\to\infty} T_{24} = 0$. Since furthermore $|T_{23} - T_{21}|$ is bounded, this implies that $\lim_{n\to\infty} (T_{23} - T_{21})T_{24} = 0$. Moreover, under the event $\{\phi_f(Q_c^m, Q_h^\ell) = 1\}$, the empirical sample $Q_f^{m+\ell}$ is the merged sample based on the independent samples Q_c^m and Q_h^ℓ. This implies that

$$
\limsup_{n\to\infty} T_{21} = \limsup_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_0,\, D(Q_c, Q_h) < \rho\Big) \le \alpha,
$$

where the inequality holds by the validity of the causality test assumed in (4.12). Combining these results gives

$$
\limsup_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \mathcal{H}_0,\, \mathcal{H}_1^f\Big) \le \limsup_{n\to\infty} T_{21} + \lim_{n\to\infty} (T_{23} - T_{21})\, T_{24} \le \alpha.
$$

The above inequality, together with (B.7) and (B.8), completes the proof. □

B.2.8 Proof of Theorem 4.3.16

Proof. Fix Q_c, Q_h, Q_t with $Q_c \neq Q_t$. Decomposing the probability of rejection in the same way as in Appendix B.2.7, we have

$$
\begin{aligned}
\Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1\Big) \;&=\; \underbrace{\Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \phi_f(Q_c^m, Q_h^\ell) = 1\Big)}_{T_1} \underbrace{\Pr\Big(\phi_f(Q_c^m, Q_h^\ell) = 1\Big)}_{T_2} \\
&\quad + \underbrace{\Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1 \,\big|\, \phi_f(Q_c^m, Q_h^\ell) = 0\Big)}_{T_3} \underbrace{\Pr\Big(\phi_f(Q_c^m, Q_h^\ell) = 0\Big)}_{1-T_2} \\
&= T_1 T_2 + T_3(1 - T_2). \qquad \text{(B.9)}
\end{aligned}
$$

The assumed condition (4.10) on the fusion test implies that $\lim_{n\to\infty} T_2 = T_2^\infty$, where

$$T_2^\infty = \begin{cases} 0, & \text{if } D(Q_c, Q_h) > \rho, \\ a, & \text{if } D(Q_c, Q_h) = \rho, \\ 1, & \text{if } D(Q_c, Q_h) < \rho, \end{cases}$$

for some constant $a \in (0, \alpha_f]$. Moreover, T_1 is the probability of rejecting the causal null hypothesis $\mathcal{H}_0$ when Q_c^m and Q_h^ℓ are merged, and T_3 the probability when they are not. By (4.13), $\lim_{n\to\infty} T_3 = 1$; by (4.14), $\lim_{n\to\infty} T_1 = 1$ when $D(Q_h, Q_c) \leq \rho$, that is, whenever $T_2^\infty > 0$. Combining these arguments with the previous decomposition gives

$$\lim_{n\to\infty} \Pr\Big(\phi_*(Q_c^m, Q_h^\ell, Q_t^n) = 1\Big) = T_2^\infty \lim_{n\to\infty} T_1 + (1 - T_2^\infty) \lim_{n\to\infty} T_3 = 1.$$

□

B.2.9 Proof of Corollary 4.3.17

Proof. It suffices to verify the conditions in Theorem 4.3.12 hold for the proposed fusion test and the causality tests in Algorithm 4.1.

Firstly, since the fusion test is an MMD equivalence test, Theorem 4.3.1 immediately shows that the condition (4.10) holds. The second condition (4.11) requires the causality test to be well-calibrated when the fusion test is *not* rejected. In this case, the causality test is a standard MMD test between Q_c^m and Q_t^n. Hence, (4.11) follows from existing result on the validity of the standard MMD permutation test; see, e.g., Schrab et al. (2023, Proposition 1).

It remains to verify (4.12) for the partial bootstrap and partial permutation causality tests. The unconditional results Theorem 4.3.5 and Theorem 4.3.8 do not apply directly, since the probability in (4.12) is *conditional* on the merge event $\{\phi_f(Q_c^m, Q_h^\ell) = 1\}$. Their conditional counterparts cover the merging case: when the fusion test merges with probability tending to one, Theorem B.1.4 gives (4.12) for the partial permutation, and Theorem B.1.3 gives it for the partial bootstrap under the additional condition $\sigma_c^2 > 0$. As the fusion test is an MMD equivalence test, Theorem 4.3.1 shows that

$$\lim_{n\to\infty} \Pr(\phi_f(Q_c^m, Q_h^\ell) = 1) = \begin{cases} 1, & D(Q_c, Q_h) < \rho, \\ 0, & D(Q_c, Q_h) > \rho. \end{cases}$$

For $D(Q_c, Q_h) < \rho$ the conditional and unconditional levels agree in the limit, giving (4.12); for $D(Q_c, Q_h) > \rho$ the level is controlled by the no-merge calibration (4.11) alone. In either case the conditions of Theorem 4.3.12 hold, proving Corollary 4.3.17. The boundary $D(Q_c, Q_h) = \rho$ is not covered: there the merge probability has a limit strictly between zero and one, and the merge event is determined by the current control Q_c^m, which also enters the causal statistic, so conditioning on it selects on the same observations and the level is no longer controlled. □

B.2.10 Proof of Corollary 4.3.18

Proof. As we conduct the MMD equivalence test for the fusion test, condition (4.10) is satisfied.

When the assumed conditions hold, under $\mathcal{H}_1$, if $\phi_f(Q_c^m, Q_h^\ell) = 0$, the historical control is not fused. In this case, both partial permutation and partial bootstrap revert back to the classical two-sample MMD testing between Q_c^m and Q_t^n. Gretton et al. (2006) shows that the MMD computed with characteristic kernel converges to a positive constant under a fixed alternative when two distributions are different. Consistency of the resulting permutation or bootstrap calibrated MMD test then follows from general results on permutation and bootstrap tests (Lehmann and Romano, 2022, Theorem 15.1; van der Vaart, 2000, Theorem 23.1). Another way to prove the consistency is by setting $\gamma = 1$ in both Appendix B.2.4 and Appendix B.2.6. Thus, (4.13) is met.

When conditioning on $\{\phi_f(Q_c^m, Q_h^\ell) = 1\}$, given $D(Q_c, Q_h) \leq \rho$, we know $\tau := \Pr(\phi_f(Q_c^m, Q_h^\ell) = 1) > 0$. Theorem 4.3.6 shows that the causality test with partial bootstrap obtains (B.1) if and only if $Q_c = Q_h$, or $Q_c \neq Q_h$ and $2(1-\gamma)D(Q_h, Q_c)\cos\beta < D(Q_c, Q_t)$. Similarly, with partial permutation, (B.1) stands if and only if $Q_c = Q_h$, or $Q_c \neq Q_h$ and $2\cos\beta\, D(Q_h, Q_c) < \left(\frac{1}{1-\gamma} + \lambda\right) D(Q_t, Q_c)$. Hence, all conditions in Proposition B.1.5 are met, thus we obtain (4.14) for both partial permutation and partial bootstrap.

Applying Theorem 4.3.16 then shows the claimed result in Corollary 4.3.18. □

B.2.11 Proof of Theorem 4.3.9 with converging sample size ratios

In this section we illustrate Remark 4.3.3, building on Appendix B.2.6, by showing how to adapt the argument to the case where the sample ratios merely converge to constants, rather than being exactly fixed. Concretely, we now assume $m/n \to c_1$ and $\ell/n \to c_2$. Correspondingly, we define

$$\gamma_n \coloneqq \frac{m}{m+\ell}, \qquad \lambda_n \coloneqq \frac{n}{m+n},$$

and denote by γ and λ the fixed ratios appearing in the main text. Under the convergence of the sample ratios we then have $\gamma_n \to \gamma$ and $\lambda_n \to \lambda$.

Proof. To adapt the proof in Appendix B.2.6 to this setting, it suffices to show, for example, that $T_{n,m,\ell} = \|(1-\gamma_n)(\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m}) - (\hat{\xi}_{t,n} - \hat{\xi}_{c,m})\|^2_{\mathcal{H}_k}$ converges to $\|(1-\gamma)(\xi_h - \xi_c) - (\xi_t - \xi_c)\|^2_{\mathcal{H}_k}$.

One could view $D^2(Q_c^m, Q_t^n, Q_h^\ell, \gamma_n)$ as a function of not only Q_c^m, Q_t^n, Q_h^ℓ, but also γ_n, and invoke Lemma B.2.1 directly. Here we instead pursue a simple decomposition which makes the technical steps explicit and can be reused when extending other results to converging ratios.

We rewrite $T_{n,m,\ell}$ as

$$\begin{aligned}
T_{n,m,\ell} &= \|(1-\gamma_n)(\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m}) - (\hat{\xi}_{t,n} - \hat{\xi}_{c,m})\|^2_{\mathcal{H}_k} \\
&= \|(1-\gamma_n)(\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m}) - (1-\gamma)(\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m}) \\
&\qquad + (1-\gamma)(\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m}) - (\hat{\xi}_{t,n} - \hat{\xi}_{c,m})\|^2_{\mathcal{H}_k} \\
&= \| \underbrace{(\gamma - \gamma_n)(\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m})}_{E_n} + \underbrace{(1-\gamma)(\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m}) - (\hat{\xi}_{t,n} - \hat{\xi}_{c,m})}_{F_n} \|^2_{\mathcal{H}_k}.
\end{aligned}$$

The term F_n is exactly the finite sample expression appearing in the fixed-ratio case (with γ fixed).

Therefore

$$\|E_n\|_{\mathcal{H}_k} = |\gamma - \gamma_n| \, \|\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m}\|_{\mathcal{H}_k}.$$

Under Assumption 4.2.2, $\mathbb{E}_R \sqrt{k(X,X)} \leq \left(\mathbb{E}_R k(X,X)\right)^{1/2} < \infty$ for $R = Q_h, Q_c$, so the embeddings ξ_h, ξ_c have finite norm. By the law of large numbers in the Hilbert space, $\hat{\xi}_{h,\ell} \to \xi_h$ and $\hat{\xi}_{c,m} \to \xi_c$ almost surely. Hence $\|\hat{\xi}_{h,\ell} - \hat{\xi}_{c,m}\|_{\mathcal{H}_k} \to \|\xi_h - \xi_c\|_{\mathcal{H}_k} < \infty$ almost surely, and since $|\gamma - \gamma_n| \to 0$, we obtain $\|E_n\|_{\mathcal{H}_k} \to 0$ almost surely.

From the fixed ratio analysis in Appendix B.2.6 we know that $F_n \to (1-\gamma)(\xi_h - \xi_c) - (\xi_t - \xi_c)$ in $\mathcal{H}_k$, hence $\|F_n\|_{\mathcal{H}_k}$ is almost surely bounded for all large n. Together with $\|E_n\|_{\mathcal{H}_k} \to 0$ this implies

$$\left| T_{n,m,\ell} - \|F_n\|^2_{\mathcal{H}_k} \right| \to 0$$

almost surely. Finally, since we already established in the fixed-ratio case that

$$\|F_n\|^2_{\mathcal{H}_k} \longrightarrow \|(1-\gamma)(\xi_h - \xi_c) - (\xi_t - \xi_c)\|^2_{\mathcal{H}_k},$$

we conclude that

$$T_{n,m,\ell} \longrightarrow \|(1-\gamma)(\xi_h - \xi_c) - (\xi_t - \xi_c)\|^2_{\mathcal{H}_k},$$

as claimed. □

B.3 Additional experiment results on normal approximation

We provide additional experimental results to examine the normal approximation. As discussed in Section 4.3.2.2, the linear kernel is unbounded, so its asymptotic behaviour emerges only when the sample size is sufficiently large.

In Figure B.1, the approximation accuracy with the normal approximation method improves with larger sample sizes. As shown in Figure B.2, which uses the same mean shift setting as Figure 4.3, the normal approximation underperforms relative to the partial bootstrap when the sample size is small, leading to reduced statistical power.

B.4 Extending Equivalence-TTP to U-statistic estimators

We discuss how to extend our proposed TTP framework when the U-statistic (4.2) is used to estimate MMD. Throughout this work, we have relied on the V-statistic estimator (4.1) because it is guaranteed to be non-negative. This property is essential in the fusion test, namely the equivalence MMD test described in Section 4.3.1. Specifically, the test statistic in the fusion test is

$$\Delta^f_{m,\ell} \coloneqq \rho - D(Q^m_c, Q^\ell_h),$$

which requires computing the *non-squared* MMD $D(Q^m_c, Q^\ell_h)$. With the V-statistic (4.1), this is computed by simply taking the square root, which is always well-defined due to non-negativity. However, with a U-statistic (4.2), the MMD estimate can take negative values, particularly when the current control Q_c and the historical control Q_h are similar so that $D(Q_h, Q_c) \approx 0$,

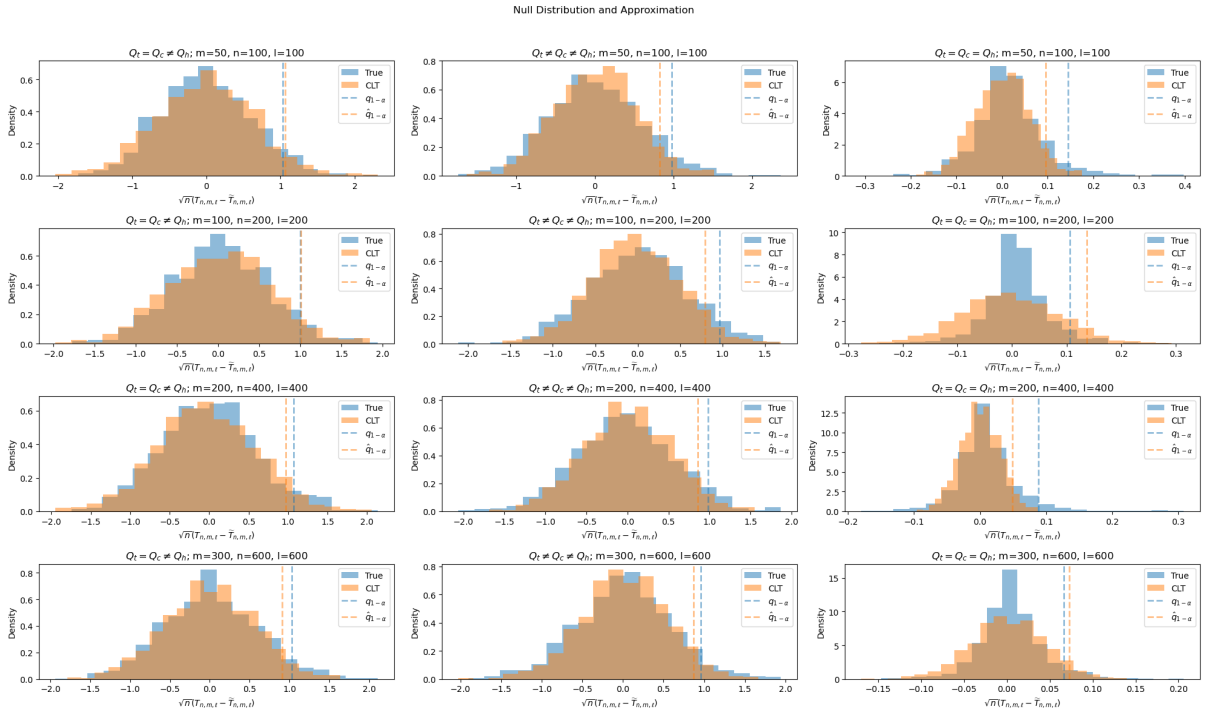


Figure B.1: Comparison of the true and approximate distributions of the test statistic under partial bootstrap using normal approximation, under different sample sizes. Results from 1000 simulations. Within each simulation, $B = 1000$ bootstraps are performed.

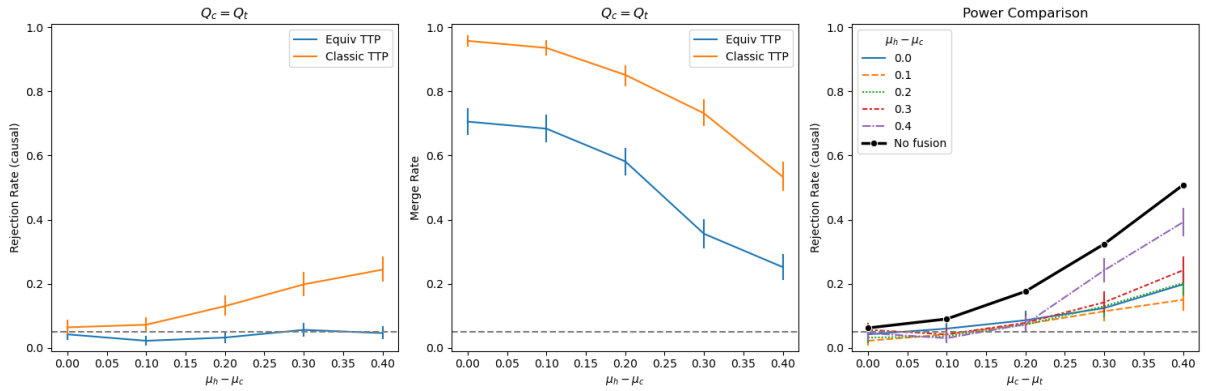


Figure B.2: ρ is fixed to be 0.4, RBF kernel, mean shift, normal approximation.

or when the sample sizes are small so that (4.2) has a large variance. In those cases, the square root required to compute $\Delta^f_{m,\ell}$ is not well-defined. A simple workaround is to never reject the null hypothesis (4.6) whenever the U-statistic is negative. However, this would prevent merging historical samples even when $D(Q_h, Q_c)$ is essentially zero, which would lead to loss of test power.

For these practical reasons, we adopted the V-statistics estimator in the fusion test. In contrast, the *causality test* can be readily extended to U-statistics, because neither the partial

bootstrap nor the partial permutation procedures require taking square roots of the MMD estimator.

Appendix C

Supplementary material for Chapter 5

C.1 Auxiliary results

This section collects auxiliary results that support the main theorems.

C.1.1 Distribution consistency from energy distance convergence

Theorem C.1.1 (Weak convergence from convergent energy distance)**.** *Let Q be a probability measure on $\mathbb{R}^d$, and let P^n be a random probability measure on $\mathbb{R}^d$. Suppose that P^n and Q satisfy Assumption 5.3.2, and that*

$$\mathrm{ED}(P^n, Q) \to 0 \qquad \textit{in probability.}$$

Then, for every bounded continuous function f,

$$\int f\, dP^n \to \int f\, dQ \qquad \textit{in probability.}$$

Equivalently, $P^n \rightsquigarrow Q$ in probability.

Proof. Fix a bounded continuous function f. By the subsequence principle for convergence in probability, it suffices to show that every subsequence of $\{\int f\, dP^n\}$ contains a further subsequence that converges almost surely to $\int f\, dQ$.

Let $\{P^{n_k}\}$ be an arbitrary subsequence. Since

$$\mathrm{ED}(P^n, Q) \to 0 \qquad \text{in probability,}$$

there exists a further subsequence, not relabeled, such that

$$\mathrm{ED}(P^{n_k}, Q) \to 0 \qquad \text{almost surely.}$$

Fix an outcome in the corresponding probability-one set. Then $\{P^{n_k}\}$ is a deterministic sequence of probability measures satisfying $\mathrm{ED}(P^{n_k}, Q) \to 0$.

By Assumption 5.3.2, for any $\varepsilon > 0$ there exists $R > 0$ such that $P^{n_k}(\|X\| > R) < \varepsilon$ uniformly in k, hence $\{P^{n_k}\}$ is tight. Hence, by Prokhorov's theorem (Billingsley, 2013, Theorem 5.1), every subsequence of $\{P^{n_k}\}$ contains a further subsequence that converges weakly to some probability measure. Let $\{P^{n_{k_j}}\}$ be such a subsequence, and write

$$P^{n_{k_j}} \rightsquigarrow \pi$$

for some probability measure π on $\mathbb{R}^d$.

Let $\varphi_P(t) = \int e^{i\langle t,x\rangle}\, dP(x)$ denote the characteristic function of a probability measure P. Since $P^{n_{k_j}} \rightsquigarrow \pi$, we have

$$\varphi_{P^{n_{k_j}}}(t) \to \varphi_\pi(t) \qquad \text{for every } t \in \mathbb{R}^d.$$

Moreover, Assumption 5.3.2 implies finite first moments, so the characteristic-function representation of the energy distance applies (Székely and Rizzo, 2013, Proposition 1):

$$\mathrm{ED}(P^{n_{k_j}}, Q) = \frac{1}{c_d} \int_{\mathbb{R}^d} \frac{\left|\varphi_{P^{n_{k_j}}}(t) - \varphi_Q(t)\right|^2}{\|t\|^{d+1}}\, dt.$$

Here $c_d > 0$ depends only on the dimension d (Székely and Rizzo, 2013). Since $\mathrm{ED}(P^{n_{k_j}}, Q) \to 0$, the right-hand side converges to zero. Therefore, after passing to a further subsequence if necessary,

$$\varphi_{P^{n_{k_j}}}(t) \to \varphi_Q(t) \qquad \text{for almost every } t \in \mathbb{R}^d.$$

Combining this with the pointwise convergence to φ_π yields

$$\varphi_\pi(t) = \varphi_Q(t) \qquad \text{for almost every } t \in \mathbb{R}^d.$$

Since characteristic functions are continuous, this implies $\varphi_\pi = \varphi_Q$ everywhere, and hence $\pi = Q$.

Thus every subsequence of $\{P^{n_k}\}$ has a further subsequence converging weakly to Q. Hence the deterministic sequence $\{P^{n_k}\}$ converges weakly to Q. In particular,

$$\int f\, dP^{n_k} \to \int f\, dQ \qquad \text{almost surely.}$$

Since the original subsequence was arbitrary, the subsequence principle for real-valued random variables implies

$$\int f\, dP^{n} \to \int f\, dQ \qquad \text{in probability.}$$

Because f was arbitrary, this proves $P^n \rightsquigarrow Q$ in probability. □

C.2 Main proofs

C.2.1 Proof of Lemma 5.2.6

Proof. When the conditions in Assumption 5.2.5 hold, we have

$$\begin{aligned}
p_{Y(x)}(y) &= p_{Y(x)|X_0}(y \,|\, x_0) && \text{(C.1)}\\
&= p_{Y(X_0,x_1)|X_0}(y \,|\, x_0) && \text{(C.2)}\\
&= \int_{\mathcal{Z}} p_{Y(X_0,x_1)|X_0Z}(y \,|\, x_0, z)\, p_{Z|X_0}(z \,|\, x_0)\, \mathrm{d}z && \text{(C.3)}\\
&= \int_{\mathcal{Z}} p_{Y(X_0,x_1)|X_0ZX_1}(y \,|\, x_0, z, x_1)\, p_{Z|X_0}(z \,|\, x_0)\, \mathrm{d}z && \text{(C.4)}\\
&= \int_{\mathcal{Z}} p_{Y|X_0ZX_1}(y \,|\, x_0, z, x_1)\, p_{Z|X_0}(z \,|\, x_0)\, \mathrm{d}z, && \text{(C.5)}
\end{aligned}$$

where (C.1) and (C.4) are from the unconfoundedness assumption, while (C.2) and (C.5) are based on the consistency assumption. Equation (C.3) is from the law of total probability; (C.1), (C.3) and (C.4) require the positivity assumption. □

C.2.2 Proof of Proposition 5.2.7

Proof. Since the energy score is a strictly proper scoring rule (Gneiting and Raftery, 2007), the objective function in (5.7) is minimised if and only if $g(\epsilon) \sim P_{ZX}$. By Assumption 5.2.4, such a model exists, we hence have $g^*(\epsilon) \sim P_{ZX}$. According to Proposition 1 of Shen and Meinshausen (2024) and Assumption 5.2.4, we have $f^*(x, h^*(z, x, \xi)) \sim P_{Y|Z=z,X=x}$ for P_{ZX}-almost every

(z, x). Because $\bar{\eta}$ and $\bar{\eta}'$ are formed from the same $\bar{X}$, the second term of (5.8) is an integrated conditional energy score over $\bar{X} \sim P_X$, so by strict propriety it is minimised if and only if

$$h^*(Z_x, x, \xi) \sim \mathcal{N}(0, I_{d_y}) \text{ when } Z_x \sim P_{Z \mid X_0 = x_0}, \tag{C.6}$$

for P_X-almost every $x = (x_0, x_1)$. By Assumption 5.2.4, a single (f', h') satisfies both the conditional fit and this reference-noise condition, so the two terms of (5.8) are minimised by a common solution.

By positivity (Assumption 5.2.5), the conditional fit established above for P_{ZX}-almost every (z, x) holds, for P_X-almost every x, at $P_{Z \mid X_0 = x_0}$-almost every z. The identification formula in Lemma 5.2.6,

$$p_{Y(x)}(y) = \int_{\mathcal{Z}} p_{Y \mid X_0 Z X_1}(y \mid x_0, z, x_1)\, p_{Z \mid X_0}(z \mid x_0)\, \mathrm{d}z.$$

suggests that to sample from $P_{Y(x)}$ for P_X-almost every $x = (x_0, x_1)$, we can

(i) first sample Z from $P_{Z \mid X_0 = x_0}$, and then

(ii) sample Y from $P_{Y \mid Z, X = x}$ by setting $\eta^* := h^*(Z, x, \xi)$ and applying $f^*(x, \eta^*)$.

According to (C.6), we know η^* follows the standard Gaussian for P_X-almost every x. Thus, we can simplify the above sampling procedure to directly sampling $\eta^* \sim \mathcal{N}(0, I_{d_y})$ and applying $f^*(x, \eta^*)$. This yields the last desired result that $f^*(x, \eta) \sim P_{Y(x)}$ for P_X-almost every x. □

C.2.3 Proof of Theorem 5.3.3

The energy distance $\mathrm{ED}(P, Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}}$ uses the mean embedding μ_P[16] in the RKHS $\mathcal{H}$ of the energy kernel $k(u, v) = \frac{1}{2}(\|u\| + \|v\| - \|u - v\|)$. On empirical inputs it has two estimators: the plug-in V-statistic $\mathrm{ED}_V \geq 0$, whose square root obeys the triangle inequality, and the unbiased U-statistic ED_U, the training loss, which can be negative. We write ED throughout, meaning ED_V on empirical inputs, and ED_U only for the training loss and the concentration bounds (Lemmas C.2.2 and C.2.3); the two differ by lower-order self-terms, absorbed below.

We first note a consequence of Assumption 5.3.2.

[16] The kernel mean embedding is written μ here; Chapters 4 and 6 write it ξ.

Lemma C.2.1 (Finite second moment)**.** *Suppose Assumption 5.3.2 holds for random variable $X \sim P_\theta$, $\forall \theta \in \Theta$ and $X \sim Q$. Then, there exists a constant $M > 0$ such that*

$$\mathbb{E}_{P_\theta}\|X\|^2 \leq M^2 \quad \textit{and} \quad \mathbb{E}_Q\|X\|^2 \leq M^2.$$

Proof of Lemma C.2.1. For $X \sim P$ with P a probability measure satisfying Assumption 5.3.2, we have

$$\mathbb{E}\|X\|^2 = \int_0^\infty 2u\Pr(\|X\| > u)\,\mathrm{d}u \leq 2C\int_0^\infty u\exp(-\kappa u)\mathrm{d}u = \frac{2C}{\kappa^2}.$$

The claim follows with $M^2 = \frac{2C}{\kappa^2}$. □

With Assumption 5.3.2 satisfied, we also have:

Lemma C.2.2 (Concentration inequality with energy distance)**.** *Let P be a probability measure on $X \subseteq \mathbb{R}^d$, P^n be the empirical measure obtained from n independently and identically distributed samples of P, i.e. $P^n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $x_i \overset{\text{i.i.d.}}{\sim} P$.*

For every $\delta > 0$, with probability at least $1 - \delta - q_{R_n}$:

$$\mathrm{ED}_U(P, P^n) < \frac{C_1 M}{\sqrt{n(1-q_{R_n})}} + 4R_n\sqrt{\frac{2}{n}\log\Big(\frac{1}{\delta}\Big)},$$

for some constant $C_1 > 0$, where $R_n = \frac{3\log n}{\kappa}$, $q_{R_n} := \Pr(\max_{1\leq i\leq n}\|x_i\| > R_n)$. We also have:

$$q_{R_n} \leq \frac{C}{n^2}.$$

To prove Lemma C.2.2, we first prove the following:

Lemma C.2.3 (Finite second moment of $\mathrm{ED}_U(P, P^n)$)**.** *For a probability measure P that satisfies Assumption 5.3.2 and Lemma C.2.1, where P^n is the empirical distribution, there exists a constant $C_0 > 0$, such that $\mathbb{E}[\mathrm{ED}_U(P, P^n)^2] \leq C_0 M^2/n$.*

Proof. We have

$$\mathrm{ED}_U(P, P^n) = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{X\sim P}\|X - x_i\| - \mathbb{E}_{X,X'\sim P}\|X - X'\| - \frac{1}{n(n-1)}\sum_{j=1}^n\sum_{\substack{k=1\\k\neq j}}^n \|x_j - x_k\|.$$

We write $u(x) = \mathbb{E}_{X\sim P}\|X - x\|$, $\nu = \mathbb{E}_{X,X'\sim P}\|X - X'\|$, $U_n = \frac{1}{n(n-1)}\sum_{j=1}^n\sum_{\substack{k=1\\k\neq j}}^n \|x_j - x_k\|$.

Clearly, $\mathbb{E}[U_n] = \nu$. We can thus write

$$\mathrm{ED}_U(P, P^n) = \underbrace{\frac{2}{n}\sum_{i=1}^{n}\big(u(x_i) - \nu\big)}_{D_1} + \underbrace{(\nu - U_n)}_{D_2}.$$

For D_1:

$$\mathrm{Var}(D_1) = \frac{4}{n}\,\mathrm{Var}\left(u(X) - \nu\right) = \frac{4}{n}\,\mathrm{Var}\left(u(X)\right).$$

By Jensen's inequality, we have $u(x)^2 \leq \mathbb{E}_{X\sim P}\|X - x\|^2$. Noting further that $\|X - x\|^2 \leq 2\|X\|^2 + 2\|x\|^2$, and applying Lemma C.2.1, we get

$$\mathbb{E}_{X\sim P}[u(X)^2] \leq 4M^2,$$

so that

$$\mathrm{Var}(u(X)) \leq \mathbb{E}[u(X)^2] \leq 4M^2$$

and

$$\mathrm{Var}(D_1) \leq \frac{4 \cdot 4M^2}{n} = \frac{16M^2}{n}.$$

For D_2, using the standard variance formula for U-statistics of order two (Hoeffding, 1948, equation 5.18), we obtain

$$\mathrm{Var}(U_n) = \frac{4}{n}\sigma^2 + O\Big(\frac{1}{n^2}\Big),$$

where $\sigma^2 = \mathrm{Var}\left[\mathbb{E}_{X'\sim P}k(X, X')\right] = \mathrm{Var}(u(X)) \leq 4M^2$. Thus,

$$\mathrm{Var}(D_2) = \mathrm{Var}(U_n) \leq \frac{16M^2}{n} + O\Big(\frac{1}{n^2}\Big),$$

which can be expressed as

$$\mathrm{Var}(D_2) \leq \frac{C_2M^2}{n},$$

for some constant $C_2 > 0$. We can therefore obtain

$$\mathbb{E}[\mathrm{ED}_U(P, P^n)^2] \leq 2\,\mathrm{Var}(D_1) + 2\,\mathrm{Var}(D_2),$$

since $\mathbb{E}D_1 = \mathbb{E}D_2 = 0$.

Thus we obtain

$$\mathbb{E}[\mathrm{ED}_U(P, P^n)^2] \leq \frac{C_0 M^2}{n}$$

for some constant $C_0 > 0$. □

Proof of Lemma C.2.2. Let $\mathcal{A}_{R_n}$ denote the event $\mathcal{A}_{R_n} := \{\max_{1\leq i\leq n} \|X_i\| \leq R_n\}$. On $\mathcal{A}_{R_n}$, any sample points x, x' satisfy $\|x\|, \|x'\| \leq R_n$, hence

$$\|x - x'\| \leq \|x\| + \|x'\| \leq 2R_n.$$

With Assumption 5.3.2, we have

$$\begin{aligned} q_{R_n} = \Pr(\mathcal{A}_{R_n}^c) = \Pr(\exists i, \|X_i\| > R_n) &\leq nC\exp(-\kappa R_n) \\ &= nC\exp\Big(-\kappa\frac{3\log n}{\kappa}\Big) \\ &= C/n^2. \end{aligned}$$

We adapt the bounded-difference argument of Briol et al. (2019). Denote $h(x_1, \ldots, x_n) = \mathrm{ED}_U(P, P^n)$, explicitly we have

$$h(x_1, \ldots, x_n) = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{X\sim P}\|X - x_i\| - \mathbb{E}_{X,X'\sim P}\|X - X'\| - \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\|x_j - x_k\|.$$

As $\mathcal{A}_{R_n} = \bigcap_{i=1}^n \{\|X_i\| \leq R_n\}$ is a product event, conditioning on it keeps the X_i independent and supported in $\{\|x\| \leq R_n\}$; we apply McDiarmid in this conditional distribution, where the bounded difference need only hold for x_i, x_i' in the ball. Only the first and last term can be affected by changing a single point from x_i to x_i'. On the event $\mathcal{A}_{R_n}$, the first term changes by at most

$$\frac{2}{n}\Big|\mathbb{E}\|X - x_i\| - \mathbb{E}\|X - x_i'\|\Big| \leq \frac{2}{n}\|x_i - x_i'\| \leq \frac{2}{n}(\|x_i\| + \|x_i'\|) \leq \frac{4R_n}{n},$$

using $|\mathbb{E}Z| \leq \mathbb{E}|Z|$ and $\|x_i\|, \|x_i'\| \leq R_n$. The last term $\frac{1}{n(n-1)}\sum_{j\neq k}\|x_j - x_k\|$ changes only in the $2(n-1)$ pairs involving x_i, each by at most $\|x_i - x_i'\| \leq 2R_n$, hence by at most $\frac{1}{n(n-1)} \cdot 2(n-1)\cdot 2R_n = \frac{4R_n}{n}$. The bounded-difference constant is therefore $8R_n/n$. By McDiarmid's inequality (McDiarmid, 1989), we get that for any $\epsilon > 0$,

$$P\Big(\operatorname{ED}_U(P,P^n) - \mathbb{E}\big[\operatorname{ED}_U(P,P^n)\,|\,\mathcal{A}_{R_n}\big] \geq \epsilon \,\big|\, \mathcal{A}_{R_n}\Big) \leq \exp\big(-2\epsilon^2/\{n(8R_n/n)^2\}\big)$$
$$= \exp\left(\frac{-n\epsilon^2}{32{R_n}^2}\right).$$

Setting this bound to δ gives $\epsilon = 4R_n\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}$. Thus, we get that on the event $\mathcal{A}_{R_n}$

$$P\left(\operatorname{ED}_U(P,P^n) - \mathbb{E}\big[\operatorname{ED}_U(P,P^n)\,|\,\mathcal{A}_{R_n}\big] < 4R_n\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \geq 1-\delta.$$

Observe that

$$\begin{aligned}
\mathbb{E}\Big[\operatorname{ED}_U(P,P^n)|\mathcal{A}_{R_n}\Big] &= \frac{\mathbb{E}[\operatorname{ED}_U(P,P^n)\mathbb{1}(\mathcal{A}_{R_n})]}{\Pr(\mathcal{A}_{R_n})} \\
&\leq \frac{\sqrt{\mathbb{E}[\operatorname{ED}_U(P,P^n)^2]}\sqrt{\Pr(\mathcal{A}_{R_n})}}{\Pr(\mathcal{A}_{R_n})} \\
&\leq \frac{\sqrt{\mathbb{E}[\operatorname{ED}_U(P,P^n)^2]}}{\sqrt{1-q_{R_n}}} \\
&\leq \frac{C_1 M}{\sqrt{n(1-q_{R_n})}};
\end{aligned}$$

here the last inequality is from Lemma C.2.3.

We thus obtain that conditioning on event $\mathcal{A}_{R_n}$:

$$P\left(\operatorname{ED}_U(P,P^n) < \frac{C_1 M}{\sqrt{n(1-q_{R_n})}} + 4R_n\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \geq 1-\delta,$$

and thus the unconditional probability is

$$P\left(\operatorname{ED}_U(P,P^n) < \frac{C_1 M}{\sqrt{n(1-q_{R_n})}} + 4R_n\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \geq (1-\delta)\Pr(\mathcal{A}_{R_n}) = (1-\delta)(1-q_{R_n}).$$

The claim follows. □

Lemma C.2.2 bounds $\operatorname{ED}_U(P,P^n)$; it transfers to $\operatorname{ED}(P,P^n)$ because the two differ by U_n/n, with $U_n = \frac{1}{n(n-1)}\sum_{j\neq k}\|x_j - x_k\|$. On $\mathcal{A}_{R_n}$ each $\|x_j - x_k\| \leq 2R_n$, so $U_n/n \leq 2R_n/n$, lower-order than the $R_n\sqrt{n^{-1}\log(1/\delta)}$ term.

Lemma C.2.4 (Energy-distance perturbation)**.** *For probability measures P, Q, R on $\mathbb{R}^d$ with finite first moments,*

$$|\operatorname{ED}(P,Q) - \operatorname{ED}(P,R)| \leq \big(\sqrt{\operatorname{ED}(P,Q)} + \sqrt{\operatorname{ED}(P,R)}\big)\sqrt{\operatorname{ED}(Q,R)}.$$

Proof. With $\mathrm{ED}(P,Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2$ in the RKHS of the energy kernel,

$$\mathrm{ED}(P,Q) - \mathrm{ED}(P,R) = \langle \mu_R - \mu_Q,\, 2\mu_P - \mu_Q - \mu_R \rangle.$$

Cauchy–Schwarz with $\|2\mu_P - \mu_Q - \mu_R\| \le \|\mu_P - \mu_Q\| + \|\mu_P - \mu_R\|$ concludes the claim. □

Lemma C.2.5 (Uniform convergence over a bounded parameter set)**.** *Let $\Theta_0 \subseteq \Theta$ be bounded with closure $\overline{\Theta_0}$, and suppose Assumption 5.3.2 holds and, for each ν, the map $\theta \mapsto g_\theta(\nu)$ is Lipschitz on $\overline{\Theta_0}$, that is, $\|g_\theta(\nu) - g_{\theta'}(\nu)\| \le L(\nu)\,\|\theta - \theta'\|$, where $L(\nu)$, not necessarily bounded, satisfies $\mathbb{E}[\sqrt{L(\nu)}] < \infty$. With $P_\theta^m = \frac{1}{m}\sum_{i=1}^m \delta_{g_\theta(\nu_i)}$ for $\nu_1, \dots, \nu_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$,*

$$\sup_{\theta \in \overline{\Theta_0}} \mathrm{ED}(P_\theta, P_\theta^m) \to 0 \quad \textit{in probability.}$$

Proof. $\overline{\Theta_0}$ is compact, and by Lemma C.2.2, $\mathrm{ED}(P_\theta, P_\theta^m) \to 0$ in probability for each fixed θ, at a rate set by the common constants of Assumption 5.3.2. It remains to make this convergence uniform over θ.

Write $D_m(\theta) = \sqrt{\mathrm{ED}(P_\theta, P_\theta^m)}$. As $\sqrt{\mathrm{ED}} = \|\mu_\cdot - \mu_\cdot\|_{\mathcal{H}}$ is a metric,

$$|D_m(\theta) - D_m(\theta')| \le \sqrt{\mathrm{ED}(P_\theta, P_{\theta'})} + \sqrt{\mathrm{ED}(P_\theta^m, P_{\theta'}^m)}.$$

For the energy kernel, $\|k(\cdot, a) - k(\cdot, b)\|_{\mathcal{H}} = \sqrt{\|a - b\|}$, so

$$\sqrt{\mathrm{ED}(P_\theta, P_{\theta'})} \le \mathbb{E}_\nu \sqrt{\|g_\theta(\nu) - g_{\theta'}(\nu)\|} \le \sqrt{\|\theta - \theta'\|}\, \mathbb{E}\Big[\sqrt{L(\nu)}\Big],$$

the last step by the Lipschitz condition; the same bound, with $\mathbb{E}_\nu$ replaced by $\frac{1}{m}\sum_{i=1}^m$, holds for $\sqrt{\mathrm{ED}(P_\theta^m, P_{\theta'}^m)}$, since the two empirical measures share the draws $\nu_1, \dots, \nu_m$. Hence $|D_m(\theta) - D_m(\theta')| \le A_m \sqrt{\|\theta - \theta'\|}$ with $A_m = \mathbb{E}[\sqrt{L(\nu)}] + \frac{1}{m}\sum_{i=1}^m \sqrt{L(\nu_i)}$, and $A_m \to 2\,\mathbb{E}[\sqrt{L(\nu)}]$ a.s., so $A_m = O_P(1)$.

Fix $\varepsilon, \gamma > 0$ and take $A < \infty$ with $\Pr(A_m \le A) \ge 1 - \gamma$ for all large m. On $\{A_m \le A\}$ the modulus bound gives $|D_m(\theta) - D_m(\theta')| \le \varepsilon$ whenever $\|\theta - \theta'\| \le (\varepsilon/A)^2$; covering $\overline{\Theta_0}$ by finitely many such balls, centred at $\theta_1, \dots, \theta_N$,

$$\sup_{\theta \in \overline{\Theta_0}} D_m(\theta) \le \max_{1 \le j \le N} D_m(\theta_j) + \varepsilon.$$

The finite maximum tends to 0 in probability, and ε, γ are arbitrary, so $\sup_\theta D_m(\theta) \to 0$, and

hence $\sup_\theta \mathrm{ED}(P_\theta, P_\theta^m) \to 0$, in probability. □

Theorem C.2.6 (Generalisation bounds)**.** *Write*

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta\in\Theta} \mathrm{ED}_U(P_\theta, Q^n) \quad \textit{and} \quad \hat{\theta}_{m,n} = \operatorname*{argmin}_{\theta\in\Theta} \mathrm{ED}_U(P_\theta^m, Q^n).$$

Under Assumption 5.3.1 and the assumptions of Lemma C.2.2, there exists a deterministic sequence $K_n = O(1 + R_n)$ such that $\mathrm{ED}(P_\theta, Q)$ and $\mathrm{ED}(P_\theta, Q^n)$ are bounded by K_n, uniformly in $\theta \in \Theta$, on the truncation event $\mathcal{A}_{R_n}$. With probability at least $1 - \delta - q_{R_n}$,

$$\mathrm{ED}(P_{\hat{\theta}_n}, Q) - \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q) \;\le\; 4\sqrt{K_n}\sqrt{\frac{C_1 M}{\sqrt{n(1-q_{R_n})}} + 4R_n\sqrt{\frac{2}{n}\log\tfrac{1}{\delta}}}.$$

For the estimator $\hat{\theta}_{m,n}$, under conditions (c)–(d) of Theorem 5.3.3,

$$\mathrm{ED}(P_{\hat{\theta}_{m,n}}, Q) - \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q) \to 0 \qquad \textit{in probability as } m, n \to \infty.$$

Proof of Theorem C.2.6. We first record the bound K_n. By Lemma C.2.1, $\mathrm{ED}(P_\theta, Q) \le 2\,\mathbb{E}\|X - Y\| \le 2(\mathbb{E}_{P_\theta}\|X\| + \mathbb{E}_Q\|Y\|) \le 4M$ for every θ; and on $\mathcal{A}_{R_n}$, where $\frac{1}{n}\sum_i \|Y_i\| \le R_n$, the same bound gives $\mathrm{ED}(P_\theta, Q^n) \le 2(M + R_n)$. Hence $K_n = O(1 + R_n)$.

On $\mathcal{A}_{R_n}$, decompose

$$\mathrm{ED}(P_{\hat{\theta}_n}, Q) - \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q) = T_1 + T_2 + T_3,$$

with $T_1 = \mathrm{ED}(P_{\hat{\theta}_n}, Q) - \mathrm{ED}(P_{\hat{\theta}_n}, Q^n)$, $T_2 = \mathrm{ED}(P_{\hat{\theta}_n}, Q^n) - \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q^n)$ and $T_3 = \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q^n) - \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q)$. The optimality of $\hat{\theta}_n$ gives $T_2 = 0$: $\mathrm{ED}_U(P_\theta, Q^n)$ and $\mathrm{ED}(P_\theta, Q^n)$ differ only by the Q^n self-term, constant in θ, so they share the minimiser $\hat{\theta}_n$. Lemma C.2.4 bounds $|T_1| \le 2\sqrt{K_n}\sqrt{\mathrm{ED}(Q, Q^n)}$; the same lemma combined with $\inf_\theta f(\theta) - \inf_\theta g(\theta) \le \sup_\theta(f(\theta) - g(\theta))$ gives $|T_3| \le 2\sqrt{K_n}\sqrt{\mathrm{ED}(Q, Q^n)}$. Hence

$$\mathrm{ED}(P_{\hat{\theta}_n}, Q) - \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q) \;\le\; 4\sqrt{K_n}\sqrt{\mathrm{ED}(Q, Q^n)} \qquad \text{on } \mathcal{A}_{R_n},$$

and Lemma C.2.2 delivers the first display in the statement.

For $\hat{\theta}_{m,n}$, condition (c) makes Θ compact, so $\hat{\theta}_{m,n} \in \Theta$. By optimality of $\hat{\theta}_{m,n}$,

$$\mathrm{ED}(P_{\hat{\theta}_{m,n}}, Q) - \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q) \le 2 \sup_{\theta\in\Theta} \big|\, \mathrm{ED}_U(P_\theta^m, Q^n) - \mathrm{ED}(P_\theta, Q)\big|.$$

Here $\mathrm{ED}_U(P_\theta^m, Q^n)$ and $\mathrm{ED}(P_\theta^m, Q^n)$ differ by the self-term correction, $o_P(1)$ uniformly in θ under conditions (c)–(d), so we may pass to ED and apply the triangle inequality. Write $d = \sqrt{\mathrm{ED}}$; then $|d(P_\theta^m, Q^n) - d(P_\theta, Q)| \le d(P_\theta^m, P_\theta) + d(Q, Q^n)$, so

$$\big|\, \mathrm{ED}(P_\theta^m, Q^n) - \mathrm{ED}(P_\theta, Q)\big| \le \big(d(P_\theta^m, P_\theta) + d(Q, Q^n)\big)\big(d(P_\theta^m, Q^n) + d(P_\theta, Q)\big).$$

The first factor is $o_P(1)$ uniformly in θ, by Lemma C.2.5 (with $\Theta_0 = \Theta$) and Lemma C.2.2. For the second, $\sup_{\theta\in\Theta} d(P_\theta, Q) = O(1)$ by Lemma C.2.1, whence $d(P_\theta^m, Q^n) \le d(P_\theta^m, P_\theta) + d(P_\theta, Q) + d(Q, Q^n)$ is $O_P(1)$ uniformly in θ. Hence $\sup_{\theta\in\Theta} |\, \mathrm{ED}(P_\theta^m, Q^n) - \mathrm{ED}(P_\theta, Q)| = o_P(1)$, and the excess risk tends to 0 in probability. □

We now prove Theorem 5.3.3.

Proof of Theorem 5.3.3. By Theorem C.2.6 applied with $\delta = 1/n^2$,

$$\Pr\{\, \mathrm{ED}(P_{\hat{\theta}_n}, Q) - \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q) \ge 4\sqrt{K_n}\,\sqrt{B_n}\} \le \tfrac{1}{n^2} + q_{R_n} = O(\tfrac{1}{n^2}),$$

where $B_n = \frac{C_1 M}{\sqrt{n(1-q_{R_n})}} + 8R_n\sqrt{\frac{\log n}{n}} = O(\frac{(\log n)^{3/2}}{\sqrt{n}})$. Since $K_n = O(\log n)$, we have $\sqrt{K_n B_n} \to 0$, and the Borel–Cantelli lemma then gives, almost surely,

$$\mathrm{ED}(P_{\hat{\theta}_n}, Q) - \inf_{\theta\in\Theta} \mathrm{ED}(P_\theta, Q) \;\to\; 0.$$

When $P_{\theta^*} \overset{d}{=} Q$, the infimum equals 0, so $\mathrm{ED}(P_{\hat{\theta}_n}, Q) \xrightarrow{p} 0$, and Theorem C.1.1 gives $P_{\hat{\theta}_n} \rightsquigarrow Q$ in probability. For $\hat{\theta}_{m,n}$, under conditions (c)–(d) of Theorem 5.3.3, the (m, n) part of Theorem C.2.6 gives $\mathrm{ED}(P_{\hat{\theta}_{m,n}}, Q) \xrightarrow{p} 0$ under the same specification, and Theorem C.1.1 again yields $P_{\hat{\theta}_{m,n}} \rightsquigarrow Q$ in probability. □

C.2.4 Proof of Corollary 5.3.4

Proof. For the first claim, apply Theorem 5.3.3 to (5.9) with target P_{ZX}; Theorem C.1.1 yields $\hat{g}(\epsilon) \rightsquigarrow P_{ZX}$ in probability.

The second and third claims both follow once the conclusions of Proposition 5.2.7 are shown to hold, off *common* null sets, for *every* population minimiser rather than for a single solution.

The proof of Proposition 5.2.7 uses strict propriety, hence characterises any minimiser. Writing Θ^* for the set of population minimisers, each $\theta^* \in \Theta^*$ satisfies $f_{\theta^*}\big(x, h_{\theta^*}(z, x, \xi)\big) \sim P_{Y \mid Z=z, X=x}$ off some P_{ZX}-null set, and $h_{\theta^*}(Z_x, x, \xi) \sim \mathcal{N}(0, I_{d_y})$ with $Z_x \sim P_{Z \mid X_0=x_0}$ off some P_X-null set, both sets depending on θ^*. As Θ is compact, Θ^* is separable; fix a countable dense $\{\theta_k\} \subset \Theta^*$ and let N_{ZX} and N_X be the unions of the corresponding P_{ZX}- and P_X-null sets, themselves null. For any $\theta^* \in \Theta^*$ take $\theta_k \to \theta^*$; since $\theta \mapsto P_\theta$ is continuous and the laws indexed by the θ_k agree, off N_{ZX} and N_X, with the same true conditional and standard Gaussian, uniqueness of weak limits gives the same two properties for θ^* off the common null sets N_{ZX} and N_X.

We assume the empirical objective (5.10) converges uniformly over Θ to its population counterpart; when X_0 is present, this also requires the auxiliary model used to draw Z from $P_{Z \mid X_0}$ to be consistent. Under the compactness, componentwise Lipschitz, and uniform tail conditions of Theorem 5.3.3 together with this consistency, the excess population risk of $\hat{\theta}$ tends to zero, so by the compactness of Θ and continuity of the population risk, $\hat{\theta}$ approaches Θ^*. For the second claim, fix $(z, x) \notin N_{ZX}$: the conditional fit then holds for every $\theta^* \in \Theta^*$, so by continuity of $\theta \mapsto P_\theta$, along any subsequence with $\hat{\theta} \to \theta^*$ we have $\hat{f}(x, \hat{h}(z, x, \xi)) \rightsquigarrow P_{Y \mid Z=z, X=x}$; as this limit is the same for all $\theta^* \in \Theta^*$, it holds in probability for P_{ZX}-almost every (z, x).

For the third claim, fix $x \notin N_X$ and outside the positivity-null set. By positivity (Assumption 5.2.5) the conditional fit holds at $P_{Z \mid X_0=x_0}$-almost every z, and the calibration gives $h_{\theta^*}(Z_x, x, \xi) \sim \mathcal{N}(0, I_{d_y})$; substituting both into the identification formula of Lemma 5.2.6 yields $f_{\theta^*}(x, \eta) \sim P_{Y(x)}$ for every $\theta^* \in \Theta^*$. Continuity of $\theta \mapsto P_\theta$ then gives, along any subsequence with $\hat{\theta} \to \theta^*$, $\hat{f}(x, \eta) \rightsquigarrow f_{\theta^*}(x, \eta) \sim P_{Y(x)}$; since this limit is the same for all $\theta^* \in \Theta^*$, $\hat{f}(x, \eta) \rightsquigarrow P_{Y(x)}$ in probability for P_X-almost every x.

□

C.2.5 Proof of Proposition 5.4.2

Proof. Since $\widetilde{\eta}$ is a continuous random variable, there exists a strictly monotone function $\widetilde{f}_1$ such that $\widetilde{\eta} \overset{d}{=} \widetilde{f}_1(\widetilde{\xi})$ where $\widetilde{\xi} \sim \mathcal{N}(0, 1)$. Let f^*, f_1^* denote the element of an optimal frengression model such that

$$f^*(x + f_1^*(\widetilde{\xi})) \sim P_{Y(x)},$$

for P_X-almost every $x \in \mathcal{X}$ by Proposition 5.2.7. Under Assumption 5.4.1 both sides depend continuously on x and P_X has full support on $\mathcal{X}$, so the relation extends to every $x \in \mathcal{X}$. This implies

$$\widetilde{f}(x + \widetilde{f}_1(\widetilde{\xi})) \stackrel{d}{=} f^*(x + f_1^*(\widetilde{\xi})),$$

for all $x \in \mathcal{X}$. Taking medians in the last display gives $\widetilde{f}(x) = f^*(x)$ on $\mathcal{X}$, so $\widetilde{f}$ and f^* are increasing or decreasing together; replacing Y by $-Y$ in the decreasing case, and choosing $\widetilde{f}_1$ and f_1^* increasing, which is possible because $\widetilde{\xi} \stackrel{d}{=} -\widetilde{\xi}$, we may take all four maps strictly increasing.

Since both sides are strictly increasing functions of the same $\widetilde{\xi} \sim \mathcal{N}(0,1)$, equality in distribution implies that they coincide as functions: for all $x \in \mathcal{X}$ and $u \in \mathbb{R}$,

$$\widetilde{f}(x + \widetilde{f}_1(u)) = f^*(x + f_1^*(u)). \tag{C.7}$$

The maps $\widetilde{f}, \widetilde{f}_1, f^*, f_1^*$ are continuously differentiable by Assumption 5.4.1; taking the derivative w.r.t. x and u respectively yields

$$\widetilde{f}'(x + \widetilde{f}_1(u)) = f^{*\prime}(x + f_1^*(u))$$
$$\widetilde{f}'(x + \widetilde{f}_1(u))\widetilde{f}_1'(u) = f^{*\prime}(x + f_1^*(u))f_1^{*\prime}(u).$$

Using the first equation to replace $f^{*\prime}(x + f_1^*(u))$ in the second gives

$$\widetilde{f}'(x + \widetilde{f}_1(u))(\widetilde{f}_1'(u) - f_1^{*\prime}(u)) = 0.$$

Since $\widetilde{f}$ is strictly increasing, its derivative cannot be zero everywhere on the interval $\{x + \widetilde{f}_1(u) : x \in [x_{\min}, x_{\max}]\}$ – otherwise $\widetilde{f}$ would be constant there. Hence for each u there is an x with $\widetilde{f}'(x + \widetilde{f}_1(u)) > 0$, so $\widetilde{f}_1'(u) = f_1^{*\prime}(u)$. As this holds for every u, integration gives $\widetilde{f}_1(u) = f_1^*(u) + c$ for a constant c. Since $f_1^*(0) = 0$ by normalisation and $\widetilde{f}_1(0) = 0$ because $\widetilde{\eta}$ has median 0, we obtain $c = 0$ and $\widetilde{f}_1 = f_1^*$.

Then according to (C.7), we have

$$\widetilde{f}(\widetilde{x}) = f^*(\widetilde{x}),$$

for all $\widetilde{x} \in [x_{\min} - \widetilde{\eta}_0, x_{\max} + \widetilde{\eta}_0]$. Note that $\widetilde{f}(x)$ and $f^*(x)$ are exactly the true and modelled conditional medians, respectively; the desired result then follows. □

C.3 Hyperparameters in frengression

All hyperparameters are kept fixed across experiments, except for the number of epochs, which is chosen so that the loss converges. The hyperparameters include

- learning_rate: The learning rate is set to be 10^{-4}.
- hidden_layer: The number of hidden layers in the neural network is set to be 3.
- hidden_dim: The size of neurons per hidden layer is set to be 100.

C.4 Experiment details

All experiments were conducted on a MacBook with an Apple M3 chip, 8-core CPU, and 32GB RAM.

C.4.1 Binary intervention

In this setting, for each replication, we train frengression for 1000 epochs.

For the AIPW estimator, we fit the propensity score model with logistic regression; because the true propensity scores are linear in the covariates, this specification performs well. We model the outcomes separately using random forests.

Next, we compare frengression to two advanced deep neural network approaches for causal inference, CausalEGM and Dragonnet. A comprehensive list of CausalEGM's hyper-parameters is available at https://causalegm.readthedocs.io/en/latest/tutorial_py.html. Because the search space is extensive, we adopt the configuration recommended for the binary treatment experiments in the original paper. The settings of key hyperparameters are given in Table C.1; all other quantities remain at their default values. Although the authors of CausalEGM suggest 30,000 epochs especially for continuous treatment setting, for fair comparison we train CausalEGM for 1000 epochs in our experiments.

Hyperparameter	Value
Latent covariate dimensions: z_dims	$[1, 1, 1, 1]$
Learning rate: lr	2×10^{-4}
Reconstruction loss weight: alpha	1
Round trip loss weight: beta	1
Gradient penalty loss weight: gamma	10
Generator units: g_units	[64, 64, 64, 64, 64]
Encoder units: e_units	[64, 64, 64, 64, 64]
F network units: f_units	[64, 32, 8]
H network units: h_units	[64, 32, 8]
Discriminator units in latent space: dz_units	[64, 32, 8]
Discriminator units in covariate space: dv_units	[64, 32, 8]

Table C.1: Values of key hyperparameters in the CausalEGM model (Python implementation).

For Dragonnet, in line with its requirements, we tuned the hyperparameters with `optuna`. The search range is shown in Table C.2. We conduct 30 hyperparameter search trials, selecting the best configuration by validation performance on a hold-out set of 400 samples, and then apply the resulting model to an independent test set of 400 samples to obtain the reported estimates.

Hyperparameter	Value
Learning rate: lr	$\text{logUniform}(10^{-5}, 10^{-2})$
Weights decay rate: wd	$\text{logUniform}(10^{-5}, 10^{-2})$
Batch size: bs	$\{32, 128, 256\}$
Number of iterations: epochs	Grid search in between 200 and 600
Number of neurons in shared layers: shared_hidden	Grid search in between 50 and 200
Number of neurons in shared layers: outcome_hidden	Grid search in between 50 and 200

Table C.2: Search ranges of key hyperparameters in the Dragonnet model (Python implementation).

C.4.2 Longitudinal analysis

C.4.2.1 Synthetic data

We provide the details of the synthetic data generation in Section 5.6.3.

C.4.2.2 Survival analysis: comparison with `dltmle`

In our implementation of `dltmle` (Shirakawa, 2024) in Section 5.6.3, we tuned the hyperparameters using their built in `dltmle.tune` function within the candidates in Table C.5.

	Setting 1	**Setting 2**	**Setting 3**
C	Bernoulli(0.5)	Exp(1)	Bernoulli(0.5)
$Z_t \mid \overline{X}_{t-1}, \overline{Z}_{t-1}, C$	$\mathcal{N}(-0.5 + 0.5X_{t-1} + 0.25C,\ 0.5)$	$\mathcal{N}(0.7Z_{t-1} + 0.2X_{t-1},\ 1)$	$\mathcal{N}(-0.5 + 0.5X_{t-1} + 0.25C,\ 0.5)$
$X_t \mid \overline{Z}_t, \overline{X}_{t-1}, C$	Bernoulli $\big(\text{expit}(0.2Z_t + 0.1C)\big)$	Bernoulli $\big(\text{expit}(-0.5 + 0.25X_{t-1} + 0.5Z_t)\big)$	$\mathcal{N}(1.2 + 0.1Z_t + 2C,\ 0.1)$
$\widetilde{Y}_t(\overline{x}_t) \mid C$	$\text{Exp}(0.3 + 0.2x_t + 0.1C)$	$\text{Exp}(0.5 + 0.2x_t + 0.2C)$	$\text{Exp}(0.1 + 0.3x_t + 0.1C)$
Gaussian copula (with Pearson correlation ρ)	$\rho = 0.4$ between Z_t and Y_t	pair-copula: $\rho_{Y_t, Z_{t-1}} = 0.2$, $\rho_{Y_t, Z_t \mid Z_{t-1}} = 0.3$ (depends on C)	$\rho = 0.4$ between Z_t and Y_t

Table C.3: Simulation settings 1–3.

	Setting 4
C	$\mathcal{N}(0, 1)$
$Z_t \mid \overline{X}_{t-1}, C$	$\mathcal{N}(-0.5 + 0.5X_{t-1} + 0.5Z_{t-1} + 0.5C,\ 0.5)$
$X_t \mid \overline{Z}_t, C$	$\mathcal{N}(2Z_t + 0.1C,\ 1)$
$Y_t(\overline{x}_t) \mid C$	$\mathcal{N}(2x_t + x_{t-1} + 0.5x_{t-2} + C,\ 1)$
Gaussian copula with correlation ρ	$\rho = 0.24$

Table C.4: Simulation setting 4.

C.4.3 LEADER

We provide the description of each variable name in our experiment on the LEADER trial in Table C.6.

The hyperparameter configurations used in `dltmle` are listed in Table C.7. Similar to its implementation in Appendix C.4.2.2, we tuned the hyperparameters with the built in `dltmle.tune` function over 20 trials.

Hyperparameter	Hyperparameter values / candidates for search
dim_emb	$\{16, 32\}$
hidden_size	$\{32, 64\}$
num_layers	$\{1, 2, 4\}$
nhead	$\{2, 3, 4\}$
dropout	$\{0, 0.1, 0.2\}$
learning_rate	$\{10^{-3}, 10^{-2}\}$
batch_size	$\{64, 128\}$
epochs	$[50, 200]$

Table C.7: Search ranges of key hyperparameters of `dltmle` model on the LEADER trial.

Hyperparameter	Hyperparameter values / candidates for search
dim_emb	$\{8, 16\}$
dim_emb_time	$\{4, 8\}$
dim_emb_type	$\{4, 8\}$
hidden_size	$\{8, 16, 32\}$
num_layers	$\{1, 2, 4\}$
nhead	$\{2, 4\}$
dropout	$\{0, 0.1, 0.2\}$
learning_rate	$\{10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-3}\}$
alpha	[0.05, 0.1, 0.5, 1]
beta	[0.05, 0.1, 0.5, 1]

Table C.5: Search ranges of key hyperparameters of `dltmle` model.

Variable	Description
SEX_F	Binary variable. If `True`, the patient is female.
SMOKER_NEVER.SMOKED	Binary variable. If `True`, the patient has never smoked before.
STENFL	Carotid $> 50\%$ stenosis on angiography. Binary.
NEPSCRFL	Diabetic Nephropathy at screening flag. Binary.
Age	Patient's age when enrolled in the trial.
DIABDUR	Diabetes duration (years) at baseline.
HDL1BL	HDL Cholesterol (mmol/L) at baseline.
LDL1BL	LDL Cholesterol (mmol/L) at baseline.
CHOL1BL	Total Cholesterol (mmol/L) at baseline.
log_TRIG1BL	Logarithm of triglycerides (mmol/L) at baseline.
log_CREATBL	Serum Creatinine (μmol/L) at baseline.

Table C.6: Description of each variable used in Section 5.7.

Appendix D

Supplementary material for Chapter 6

D.1 Proof of Proposition 6.3.1

Proof. Let $T = t(X)$ be any statistic through which the weight factors, $w = \bar{w}(T)$. Then reweighting gives $p_w(t, y) = \bar{w}(t)\, p(t, y)$. Integrating over y gives $p_w(t) = \bar{w}(t)\, p(t)$, thus

$$p_w(y \,|\, t) = \frac{p_w(t, y)}{p_w(t)} = \frac{\bar{w}(t)\, p(t, y)}{\bar{w}(t)\, p(t)} = p(y \,|\, t),$$

that is, $P_w(Y \,|\, T) = P(Y \,|\, T)$.

The weight $w(x) = \tilde{w}(s(x))$ factors through V, and, as a function of x, through X; taking $T = V$ and $T = X$ gives $P_w(Y \,|\, V) = P(Y \,|\, V)$ and $P_w(Y \,|\, X) = P(Y \,|\, X)$. The same holds for any statistic through which w factors.

For a covariate subset X_S with w depending only on X_{S_w} and $S_w \subseteq S$, writing the remaining covariates as $x_{-S} = (x_j)_{j \notin S}$, the step $T = X_S$ is the marginalisation over x_{-S},

$$p_w(x_S, y) = \int \tilde{w}(x_{S_w})\, p(x, y)\, \mathrm{d}x_{-S} = \tilde{w}(x_{S_w})\, p(x_S, y),$$

since $\tilde{w}(x_{S_w})$ is constant in x_{-S}. Integrating over y gives $p_w(x_S) = \tilde{w}(x_{S_w})\, p(x_S)$, thus

$$p_w(y \,|\, x_S) = \frac{\tilde{w}(x_{S_w})\, p(x_S, y)}{\tilde{w}(x_{S_w})\, p(x_S)} = p(y \,|\, x_S),$$

that is, $P_w(Y \,|\, X_S) = P(Y \,|\, X_S)$. □

D.2 Proof of Proposition 6.3.2

Proof. If $g \in \mathcal{G}_I$, then $\sigma_g^2(F_g) = c_g$ almost surely, and taking expectations gives $c_g = \bar{\sigma}_g^2$. Since $\sigma_g^2(F_g) = \mathbb{E}[R_g \mid F_g]$, the law of iterated expectation gives

$$\mathbb{E}[R_g \, \mathbb{1}\{F_g \le s\}] = \mathbb{E}[\sigma_g^2(F_g) \, \mathbb{1}\{F_g \le s\}] = \bar{\sigma}_g^2 \Pr(F_g \le s).$$

Conversely, suppose this equality holds for all $s \in \mathbb{R}$. As above, $\mathbb{E}[R_g \, \mathbb{1}\{F_g \le s\}] = \mathbb{E}[\sigma_g^2(F_g) \, \mathbb{1}\{F_g \le s\}]$, so the supposed equality states that the two finite measures

$$\mathcal{A} \mapsto \mathbb{E}[\sigma_g^2(F_g) \, \mathbb{1}\{F_g \in \mathcal{A}\}] \qquad \text{and} \qquad \mathcal{A} \mapsto \bar{\sigma}_g^2 \Pr(F_g \in \mathcal{A})$$

have the same distribution function. A finite measure on $\mathbb{R}$ is determined by its distribution function, so the two measures coincide on every Borel set $\mathcal{A}$, giving $\mathbb{E}[(\sigma_g^2(F_g) - \bar{\sigma}_g^2) \, \mathbb{1}\{F_g \in \mathcal{A}\}] = 0$. On $\mathcal{A} = \{t : \sigma_g^2(t) > \bar{\sigma}_g^2\}$ and $\mathcal{A} = \{t : \sigma_g^2(t) < \bar{\sigma}_g^2\}$ the integrand has a strict sign, so each set is P_{F_g}-null. Hence $\sigma_g^2(t) = \bar{\sigma}_g^2$ for P_{F_g}-almost every t, that is, $g \in \mathcal{G}_I$. □

D.3 Proof of Proposition 6.4.1

Proof. If $g \in \mathcal{G}_I$, then $\sigma_g^2(t) = \bar{\sigma}_g^2$ for P_{F_g}-almost every t, and the integrand of (6.2) at threshold s is

$$\mathbb{E}[R_g \, \mathbb{1}\{F_g \le s\}] - \bar{\sigma}_g^2 \Pr(F_g \le s) = \int_{(-\infty, \, s]} (\sigma_g^2(t) - \bar{\sigma}_g^2) \, \mathrm{d}P_{F_g}(t) = 0,$$

so $\mathcal{C}(g) = 0$.

Conversely, suppose $\mathcal{C}(g) = 0$. Let $U = H_g(F_g)$, and define

$$\sigma_U^2(u) \coloneqq \sigma_g^2(H_g^{-1}(u)), \qquad \psi_g(u) \coloneqq \sigma_U^2(u) - \bar{\sigma}_g^2, \qquad \Psi_g(u) \coloneqq \int_0^u \psi_g(v) \, \mathrm{d}v.$$

Since H_g is continuous, $U \sim \mathrm{Uniform}(0,1)$ by the probability integral transform. Moreover, an interval on which H_g is constant carries no P_{F_g}-mass. Hence $F_g = H_g^{-1}(U)$ almost surely and, for every s, $\mathbb{1}\{F_g \le s\} = \mathbb{1}\{U \le H_g(s)\}$ almost surely. By the law of iterated expectation, $\int_0^1 \psi_g(v) \, \mathrm{d}v = \mathbb{E}[\sigma_g^2(F_g)] - \bar{\sigma}_g^2 = 0$, so $\Psi_g(0) = \Psi_g(1) = 0$. For every threshold s, iterated expectation likewise gives

$$\mathbb{E}[R_g\,\mathbb{1}\{F_g \le s\}] = \mathbb{E}[\sigma_g^2(F_g)\,\mathbb{1}\{F_g \le s\}] = \mathbb{E}[\sigma_U^2(U)\,\mathbb{1}\{U \le H_g(s)\}] = \int_0^{H_g(s)} \sigma_U^2(v)\,\mathrm{d}v,$$

and $\Pr(F_g \le s) = H_g(s)$, so the integrand of (6.2) at threshold s equals $\int_0^{H_g(s)} \psi_g(v)\,\mathrm{d}v = \Psi_g(H_g(s))$. Averaging its square over $s \sim P_{F_g}$ and using $H_g(F_g) = U \sim \mathrm{Uniform}(0,1)$,

$$\mathcal{C}(g) = \mathbb{E}\Big[\Psi_g(H_g(F_g))^2\Big] = \int_0^1 \Psi_g(u)^2\,\mathrm{d}u.$$

Since Ψ_g is absolutely continuous on $[0,1]$, being the integral of $\psi_g \in L^1([0,1])$, the supposition $\mathcal{C}(g) = 0$ forces $\Psi_g(u) = 0$ for all $u \in [0,1]$, and hence $\psi_g(u) = \Psi_g'(u) = 0$ for Lebesgue-almost every $u \in [0,1]$. Translating back via $u = H_g(t)$, $\sigma_g^2(t) = \bar{\sigma}_g^2$ for P_{F_g}-almost every t, i.e., $g \in \mathcal{G}_I$. □

D.4 Proof of Lemma 6.5.1

Proof. Fix $u \in [0,1]$. The function $k_0(\cdot, u)$ satisfies $k_0(1,u) = 0$ and is absolutely continuous with $\partial_v k_0(v,u) = -\mathbb{1}\{v > u\} \in L^2[0,1]$, so $k_0(\cdot,u) \in \mathcal{H}$. For every $h \in \mathcal{H}$,

$$\langle h, k_0(\cdot,u)\rangle_{\mathcal{H}} = -\int_u^1 h'(v)\,\mathrm{d}v = h(u) - h(1) = h(u).$$

Evaluation at u is thus a bounded linear functional represented by $k_0(\cdot,u)$, so $\mathcal{H}$ is an RKHS with kernel k_0. Equivalently, $k_0(v,w) = \min(1-v, 1-w)$ arises from the kernel $\min(a,b)$ of Berlinet and Thomas-Agnan (2011, Chapter 1, Exercise 14) by the change of variables $a = 1-v$, $b = 1-w$, under which the boundary condition $f(0) = 0$ becomes $h(1) = 0$.

Substituting the explicit form of k_0,

$$\xi_g(u) = (1-u)\,\Psi_g(u) + \int_u^1 \psi_g(v)\,(1-v)\,\mathrm{d}v,$$

so $\xi_g(1) = 0$, ξ_g is absolutely continuous, and $\xi_g'(u) = -\Psi_g(u) + (1-u)\,\psi_g(u) - (1-u)\,\psi_g(u) = -\Psi_g(u)$ for almost every u. Since $|\Psi_g| \le \int_0^1 |\psi_g| \le 2\bar{\sigma}_g^2$, we have $\xi_g' \in L^2[0,1]$ and hence $\xi_g \in \mathcal{H}$. For $h \in \mathcal{H}$, integration by parts with $\Psi_g(0) = \Psi_g(1) = 0$ gives

$$\langle h, \xi_g\rangle_{\mathcal{H}} = \int_0^1 h'(u)\,\xi_g'(u)\,\mathrm{d}u = -\int_0^1 h'(u)\,\Psi_g(u)\,\mathrm{d}u = \int_0^1 h(u)\,\psi_g(u)\,\mathrm{d}u.$$

Finally, $\|\xi_g\|_{\mathcal{H}}^2 = \int_0^1 (\xi_g')^2 = \int_0^1 \Psi_g(u)^2\,\mathrm{d}u$, which equals $\mathcal{C}(g)$ by the change of variable $u = H_g(s)$

in Appendix D.3. □

D.5 Proof of Theorem 6.5.2

Proof. Let $U = H_g(F_g) \sim \text{Uniform}(0,1)$. For any $h \in \mathcal{H}$, writing $w_h = \mathrm{d}Q_h/\mathrm{d}P$ for the corresponding density ratio,

$$\mathbb{E}_{Q_h}[R_g] - \mathbb{E}_P[R_g] = \mathbb{E}_P\big[R_g\,(w_h(X)-1)\big] = \mathbb{E}_P\Big[R_g\Big(h(U) - \int_0^1 h\Big)\Big].$$

By iterated expectation, $\mathbb{E}_P[R_g\,h(U)] = \int_0^1 \sigma_g^2(H_g^{-1}(u))\,h(u)\,\mathrm{d}u$, and combining with $\int_0^1 h = \mathbb{E}[h(U)]$ and Lemma 6.5.1 gives

$$\mathbb{E}_{Q_h}[R_g] - \mathbb{E}_P[R_g] = \int_0^1 \psi_g(u)\,h(u)\,\mathrm{d}u = \langle h, \xi_g\rangle_{\mathcal{H}}. \tag{D.1}$$

By Fenchel duality, the supremum of $\langle h, \xi_g\rangle_{\mathcal{H}} - \|h\|_{\mathcal{H}}^2/(4\lambda)$ over $h \in \mathcal{H}$ is the convex conjugate of $h \mapsto \|h\|_{\mathcal{H}}^2/(4\lambda)$, evaluated at ξ_g. Completing the square in $\mathcal{H}$,

$$\langle h, \xi_g\rangle_{\mathcal{H}} - \frac{\|h\|_{\mathcal{H}}^2}{4\lambda} = \lambda\,\|\xi_g\|_{\mathcal{H}}^2 - \frac{\|h - 2\lambda\,\xi_g\|_{\mathcal{H}}^2}{4\lambda},$$

so this supremum equals $\lambda\,\|\xi_g\|_{\mathcal{H}}^2$ and is attained only at $h^* = 2\lambda\,\xi_g$. Hence, with $\|\xi_g\|_{\mathcal{H}}^2 = \mathcal{C}(g)$ from Lemma 6.5.1,

$$\sup_{h\in\mathcal{H}}\Big\{\mathbb{E}_{Q_h}[R_g] - \tfrac{\|h\|_{\mathcal{H}}^2}{4\lambda}\Big\} = \mathbb{E}_P[R_g] + \sup_{h\in\mathcal{H}}\Big\{\langle h, \xi_g\rangle_{\mathcal{H}} - \tfrac{\|h\|_{\mathcal{H}}^2}{4\lambda}\Big\} = \mathbb{E}_P[R_g] + \lambda\,\mathcal{C}(g),$$

which is the claimed identity, with the supremum attained only at $h^* = 2\lambda\,\xi_g$. Since the identity holds for every $g \in \mathcal{G}$, minimising either side over $\mathcal{G}$ returns the same minimisers, which gives the stated characterisation of g_λ. □

D.6 Proof of Proposition 6.6.2

The limiting statement of Proposition 6.6.2 uses the following continuity property of the penalty.

Lemma D.6.1 (Continuity of $\mathcal{C}$ in $L_2(P_X)$)**.** *Let $g^* \in \mathcal{G}$ and let $(g_k)_{k\geq 1}$ be any sequence in $\mathcal{G}$ with $\|g_k - g^*\|_{L_2(P_X)} \to 0$ as $k \to \infty$. Then $\mathcal{C}(g_k) \to \mathcal{C}(g^*)$.*

Proof. For each $g \in \mathcal{G}$ and $s \in \mathbb{R}$, define the population deviation

$$\nu(g,s) \coloneqq \mathbb{E}\big[R_g\,\mathbb{1}\{F_g \le s\}\big] - \bar{\sigma}_g^2 \Pr(F_g \le s),$$

so that $\mathcal{C}(g) = \mathbb{E}\big[\nu(g, g(X))^2\big]$ and $|\nu(g,s)| \le 2\bar{\sigma}_g^2$ for all s. Since $g_k \to g^*$ in $L_2(P_X)$, we have $g_k(X) \to g^*(X)$ in probability, and hence, for every s,

$$\mathbb{1}\{g_k(X) \le s\} \to \mathbb{1}\{g^*(X) \le s\} \quad \text{in probability,} \tag{D.2}$$

because $P\big(g^*(X) = s\big) = 0$ by the continuity of H_{g^*} (Definition 6.2.1).

We first prove the uniform convergence $\sup_s |\nu(g_k,s) - \nu(g^*,s)| \to 0$. For any fixed s,

$$\begin{aligned}|\nu(g_k,s) - \nu(g^*,s)| \le{}& \underbrace{\Big|\mathbb{E}\big[R_{g_k}\mathbb{1}\{F_{g_k} \le s\}\big] - \mathbb{E}\big[R_{g^*}\mathbb{1}\{F_{g^*} \le s\}\big]\Big|}_{①} \\ &+ \underbrace{\big|\bar{\sigma}_{g_k}^2 - \bar{\sigma}_{g^*}^2\big|}_{②} + \underbrace{\bar{\sigma}_{g_k}^2\,\big|\Pr(F_{g_k} \le s) - \Pr(F_{g^*} \le s)\big|}_{③}.\end{aligned}$$

For ①, since $0 \le \mathbb{1}\{F_{g_k} \le s\} \le 1$ pointwise,

$$① \le \mathbb{E}\big|R_{g_k} - R_{g^*}\big| + \Big|\mathbb{E}\big[R_{g^*}\mathbb{1}\{F_{g_k} \le s\}\big] - \mathbb{E}\big[R_{g^*}\mathbb{1}\{F_{g^*} \le s\}\big]\Big|.$$

Expanding $R_{g_k} - R_{g^*}$ and applying the Cauchy–Schwarz inequality,

$$\begin{aligned}\mathbb{E}\big|R_{g_k} - R_{g^*}\big| &= \mathbb{E}\Big[\big|g^*(X) - g_k(X)\big|\,\big|2Y - g_k(X) - g^*(X)\big|\Big] \\ &\le \big\|g_k - g^*\big\|_{L_2(P_X)}\Big(2\,\mathbb{E}[Y^2]^{1/2} + \mathbb{E}\big[g_k(X)^2\big]^{1/2} + \mathbb{E}\big[g^*(X)^2\big]^{1/2}\Big),\end{aligned}$$

which tends to zero, the second factor being bounded since $\mathbb{E}[Y^2] < \infty$ and g_k converges in $L_2(P_X)$. This bound does not depend on s, and it also gives ② $\le \mathbb{E}|R_{g_k} - R_{g^*}| \to 0$; in particular $\bar{\sigma}_{g_k}^2 \to \bar{\sigma}_{g^*}^2$.

For the remaining part of ①, define a tilted probability measure $\tilde{P}$ with density

$$\frac{\mathrm{d}\tilde{P}}{\mathrm{d}P} = \frac{R_{g^*}}{\bar{\sigma}_{g^*}^2},$$

which is non-negative with $\mathbb{E}_P[R_{g^*}/\bar{\sigma}_{g^*}^2] = 1$. Under $\tilde{P}$,

$$\frac{\mathbb{E}[R_{g^*}\mathbb{1}\{F_{g_k} \le s\}]}{\bar{\sigma}^2_{g^*}} = \tilde{P}(g_k(X) \le s) \quad \text{and} \quad \frac{\mathbb{E}[R_{g^*}\mathbb{1}\{F_{g^*} \le s\}]}{\bar{\sigma}^2_{g^*}} = \tilde{P}(g^*(X) \le s),$$

the distribution functions of $g_k(X)$ and $g^*(X)$ under $\tilde{P}$. By (D.2) and dominated convergence with dominating function $R_{g^*} \in L_1(P)$, $\tilde{P}(g_k(X) \le s) \to \tilde{P}(g^*(X) \le s)$ for every s. Since $\tilde{P}$ is absolutely continuous with respect to P and $\Pr(g^*(X) = s) = 0$ for every s, the distribution of $g^*(X)$ under $\tilde{P}$ has no atoms. Its distribution function is therefore continuous, and Pólya's theorem (van der Vaart, 2000, Lemma 2.11) yields

$$\sup_s \left|\mathbb{E}[R_{g^*}\mathbb{1}\{F_{g_k} \le s\}] - \mathbb{E}[R_{g^*}\mathbb{1}\{F_{g^*} \le s\}]\right| \to 0.$$

For ③, $\Pr(F_{g_k} \le s) \to \Pr(F_{g^*} \le s)$ for every s by (D.2) and bounded convergence, and H_{g^*} is continuous. Pólya's theorem again gives $\sup_s |\Pr(F_{g_k} \le s) - \Pr(F_{g^*} \le s)| \to 0$, while the factor $\bar{\sigma}^2_{g_k}$ is bounded. Combining the three terms proves the uniform convergence.

By the triangle inequality,

$$\left|\nu(g_k, g_k(X)) - \nu(g^*, g^*(X))\right| \le \sup_s |\nu(g_k, s) - \nu(g^*, s)| + \left|\nu(g^*, g_k(X)) - \nu(g^*, g^*(X))\right|.$$

The map $s \mapsto \nu(g^*, s) = \bar{\sigma}^2_{g^*}\{\tilde{P}(g^*(X) \le s) - H_{g^*}(s)\}$ is bounded and continuous, being a difference of two continuous distribution functions. Since $g_k(X) \to g^*(X)$ in probability, the second term tends to zero in probability. Combining with the uniform bounds $|\nu(g_k, \cdot)| \le 2\bar{\sigma}^2_{g_k}$ and $|\nu(g^*, \cdot)| \le 2\bar{\sigma}^2_{g^*}$, dominated convergence gives

$$\mathbb{E}\Big[\nu(g_k, g_k(X))^2\Big] \to \mathbb{E}\Big[\nu(g^*, g^*(X))^2\Big],$$

that is, $\mathcal{C}(g_k) \to \mathcal{C}(g^*)$. □

Proof of Proposition 6.6.2. By Assumption 6.2.2, $R_{g_0} = \varepsilon^2$ with $\varepsilon \perp\!\!\!\perp g_0(X)$, so the conditional MSE $\sigma^2_{g_0}(t) = \mathbb{E}[\varepsilon^2 \mid F_{g_0} = t] = \sigma^2$ is constant; hence $g_0 \in \mathcal{G}_I$, and $\mathcal{C}(g_0) = 0$ by Proposition 6.4.1.

For any $g \in \mathcal{G}_I$,

$$\mathbb{E}[(Y - g(X))^2] = \mathbb{E}[(g_0(X) - g(X))^2] + \sigma^2 + 2\,\mathbb{E}[(g_0(X) - g(X))\,\varepsilon].$$

The cross term is zero: $\mathbb{E}[g_0(X)\,\varepsilon] = 0$ by $\varepsilon \perp\!\!\!\perp g_0(X)$ with $\mathbb{E}[\varepsilon] = 0$, and $\mathbb{E}[g(X)\,\varepsilon] = 0$ by Assumption 6.6.1. Therefore

$$\mathbb{E}[(Y - g(X))^2] = \mathbb{E}[(g_0(X) - g(X))^2] + \sigma^2 \geq \sigma^2 = \mathbb{E}[(Y - g_0(X))^2],$$

with equality if and only if $g(X) = g_0(X)$ almost surely. Hence g_0 is the unique minimiser of $\mathbb{E}[(Y - g(X))^2]$ over $\mathcal{G}_I$.

For the limiting statement, let g_λ minimise $\mathbb{E}[(Y - g(X))^2] + \lambda \mathcal{C}(g)$, and let $\tilde{g} \in \mathcal{G}$ be an $L_2(P_X)$-limit point, with $g_{\lambda_k} \to \tilde{g}$ and $\lambda_k \to \infty$. Since $\mathcal{C}(g_0) = 0$, comparing g_λ with g_0 gives

$$\mathbb{E}[(Y - g_\lambda(X))^2] + \lambda \mathcal{C}(g_\lambda) \leq \mathbb{E}[(Y - g_0(X))^2],$$

so both $\mathcal{C}(g_\lambda) \leq \mathbb{E}[(Y - g_0(X))^2]/\lambda \to 0$ and $\mathbb{E}[(Y - g_\lambda(X))^2] \leq \mathbb{E}[(Y - g_0(X))^2] = \sigma^2$.

By Lemma D.6.1, $\mathcal{C}(\tilde{g}) = \lim_k \mathcal{C}(g_{\lambda_k}) = 0$, so $\tilde{g} \in \mathcal{G}_I$ by Proposition 6.4.1. Moreover, $\mathbb{E}|R_{g_{\lambda_k}} - R_{\tilde{g}}| \to 0$ by the Cauchy–Schwarz bound in the proof of Lemma D.6.1. Hence $\mathbb{E}[(Y - \tilde{g}(X))^2] = \lim_k \mathbb{E}[(Y - g_{\lambda_k}(X))^2] \leq \sigma^2$. The first part then gives $\mathbb{E}[(Y - \tilde{g}(X))^2] = \mathbb{E}[(g_0(X) - \tilde{g}(X))^2] + \sigma^2 \leq \sigma^2$, forcing $\tilde{g} = g_0$ in $L_2(P_X)$. □

D.7 Proof of Lemma 6.7.2

Proof. Fix e. Under Assumption 6.7.1, $Y^e - g_0(X^e) = \varepsilon_e$, so

$$\sigma^2_{g_0,e}(t) = \mathbb{E}_{P_e}\left[(Y - g_0(X))^2 \,\middle|\, g_0(X) = t\right] = \mathbb{E}_{P_e}\left[\varepsilon_e^2 \,\middle|\, g_0(X) = t\right] = \mathbb{E}_{P_e}[\varepsilon_e^2],$$

constant in t by $\varepsilon_e \perp\!\!\!\perp g_0(X^e)$. Hence $g_0 \in \mathcal{G}_e$, and since e is arbitrary, $g_0 \in \mathcal{G}_\cap$. □

D.8 Proof of Proposition 6.7.4

Proof. By Lemma 6.7.2, $g_0 \in \mathcal{G}_\cap$. Conversely, take $g \in \mathcal{G}_\cap$; then $g \in \mathcal{G}_e$ for every $e \in \{1, \ldots, K\}$, hence $\mathcal{C}^e(g) = 0$ for every $e \in \{1, \ldots, K\}$ by Proposition 6.4.1 applied under P_e. If $g \neq g_0$, Assumption 6.7.3 gives an environment e^* with $\mathcal{C}^{e^*}(g) > 0$, contradicting $\mathcal{C}^{e^*}(g) = 0$. Hence $\mathcal{G}_\cap = \{g_0\}$. Since each $\mathcal{C}^e \geq 0$ and $\omega_e > 0$, $\mathcal{C}^{\text{ew}}(g) = 0$ if and only if $\mathcal{C}^e(g) = 0$ for every $e \in \{1, \ldots, K\}$, that is $g \in \mathcal{G}_\cap$; therefore g_0 is the only $g \in \mathcal{G}$ with $\mathcal{C}^{\text{ew}}(g) = 0$. □

Bibliography

A. Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, 97(457):284–292, 2002.

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

A. Araujo and E. Giné. *The central limit theorem for real and Banach valued random variables*. Wiley, 1980.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):597–623, 2018.

S. Athey, G. W. Imbens, J. Metzger, and E. Munro. Using Wasserstein generative adversarial networks for the design of Monte Carlo simulations. *Journal of Econometrics*, 240(2):105076, 2024.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

D. Benkeser, W. Cai, and M. J. van der Laan. A nonparametric super-efficient estimator of the average treatment effect. 2020.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

M. P. Bitler, J. B. Gelbach, and H. W. Hoynes. Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? evidence from a social experiment. *Review of Economics and Statistics*, 99(4):683–697, 2017.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

F.-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*, 2019.

M. A. Brookhart, S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer. Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12): 1149–1156, 2006.

M. C. Bryson. Heavy-tailed distributions: properties and tests. *Technometrics*, 16(1):61–68, 1974. Publisher: Taylor & Francis.

P. Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

H. U. Burger, C. Gerlinger, C. Harbron, A. Koch, M. Posch, J. Rochon, and A. Schiel. The use of external controls: To what extent can it currently be recommended? *Pharmaceutical Statistics*, 20(6):1002–1016, 2021.

S. A. Busgang, L. A. Waller, E. Colicino, R. D'Agostino Jr, I. Hertz-Picciotto, and C. Gennings. Selecting external controls for internal cases using stratification score matching methods. *International Journal of Environmental Research and Public Health*, 19(5):2549, 2022.

S. Calonico, M. D. Cattaneo, and M. H. Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113 (522):767–779, 2018.

S. Calonico, M. D. Cattaneo, and M. H. Farrell. nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *arXiv preprint arXiv:1906.00198*, 2019.

S. L. Chau, A. Schrab, A. Gretton, D. Sejdinovic, and K. Muandet. Credal two-sample tests of epistemic uncertainty. *arXiv preprint arXiv:2410.12921*, 2024.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

A. M. Christgau and N. R. Hansen. Efficient adjustment for complex covariates: Gaining efficiency with DOPE. *arXiv preprint arXiv:2402.12980v1*, 2024.

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

A. P. Dawid and V. Didelez. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231, 2010.

D. de Vassimon Manela, L. Battaglia, and R. Evans. Marginal causal flows for validation and inference. *Advances in Neural Information Processing Systems*, 37:9920–9949, 2024.

D. de Vassimon Manela, L. Yang, and R. J. Evans. Testing generalizability in causal inference. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448): 1053–1062, 1999.

I. Díaz. Efficient estimation of quantiles in missing data models. *Journal of Statistical Planning and Inference*, 190:39–51, 2017.

I. Díaz. Doubly robust estimators for the average treatment effect under positivity violations: introducing the *e*-score. *arXiv preprint arXiv:1807.09148*, 2018.

J. C. Duchi and H. Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956. Publisher: JSTOR.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

R. J. Evans. causl. https://github.com/rje42/causl, 2021a.

R. J. Evans. survivl. https://github.com/rje42/survivl, 2021b.

R. J. Evans and V. Didelez. Parameterizing and simulating from causal models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):535–568, 2024.

K. M. Flegal, D. Kruszon-Moran, M. D. Carroll, C. D. Fryar, and C. L. Ogden. Trends in obesity among adults in the United States, 2005 to 2014. *JAMA*, 315(21):2284–2291, 2016.

D. Frauen, T. Hatt, V. Melnychuk, and S. Feuerriegel. Estimating average causal effects from patient trajectories. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 7586–7594, 2023.

D. S. Freedman, L. K. Khan, W. H. Dietz, S. R. Srinivasan, and G. S. Berenson. Relationship of childhood obesity to coronary heart disease risk factors in adulthood: the Bogalusa heart study. *Pediatrics*, 108(3):712–718, 2001.

S. Friedrich and T. Friede. On the role of benchmarking data sets and simulations in method comparison studies. *Biometrical Journal*, 66(1):2200212, 2024.

N. Galwey. Supplementation of a clinical trial by historical control data: is the prospect of dynamic borrowing an illusion? *Statistics in Medicine*, 36(6):899–916, 2017.

J. L. Gamella, J. Peters, and P. Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, pages 1–12, 2025.

C. Gao and S. Yang. Pretest estimation in combining probability and non-probability samples. *Electronic Journal of Statistics*, 17(1):1492–1546, 2023.

M. Giuffrè and D. L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, 6(1):186, 2023.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.

A. Gretton, K. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems*, 19, 2006.

A. Gretton, K. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

L. Gui and V. Veitch. Causal estimation for text data with (apparent) overlap violations. *arXiv preprint arXiv:2210.00079*, 2022.

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12(0):1–115, 1944.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

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

W. Havercroft and V. Didelez. Simulating from marginal structural models with time-dependent confounding. *Statistics in Medicine*, 31(30):4190–4206, 2012.

C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

L. Henckel, E. Perković, and M. H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, 2022.

M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2025.

K. Hirano and G. W. Imbens. *The Propensity Score with Continuous Treatments*, chapter 7, pages 73–84. John Wiley & Sons, Ltd, 2004.

W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. Publisher: Institute of Mathematical Statistics.

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

A. Holovchak, S. Saengkyongam, N. Meinshausen, and X. Shen. Distributional instrumental variable method. *arXiv preprint arXiv:2502.07641*, 2025.

C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

P. J. Huber. *Robust statistics*. Wiley series in probability and mathematical statistics. Wiley, New York ;, 1981.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 2014.

G. Imbens and Y. Xu. Lalonde (1986) after nearly four decades: Lessons learned. *arXiv preprint arXiv:2406.00827*, 2024.

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

C. Ju, R. Wyss, J. M. Franklin, S. Schneeweiss, J. Häggström, and M. J. van der Laan. Collaborative-controlled lasso for constructing propensity score-based estimators in high-dimensional data. *Statistical Methods in Medical Research*, 28(4):1044–1063, 2019.

S. Kamthe, S. Assefa, and M. Deisenroth. Copula flows for synthetic data generation. *arXiv preprint arXiv:2101.00598*, 2021.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pages 207–236, 2024.

E. H. Kennedy, S. Balakrishnan, and L. Wasserman. Semiparametric counterfactual density estimation. *Biometrika*, 110(4):875–896, 2023.

C. T. Kenny, S. Kuriwaki, C. McCartan, E. T. Rosenman, T. Simko, and K. Imai. The use of differential privacy for census data and its impact on redistricting: The case of the 2020 US Census. *Science Advances*, 7(41):eabk3283, 2021.

M. Kojima. Dynamic borrowing method for historical information using a frequentist approach for hybrid control design. *arXiv preprint arXiv:2305.14679*, 2023.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer, 2022.

S. D. Lendle, J. Schwab, M. L. Petersen, and M. J. van der Laan. ltmle: an R package implementing targeted minimum loss-based estimation for longitudinal data. *Journal of Statistical Software*, 81:1–21, 2017.

F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

W. Li, F. Liu, and D. Snavely. Revisit of test-then-pool methods and some practical considerations. *Pharmaceutical Statistics*, 19(5):498–517, 2020.

Q. Liu and W. H. Wong. An AI-powered Bayesian generative modeling approach for causal inference in observational studies. *arXiv preprint arXiv:2501.00755*, 2025.

Q. Liu, Z. Chen, and W. H. Wong. An encoding generative modeling approach to dimension reduction and covariate adjustment in causal inference with observational studies. *Proceedings of the National Academy of Sciences*, 121(23):e2322376121, 2024.

X. Liu and A. Gandy. Kernel tests of equivalence. *arXiv preprint arXiv:2603.10886*, 2026.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems*, 30, 2017.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

S. P. Marso, G. H. Daniels, K. Brown-Frandsen, P. Kristensen, J. F. Mann, M. A. Nauck, S. E. Nissen, S. Pocock, N. R. Poulter, L. S. Ravn, et al. Liraglutide and cardiovascular outcomes in type 2 diabetes. *New England Journal of Medicine*, 375(4):311–322, 2016.

C. McCartan, T. Simko, and K. Imai. Making differential privacy work for census data users. *Harvard Data Science Review*, 10, 2023.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 141(1): 148–188, 1989.

N. Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

T. Meng, X. Jing, Z. Yan, and W. Pedrycz. A survey on machine learning for data fusion. *Information Fusion*, 57:115–129, 2020.

H. Motoyama. Extended Glivenko—Cantelli theorem for simple random sampling without replacement from a finite population. *Communications in Statistics-Theory and Methods*, 53 (16):5924–5934, 2024.

K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2): 1–141, 2017.

K. Muandet, M. Kanagawa, S. Saengkyongam, and S. Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

B. Neal, C.-W. Huang, and S. Raghupathi. RealCause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

J. Neyman. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in English. *Statistical Science*, 5(463-472):4, 1923.

Y. Ning, P. Sida, and K. Imai. Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554, 2020.

Novo Nordisk A/S. Liraglutide effect and action in diabetes: Evaluation of cardiovascular outcome results (LEADER). https://clinicaltrials.gov/study/NCT01179048, 2015. NLM Identifier: NCT01179048.

H. Parikh, C. Varjao, L. Xu, and E. T. Tchetgen. Validating causal inference methods. In *International Conference on Machine Learning*, pages 17346–17358. PMLR, 2022.

J. Park, U. Shalit, B. Schölkopf, and K. Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In *International Conference on Machine Learning*, pages 8401–8412. PMLR, 2021.

J. Pearl. *Causality : models, reasoning, and inference*. Cambridge University Press, Cambridge, 2nd ed. edition, 2009.

J. Pearl. On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pages 417–424, Arlington, Virginia, USA, 2010. AUAI Press.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54, 2012.

V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, and D. I. Fotiadis. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal*, 2024.

A. Ramdas and R. Wang. Hypothesis testing with E-values. *Foundations and Trends in Statistics*, 1(1-2):1–390, 2025.

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

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

J. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, 22(4): 544–559, 2007.

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

P. R. Rosenbaum. *Overt bias in observational studies*. Springer, 2002.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

A. Rotnitzky and E. Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86, 2020.

D. Rubin. Author's reply (to Judea Pearl's and Arvid Sjölander's letters to the editor). *Statistics in Medicine*, 28:1420–1423, 2009.

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

D. B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

V. Sadhanala, Y.-X. Wang, A. Ramdas, and R. J. Tibshirani. A higher-order kolmogorov-smirnov test. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2621–2630. PMLR, 2019.

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning (ICML)*, pages 515–521, 1998.

H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

H. Schmidli, D. A. Häring, M. Thomas, A. Cassidy, S. Weber, and F. Bretz. Beyond randomized clinical trials: use of external controls. *Clinical Pharmacology & Therapeutics*, 107(4):806–816, 2020.

A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

X. Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 105 (489), 2010.

S. Shekhar, I. Kim, and A. Ramdas. A permutation-free kernel two-sample test. *Advances in Neural Information Processing Systems*, 35:18168–18180, 2022.

X. Shen and N. Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae108, 2024.

X. Shen, P. Bühlmann, and A. Taeb. Causality-oriented robustness: exploiting general noise interventions in linear structural causal models. *Journal of the American Statistical Association*, 2025.

C. Shi, D. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32, 2019.

T. Shirakawa. dltmle python package (2024). https://github.com/shirakawatoru/dltmle, 2024.

T. Shirakawa, Y. Li, Y. Wu, S. Qiu, Y. Li, M. Zhao, H. Iso, and M. Van Der Laan. Longitudinal targeted minimum loss-based estimation with temporal-difference heterogeneous transformer. *arXiv preprint arXiv:2404.04399*, 2024.

A. Sinha, H. Namkoong, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

E. Skoufias. *PROGRESA and its impacts on the welfare of rural households in Mexico*, volume 139. Intl Food Policy Res Inst, 2005.

L. Song, K. Fukumizu, and A. Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

T. Stürmer, K. J. Rothman, J. Avorn, and R. J. Glynn. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *American Journal of Epidemiology*, 172(7):843–854, 2010.

W. Sun, P. Wang, D. Yin, J. Yang, and Y. Chang. Causal inference via sparse additive models with application to online advertising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

M. F. Taufiq, A. Doucet, R. Cornish, and J.-F. Ton. Marginal density ratio for off-policy evaluation in contextual bandits. *Advances in Neural Information Processing Systems*, 36, 2024.

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

M. J. van der Laan and S. Gruber. Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1), 2010.

M. J. van der Laan and S. Gruber. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1), 2012.

M. J. Van der Laan, S. Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

T. J. Vander Weele and I. Shpitser. A new criterion for confounder selection. *Biometrics*, 67 (4):1406–1413, 2011.

S. Venturini, F. Dominici, and G. Parmigiani. Generalized quantile treatment effect: A flexible Bayesian approach using quantile ratio smoothing. *Bayesian Analysis*, 10(3):523–552, 2015.

K. Viele, S. Berry, B. Neuenschwander, B. Amzal, F. Chen, N. Enas, B. Hobbs, J. G. Ibrahim, N. Kinnersley, S. Lindborg, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54, 2014.

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

Y.-X. Wang, A. Smola, and R. Tibshirani. The falling factorial basis and its statistical applications. In *International Conference on Machine Learning*, pages 730–738. PMLR, 2014.

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

G. Wolfer and P. Alquier. Variance-aware estimation of kernel mean embedding. *Journal of Machine Learning Research*, 26(57):1–48, 2025.

D. Wormser, S. Kaptoge, E. Di Angelantonio, A. Wood, L. Pennells, A. Thompson, N. Sarwar, J. Kizer, D. Lawlor, B. Nordestgaard, et al. Emerging risk factors collaboration separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58 prospective studies. *The Lancet*, 377(9771):1085–95, 2011.

D. Xu, M. J. Daniels, and A. G. Winterstein. A Bayesian nonparametric approach to causal inference on quantiles. *Biometrics*, 74(3):986–996, 2018.

L. Yang and R. J. Evans. Outcome-Informed Weighting for Robust ATE Estimation. *arXiv preprint arXiv:2503.15989*, 2025.

L. Yang, R. J. Evans, and X. Shen. Frugal, Flexible, Faithful: Causal Data Simulation via Frengression. *arXiv preprint arXiv:2508.01018*, 2025.

L. Yang, X. Liu, and R. J. Evans. Data Fusion with Distributional Equivalence Test-then-pool. *arXiv preprint arXiv:2603.11867*, 2026.

S. Yang and P. Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493, 2018.

J. Yoon, L. N. Drumright, and M. Van Der Schaar. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, 2020.