

Essays in Microeconometrics

by

Muyang Ren

Department of Economics  
Duke University

Defense Date: March 24th, 2025

Approved:

Matthew A. Masten, Supervisor

Federico A. Bugni

Arnaud Maurel

Adam M. Rosen

Dissertation submitted in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy in the Department of Economics  
in the Graduate School of Duke University  
2025

ABSTRACT

Essays in Microeconometrics

by

Muyang Ren

Department of Economics  
Duke University

Defense Date: March 24th, 2025

Approved:

Matthew A. Masten, Supervisor

Federico A. Bugni

Arnaud Maurel

Adam M. Rosen

An abstract of a dissertation submitted in partial fulfillment of the requirements for  
the degree of Doctor of Philosophy in the Department of Economics  
in the Graduate School of Duke University  
2025

Copyright © 2025 by  
Muyang Ren  
All rights reserved

## **Abstract**

This dissertation consists of three essays in microeconometrics.

The first essay establishes the first set of inference results for using marginal treatment effects (MTEs) to extrapolate local average treatment effects (LATEs) that are robust to limited instrument variation. These results apply not only to inference on the MTE itself but also to other causal parameters, such as policy-relevant treatment effects, which are of particular interest to policymakers.

The second essay, co-authored with Federico Bugni and Jackson Bunting, studies hypothesis testing for the marginal homogeneity assumption, an assumption where time-specific marginal distributions of the panel data remain homogeneous or time-invariant. We develop three inference methods to test this hypothesis based on asymptotic approximation, bootstrap, and permutations.

The third essay, co-authored with Matthew Masten and Alexandre Poirier, introduces a general class of relaxations of the unconfoundedness assumption, encompassing several existing approaches as special cases. We use this class to derive a variety of new identification results which can be used to assess sensitivity to unconfoundedness.

# Contents

Abstract . . . . .	iv
List of Tables . . . . .	ix
List of Figures . . . . .	x
Acknowledgements . . . . .	xi
1 Introduction . . . . .	1
2 Extrapolating LATE with Weak IVs . . . . .	3
2.1 Introduction . . . . .	3
2.2 MTE Model with Discrete IVs . . . . .	11
2.2.1 Setup . . . . .	11
2.2.2 Weak identification . . . . .	16
2.2.3 Parameter space restriction . . . . .	17
2.2.4 Notation and preliminaries . . . . .	18
2.3 Robust Inference in Linear MTE Models . . . . .	19
2.3.1 Inference with known weights . . . . .	21
2.3.2 Inference with estimated weights . . . . .	25
2.4 Robust Inference in Polynomial MTE Models . . . . .	28
2.4.1 Improved projection inference . . . . .	28
2.4.2 Contributions and modifications . . . . .	31
2.4.3 Implementation of the MLC test . . . . .	34
2.4.4 Uniform validity . . . . .	36
2.5 Incorporating Covariates . . . . .	37
2.5.1 Bias from additive separability . . . . .	39
2.5.2 Šidák-Bonferroni's correction . . . . .	42
2.5.3 Discussion of inference under additive separability . . . . .	47
2.6 Monte Carlo Simulation . . . . .	50

2.7	Empirical Application to Misdemeanor Prosecution . . . . .	53
2.7.1	Marginal policy relevant treatment effects . . . . .	56
2.7.2	Quota . . . . .	57
2.8	Conclusion . . . . .	60
3	Marginal Homogeneity Tests with Panel Data . . . . .	62
3.1	Introduction . . . . .	62
3.2	The Hypothesis Testing Problem . . . . .	66
3.2.1	Test statistics . . . . .	66
3.2.2	Asymptotic distribution under the null hypothesis . . . . .	69
3.3	Critical Values and Validity of Inference . . . . .	70
3.3.1	Asymptotic approximation . . . . .	70
3.3.2	Bootstrap . . . . .	71
3.3.3	Permutations . . . . .	73
3.4	Power Analysis . . . . .	77
3.5	Monte Carlo Simulations . . . . .	78
3.6	Empirical Application . . . . .	82
3.7	Conclusions . . . . .	85
4	A General Approach to Relaxing Unconfoundedness . . . . .	87
4.1	Introduction . . . . .	87
4.2	Setup and Target Parameters . . . . .	91
4.3	Relaxing Unconfoundedness . . . . .	96
4.3.1	The marginal sensitivity Model . . . . .	98
4.3.2	Conditional $c$ -dependence . . . . .	100
4.4	General Identification Results . . . . .	101
4.4.1	Bounds on marginal distributions . . . . .	101
4.4.2	Bounds on monotonic parameters . . . . .	105

4.5	Analytical Bounds on Treatment Effect Parameters . . . . .	108
4.5.1	Average treatment effects (Example 4.2.3) . . . . .	109
4.5.2	Quantile treatment effects (Example 4.2.5) . . . . .	110
4.5.3	Average weighted welfare (Example 4.2.8) . . . . .	110
4.5.4	Copula-dependent parameters . . . . .	111
4.6	Conclusion . . . . .	113
5	Conclusions . . . . .	115
	Appendix A Additional Results for Chapter 2 . . . . .	116
A.1	Proofs . . . . .	116
A.1.1	Notation . . . . .	116
A.1.2	Proof of uniform validity in Section 2.3 . . . . .	116
A.1.3	Proof of uniform validity in Section 2.4 . . . . .	124
A.1.4	Proof of bias formula in section 2.5.1 . . . . .	130
A.1.5	Lemmas for main results . . . . .	134
A.1.6	WLLN and CLT for triangular array . . . . .	149
A.2	Power Analysis of MLC Tests . . . . .	150
A.2.1	Main results . . . . .	150
A.2.2	Additional lemmas . . . . .	152
A.2.3	Proofs . . . . .	153
A.3	Inference with Estimated Weights for MLC Tests . . . . .	162
A.4	Power Comparison in Linear MTE Models . . . . .	164
A.5	Testing IV Strength . . . . .	166
A.5.1	Pre-testing weak identification by size distortion . . . . .	166
A.5.2	Testing exact under-identification . . . . .	168
A.6	Numerical Example of Additive Separability Bias . . . . .	170
A.7	Additional Literature Discussion . . . . .	171

A.7.1	Other approaches to inference on subvectors or functions of parameters . . . .	171
A.7.2	Covariates in weakly identified models . . . . .	175
A.8	Two-stage Regression Approach . . . . .	177
	Appendix B Additional Results for Chapter 3 . . . . .	182
B.1	Proofs . . . . .	182
	Appendix C Additional Results for Chapter 4 . . . . .	201
C.1	Bound Expressions . . . . .	201
C.1.1	Lower bound on $Y_1$ . . . . .	201
C.1.2	Upper bound on $Y_1$ . . . . .	202
C.1.3	Lower bound on $Y_0$ . . . . .	203
C.1.4	Upper bound for $Y_0$ . . . . .	204
C.2	Proofs for Section 4.3 . . . . .	205
C.3	Proofs for Section 4.4 . . . . .	206
C.3.1	Proof of Theorem 4.4.1 . . . . .	209
C.3.2	Proof of Theorem 4.4.2 . . . . .	232
C.3.3	Proofs for Section 4.4.2 . . . . .	264
C.4	Proofs for Section 4.5 . . . . .	268
C.5	Appendix: Explicit bounds on expected potential outcomes . . . . .	270
C.5.1	Proofs of analytical bounds on expectations . . . . .	272
	Bibliography . . . . .	276

## List of Tables

2.1	Weights for Treatment Effects . . . . .	15
2.2	Summary Statistics of Nonprosecution Rates across Courts . . . . .	55
2.3	95% (top) and 90% (bottom) Confidence Sets for Additive MP RTE $\alpha_+(0)$ . . . . .	57
3.1	Empirical Rejection Rates (in %) under $H_0$ . . . . .	80
3.2	Empirical Rejection Rates (in %) under $H_1$ . . . . .	81
3.3	Marginal Homogeneity Tests for Igami and Yang (2016) . . . . .	85
A.1	Important Notation in Appendix A.1 . . . . .	116
A.2	Testing Weak Identification of Additive MP RTE at $\gamma = 10\%$ and $\alpha = 5\%$ . . . . .	169
A.3	Analytical Bootstrap Test for $H_0^{\text{rank}} : \text{rank}(A) \leq 2M + 1$ . . . . .	170

## List of Figures

2.1	Estimators under Different Types of IV Strengths . . . . .	5
2.2	3D Plot of Empirical Size for MLC and Wald Tests . . . . .	51
2.3	Power Curves of the MLC and Wald Test . . . . .	52
2.4	Average Confidence Sets for Additive PRTE . . . . .	58
2.5	Average Confidence Sets for Threshold Policy Effects . . . . .	59
3.1	Average Number of Stores per Market over Time for Each Chain . . . . .	84
3.2	Average Type of Market over Time . . . . .	84
A.1	Power Curves of the Conditional Wald, MLC, and Wald Tests . . . . .	167
A.2	Bias of ATE Estimand under Additive Separability . . . . .	172

## Acknowledgements

I am deeply grateful to my advisor, Matt Masten, for his unwavering support and dedicated mentorship throughout my doctoral journey. He not only guided me in developing rigorous research skills as an econometrician, but also helped me refine my ideas and present them effectively to a broader audience. This dissertation would have been impossible without his help and guidance, and I could not have asked for a better advisor in my career.

I would also like to extend my sincere thanks to my committee members—Federico Bugni, Arnaud Maurel, and Adam Rosen—for their steady support and extensive feedback. Their valuable advice and thoughtful insights have significantly influenced my research and strengthened my work.

Beyond my committee, I have been fortunate to receive tremendous support from many people at Duke and UNC Chapel Hill, especially Huan Wu, Weiting Miao, Tianqi Li, and Xinyue Bei. Their friendship, camaraderie, and enthusiasm have brought great joy to my life.

Finally, I am thankful to my parents for their constant and silent support throughout my PhD study at Duke.

# 1. Introduction

This dissertation consists of three essays in microeconometrics. The first essay provides a set of valid inference results for using MTEs to extrapolate LATEs that are robust to limited instrument variation. These results lead to asymptotically valid confidence sets for various linear functionals of MTEs, including the LATE, the average treatment effect (ATE), the average treatment effect on the treated (ATT), and policy-relevant treatment effects, regardless of identification strength. This is the first paper to provide weak instrument robust inference results for this class of parameters. Finally, I illustrate my results using data from Agan et al. (2023) to analyze counterfactual policies of changing prosecutors' leniency and their effects on reducing recidivism.

The second essay, coauthored with Federico Bugni and Jackson Bunting, studies hypothesis testing for marginal homogeneity, an assumption that requires the time-specific marginal distributions of the panel data to be homogeneous or time-invariant. Marginal homogeneity is relevant in economic settings such as dynamic discrete games. In this paper, we propose several tests for the hypothesis of marginal homogeneity and investigate their properties. We consider an asymptotic framework in which the number of individuals  $n$  in the panel diverges, and the number of periods  $T$  is fixed. We implement our tests by comparing a studentized or non-studentized  $T$ -sample version of the Cramér-von Mises statistic with a suitable critical value. We propose three methods to construct the critical value: asymptotic approximations, the bootstrap, and time permutations. We show that the first two methods result in asymptotically exact hypothesis tests. The permutation test based on a non-studentized statistic is asymptotically exact when  $T = 2$ , but is asymptotically invalid when  $T > 2$ . In contrast, the permutation test based on a studentized statistic is always asymptotically exact. Finally, under a time-exchangeability assumption, the permutation test is exact in finite samples, both with and without studentization.

The third essay, coauthored with Matthew Masten and Alexandre Poirier, defines a general class of relaxations of the unconfoundedness assumption. This class includes several previous approaches as special cases, including the marginal sensitivity model of Tan (2006).

This class therefore allows us to precisely compare and contrast these previously disparate relaxations. We use this class to derive a variety of new identification results which can be used to assess sensitivity to unconfoundedness. In particular, the prior literature focuses on average parameters, like the ATE. We move beyond averages by providing sharp bounds for a large class of parameters, including both the quantile treatment effect (QTE) and the distribution of treatment effects (DTE), results which were previously unknown even for the marginal sensitivity model.

## 2. Extrapolating LATE with Weak IVs

In this essay, I provide uniformly valid inference results for studying policy-relevant treatment effects extrapolated by the MTE model if instruments have limited variation.

### 2.1 Introduction

This paper provides the first formal treatment of weak identification analysis in the marginal treatment effect model. Originally developed in the seminal works of Björklund and Moffitt (1987) and J. J. Heckman and Vytlacil (2001, 2005) and J. J. Heckman and Vytlacil (1999), the MTE model has been widely adopted in various studies for extrapolating treatment effects, such as returns to schooling (Carneiro et al., 2011; Moffitt, 2008), analysis of recidivism effects (Agan et al., 2023; Bhuller et al., 2020), and evaluation of social insurance programs (Aizawa et al., 2023; Maestas et al., 2013) (see Table 6 of Mogstad and Torgovitsky, 2024 for a broad survey of MTE applications). Despite its widespread use in applied economic research, traditional confidence sets for the extrapolated causal effects within this model are often too short when the probability of receiving treatment (i.e., the propensity score) exhibits limited variation across the instruments support. Moreover, this variation can be very weak even if the usual  $F$ -test statistic is very large.

To achieve valid coverage of causal effects, this paper establishes the first set of inference results that are robust against weak IV variation in MTE models. For linear MTE models, I develop an asymptotically similar conditional Wald test that delivers uniformly valid confidence sets with exact coverage. For a more general class of polynomial MTE models, I propose a modified linear combination (MLC) test that produces uniformly valid confidence sets while achieving approximate asymptotic efficiency under strong identification. Additionally, I highlight limitations of the additive separability assumption, a functional form often used to address weak variation of propensity scores, by deriving explicit formulas for the bias of estimands when this specification is incorrectly specified.

## Intuition for weak IVs in MTE models

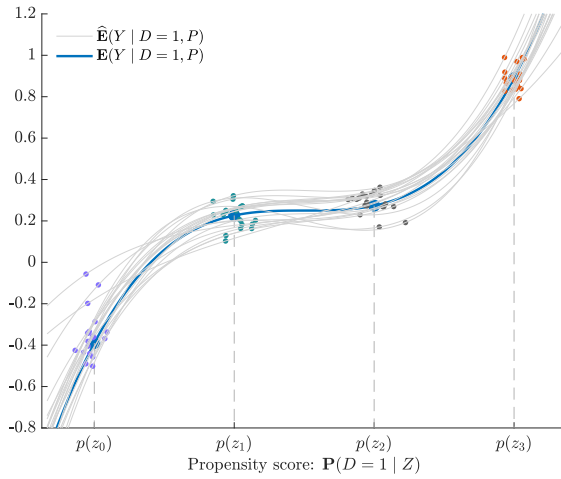
To demonstrate the consequences of weak IVs in MTE models and to illustrate why the  $F$ -statistic fails, I plot regression estimates of the outcome variable on propensity scores for treated samples in Figure 2.1. These estimators are computed from independent simulations based on a cubic MTE model with a discrete IV that takes four values. As will be shown in section 2.2, the MTE can be directly recovered from the conditional regression, allowing us to study the weak IV problem by examining the finite-sample behavior of the estimates in Figure 2.1.

Figures 2.1a through 2.1c demonstrate that as propensity score variation diminishes, the regression estimates (shown in gray) become increasingly volatile and diverge from the true population quantity (shown in blue). This behavior implies that the MTE estimator loses consistency when propensity scores exhibit limited variation—a common occurrence with instruments that weakly influence treatment selection. As a result, the traditional Wald confidence interval, which build on these compromised estimates, may fail to achieve its desired coverage probability. These estimation and inference problems arise not only when propensity scores cluster around one value but also when they approximate binary variation (Figure 2.1d) under the cubic MTE design. In such cases, classical confidence intervals become unreliable, but the  $F$ -statistic can be misleadingly large because it detects the deviation from the null where all propensity scores are equal.

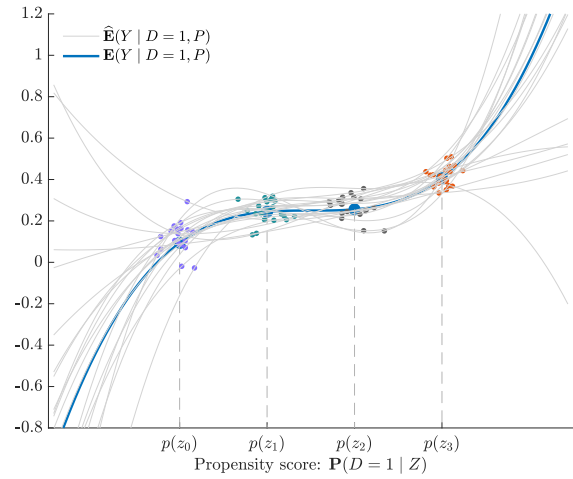
## Organization of this paper

In section 2.2, I describe the setup in Brinch et al. (2017) and Kline and Walters (2019) and present the MTE model with a discrete instrument. By imposing a parametric structure on the marginal treatment response (MTR) functions, the MTE is fully characterized by a finite-dimensional parameter, which can be point identified using discrete variation from the instrument. In finite samples, the MTE can be estimated by running separate regressions for both treated and control groups. However, these separate regression estimators, along with the corresponding Wald confidence sets, are highly vulnerable to limited variation

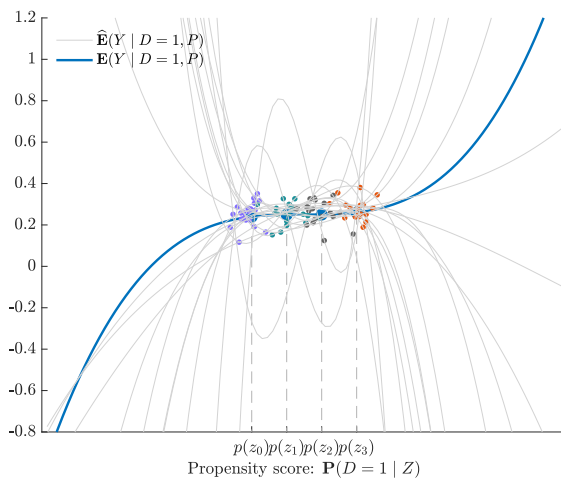
(a) Strong Variation



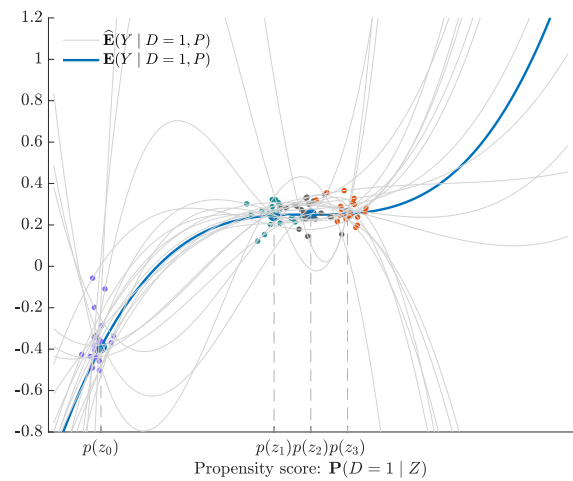
(b) Intermediate Variation



(c) Limited Variation



(d) Binary Variation



Note: The figure shows the estimated expected outcomes for treated units conditional on propensity scores under different designs of propensity score variation, based on 20 independent simulations. All designs and simulations use a cubic specification for the MTE curve with a discrete IV that takes four values in its support. The sample size is 2,000.

FIGURE 2.1: Estimators under Different Types of IV Strengths

in propensity scores. In light of this potential identification failure, this paper constructs uniformly valid confidence sets for causal parameters that are linear functionals of the MTE function, which cover a broad range of causal effects of interest including the MTE function itself, the average treatment effect (ATE), the average treatment effect on the treated

(ATT), the local average treatment effect (LATE), and the policy-relevant treatment effect (PRTE).

In section 2.3, I develop a simple inference method for treatment effects in the linear MTE model, where the MTR functions vary linearly with the selection unobservable (Brinch et al., 2017). This special structure offers a novel moment condition for constructing estimators of treatment effects of interest, bypassing the necessity for estimating the full model. Building on this new moment condition, I construct a simple conditional Wald test and demonstrate its uniform validity under weak identification. This approach circumvents challenges posed by weakly identified nuisance parameters in the MTE, thus leading to confidence sets that achieve asymptotic similarity, a key advantage over existing approaches to subvector inference with weak identification (see Appendix A.7.1).

For a generic MTE model with a discrete instrument, it is infeasible to directly estimate causal effects of interest without relying on other primitive parameters in the MTE function. In such cases, these weakly identified primitive parameters hamper the ability to perform valid inference on linear functions of them. In section 2.4, I build on the improved projection approach developed in I. Andrews (2018) and propose a MLC test that achieves uniform validity regardless of identification strength (see page 7 for a detailed comparison of my work with the literature). The inverted confidence set from this test has correct coverage under weak identification, is straightforward to compute, and is shown to be approximately as efficient as the Wald confidence set under strong identification.

In section 2.5, I incorporate covariates into the MTE model. First, I demonstrate that the commonly used additive separability condition can substantially bias causal effect estimands when it fails. Such bias does not vanish unless (1) unobserved treatment effect heterogeneity does not vary across covariates, or (2) individuals' treatment decisions do not depend on those covariates. If neither of these two assumptions hold, I show that causal effect estimands under additive separability differ from the true effects and potentially have the opposite sign. To deal with the bias induced by imposing additive separability, researchers may use the proposed methods in this paper to conduct inference by condition-

ing on covariates, while also achieving robustness against limited variation of propensity scores from conditioning. Additionally, I propose a Bonferroni-type size correction for valid inference on aggregated effects in the absence of additive separability. For researchers interested in performing inference under additive separability, I also present an extension of the proposed inference methods to accommodate this framework at the end of the section.

In section 2.6, I examine the performance of MLC tests via a sequence of Monte-Carlo simulations in a quadratic MTE model with varying degrees of identification strength. The simulation results show that the MLC test is asymptotically size-correct whenever the model is strongly identified, weakly identified, or partially identified. In contrast, the classical Wald test reports over-rejections of the null value at 15% frequency versus the 5% significance level under partial identification and exhibits trivial power under weak identification. When the instrument strength is sufficiently strong, the proposed MLC test has power close to the asymptotically efficient Wald test. Theoretical justification is provided in Appendix A.2.

In section 2.7, I illustrate the proposed methods by using data from Agan et al. (2023), who show that not prosecuting defendants with misdemeanor offenses reduces recidivism. While their analysis uses LATE estimates and ATE/ATT estimates under additive separability, I complement their findings by examining recidivism effects under different counterfactual prosecution policies. Specifically, I construct confidence sets for the reduction in recidivism under two scenarios: (1) a homogeneous marginal increase in nonprosecution rates across prosecutors, (2) implementation of a minimum nonprosecution rate threshold. The results highlight that weak identification is a significant concern in higher-order polynomial models, and empirical conclusions differ substantially between robust and classical inference methods. Section 2.8 concludes with a discussion of potential future extensions.

## **Related literature**

This paper contributes to three strands of the literature: marginal treatment effects, subvector inference in weakly identified models, and judge/examiner designs. For the rest

of the introduction, I review the related literature.

Although the MTE model has been used to extrapolate treatment effects in a variety of fields, there are relatively few studies on the theory of estimation and inference in MTE models. In a semiparametric MTE model with continuous propensity scores, J. Heckman et al. (2006) use bootstrap methods to construct confidence bands for MTE functions. This approach is now common practice<sup>1</sup>, but its theoretical validity has not yet been established. Carneiro and Lee (2009) analyze the pointwise limit distributions of MTE estimators using a separate regression approach. Building on this work, Sasaki and Ura (2023) further derive asymptotic theory for the PRTE using an orthogonalized score for double debiased estimation. Both papers assume a semiparametric MTE model with additive separability and strong IV/covariate variation to achieve  $\sqrt{n}$ -consistent estimation. Mogstad et al. (2017, 2018) propose a bootstrap procedure for conducting inference on the PRTE when the MTE is partially identified. None of these papers study weak identification problems. Thus, compared to this existing literature, the inference approach I propose is the first to achieve robustness against weak instruments in MTE models.

To estimate the MTE over a large support of propensity scores, empirical researchers usually assume that the MTR functions—and their difference, the MTE—are additively separable in covariates and unobserved costs of treatment selection Brinch et al. (2017, Assumption 2). This assumption allows the MTE to be estimated on the unconditional support of propensity scores, by pooling variation in the propensity scores across different covariate values. Despite its increasing popularity in empirical work, there is little theoretical justification for additive separability. In fact, this assumption can be strong enough to point identify the MTE without exogenous variation from instruments (Pan et al., 2024). In addition, if the MTE is misspecified as linearly additively separable, Devereux (2022) provides numerical evidence indicating that omitting higher-order covariate terms can introduce significant bias in the estimated MTE slopes. I contribute to this knowledge by

---

<sup>1</sup> The Stata implementation of MTE estimation by Brave and Walstrum (2014) and Andresen (2018) uses bootstrap methods to produce confidence sets.

providing the first analytical bias formula for the causal parameters (e.g., ATE, conditional ATE, and MTE slope) when additive separability is misspecified in a widely used class of latent threshold crossing models (Kline & Walters, 2019). To avoid this bias, researchers should extrapolate treatment effects conditional on covariates instead of relying on additive separability. Moreover, the robust inference procedures proposed in this paper can help address potential weak IV variation that may arise after conditioning on covariates.

The inference problem in this paper is also related to the literature on subvector inference in weakly identified models, where a subvector is a subset of the structural parameters. For inference on subvectors, Dufour and Taamouti (2005) suggest projecting the robust confidence sets of the full vector onto the subvector of interest. However, this procedure can be very conservative especially when the dimension of the full vector is much larger than that of the target subvector. To this end, there is a sequence of studies trying to reduce the conservativeness of projection inference. Chaudhuri and Zivot (2011) consider modifying the (Robust) Lagrangian Multiplier statistic (Kleibergen, 2005) such that it is locally equivalent to asymptotic efficient subvector tests under strong identification and propose a Bonferroni method to improve the power of their tests at distant alternatives. Building on this idea, D. W. Andrews (2017) improves the Bonferroni method such that the refined tests are asymptotically non-conservative and uniformly valid. However, D. W. Andrews (2017, p.2) acknowledges six key limitations, including computational challenges and the need for additional tuning parameters to categorize the identification strength, neither of which are required for my proposed MLC test. Moreover, his method does not directly address inference on a linear function of parameters, which is the primary focus of this paper. While his method could conceptually extend to inference on a linear function through model reparametrization, finding a universal reparametrization rule that works for all linear hypotheses of interest while maintaining tractable asymptotic analysis remains challenging.

For inference on a function of parameters in a weakly identified model, I. Andrews (2018) generalizes the results in Chaudhuri and Zivot (2011) and proposes a two-step confidence

set that achieves sequential validity with controlled coverage distortions under a set of high-level conditions imposed on a sequence of data generating processes (DGPs). The MLC test considered in this paper builds on the idea from I. Andrews (2018) but differs in a few ways: (1) Most importantly, Andrews paper only provided *sequential* validity results, concluding that “conditions for *uniform* asymptotic validity are an interesting open question in Section V (page 347). I demonstrate that, with a minor modification to the test statistics, the robust test achieves uniform validity for inference on a scalar function. (2) While Andrews derived results for a general class of models using high-level conditions, my focus on the MTE model allows for more specific, primitive conditions for validity. (3) Instead of employing Andrews’ two-step approach that alternates between non-robust Wald and robust confidence sets based on identification strength, I focus exclusively on his robust confidence sets. (4) I adapt his robust confidence sets, originally developed for GMM models, to the minimum distance framework arising from separate regressions in the MTE setting. (5) By using the robust confidence set, this approach does not involve pretesting distortion in the two-step approach and maintains the desired asymptotic coverage level  $1 - \alpha$ . (6) I prove that the local power difference from the asymptotically efficient Wald test can be arbitrarily small under strong identification. Applied to MTE models, this approach allows for valid and powerful inference on treatment effects even under limited variation of propensity scores. In Appendix A.7.1, I also discuss other approaches to inference on functions (or subvectors) of parameters in weakly identified models and explain their inapplicability to the MTE model studied here.

Finally, the empirical analysis of this paper also connects to the judge/examiner design problems (see the survey by Chyn et al. (2024)). By using the quasi-random assignment of examiners, researchers can identify the causal effects of judicial decisions for defendants at the margin of being treated, and then extrapolate these effects to the broader population under the assumption of pairwise monotonicity (Frandsen et al., 2023). While the MTE model is commonly employed for such extrapolation, researchers often report LATE, ATE, ATUT, and ATT to inform potential policy decisions (Agan et al., 2023; Baron & Gross,

2023; Bhuller et al., 2020). My paper provides methods for robust inference on all of these parameters, as well as the MTE function itself, and the policy counterfactual parameters studied in J. J. Heckman and Vytlacil (2001, 2005), and Carneiro et al. (2010).

## **2.2 MTE Model with Discrete IVs**

In this section, I describe the MTE model and the related identification result following Brinch et al. (2017). Based on this result, the weakness of IVs can be characterized by limited variation of propensity scores. Then I introduce the relevant parameter space on which we achieve uniform validity of the proposed inference procedures.

### **2.2.1 Setup**

Let  $Y_1$  be the potential outcome of an individual who receive a binary treatment ( $D = 1$ ) and  $Y_0$  denote her potential outcome in the untreated state ( $D = 0$ ). The observed outcome  $Y$  is realized through

$$Y = (1 - D)Y_0 + DY_1. \tag{2.1}$$

We further specify

$$Y_d = \mu_d + V_d, \quad d = 0, 1,$$

where  $\mu_d \equiv \mathbb{E}[Y_d]$  is the mean of potential outcome. For simplicity, I leave out additional covariates and discuss them in section 2.5. The treatment is determined by a weakly separable selection equation

$$D = \mathbb{1}[U \leq \nu(Z)], \tag{2.2}$$

where  $\nu(\cdot)$  is an unknown function,  $U$  is a continuous random variable representing the unobserved cost of selection into treatment, and  $Z \in \text{supp}(Z) = \{z_0, z_1, \dots, z_K\}$  is the excluded discrete instrument. Researchers observe the outcome  $Y$ , the binary treatment  $D$ , and the excluded instrument  $Z$  from data, while the unobservables are the potential outcomes  $(Y_0, Y_1)$  and the variable  $U$  in the selection equation. The MTE model allows individuals to be selected into treatment based on their information on potential outcomes, which leads to potential dependence between  $(V_1, V_0)$  and  $U$ .

The key identifying assumption from Brinch et al. (2017, proposition 1.ii) is as follows:

**Assumption 1** (MTE model with discrete IVs).

1.  $Z \perp\!\!\!\perp U$ .
2.  $\mathbb{E}[Y_d | Z, U] = \mathbb{E}[Y_d | U]$  and  $\mathbb{E}|Y_d| < \infty$  for  $d \in \{0, 1\}$ .
3.  $U$  is continuously distributed.
4.  $0 < \mathbb{P}(D = 1 | Z = z) < 1$  for all  $z \in \text{supp}(Z)$ .
5. Let  $\{h_m(\cdot)\}_{m=1}^M$  be a set of known continuous functions defined on  $(0, 1)$ . For  $d \in \{0, 1\}$ , the MTR function is given by

$$\mathbb{E}[Y_d | F_U(U) = u] = \mu_d + \sum_{m=1}^M \rho_{dm} h_m(u) \quad \text{for } u \in (0, 1),$$

where  $F_U(\cdot)$  denotes the distribution function of  $U$ .

6. Let  $\lambda_{00}(\cdot) = \lambda_{10}(\cdot) \equiv 1$ , and

$$\lambda_{1m}(p) \equiv \frac{1}{p} \int_0^p h_m(u) du \quad \text{and} \quad \lambda_{0m}(p) \equiv \frac{1}{1-p} \int_p^1 h_m(u) du \quad \text{for } m = 1, \dots, M.$$

Then  $\{\lambda_{1m}(\cdot)\}_{m=0}^M$  and  $\{\lambda_{0m}(\cdot)\}_{m=0}^M$  are unisolvent<sup>2</sup> on  $(0, 1)$ .

7.  $\bar{K} \equiv |\{\mathbb{P}(D = 1 | Z = z) : z = z_0, z_1, \dots, z_K\}| \geq M + 1$ .

Assumptions 1.1 and 1.2 require the excluded instrument  $Z$  to be exogenous to both the selection and outcome processes. Assumption 1.3 allows us to normalize the marginal distribution of  $U$  to be uniformly distributed over  $[0, 1]$ . That is, we can transform  $U$  to a uniformly distributed variable  $\tilde{U} = F_U(U)$ . Under the exogeneity of  $Z$  in Assumption 1.1, the function  $F_U(\nu(Z))$  can then be interpreted as propensity score:

$$\begin{aligned} p(z) &\equiv \mathbb{P}(D = 1 | Z = z) \\ &= \mathbb{P}(U \leq \nu(Z) | Z = z) \\ &= \mathbb{P}(U \leq \nu(z)) \\ &= F_U(\nu(z)) \end{aligned}$$

---

<sup>2</sup> A set of  $n$  functions  $f_1, f_2, \dots, f_n$  is unisolvent on domain  $\Omega$  if the matrix  $F \in \mathbb{R}^{n \times n}$  with entries  $f_i(x_j)$  has nonzero determinant for any choice of  $n$  distinct points  $x_1, x_2, \dots, x_n$  in  $\Omega$ .

where the third line uses the Assumption 1.1. Define  $\tilde{\nu}(z) \equiv F_U(\nu(Z))$ , and then we can work with  $(\tilde{U}, \tilde{\nu}(z))$  in place of  $(U, \nu(z))$  without affecting the empirical content of the selection model. For simplicity, we drop out the tilde and assume  $U$  is uniformly distributed throughout our analysis, and therefore  $\nu(z) = p(z)$ . Assumption 1.4 validates the overlap condition, so that we observe both treated and untreated individuals for each group defined by the value of the instrument. Assumption 1.5 imposes a parametric restriction on the MTR function  $\mathbb{E}[Y_d | F_U(U) = u]$  so that the MTE can be extrapolated outside the discrete support of the instrument. Assumption 1.6 is a weak condition that rules out redundant specifications in  $\{h_m(\cdot)\}_{m=0}^M$  that cause multicollinearity in  $\{\lambda_{dm}(\cdot)\}_{m=0}^M$ . In particular, the usual polynomial specification  $h_m(u) = u^m - \frac{1}{m+1}$  satisfies this condition. Assumption 1.7 requires sufficient variation of the exogenous instrument to point identify the structural parameters. While  $\bar{K} \leq K + 1$ , with strict inequality when multiple instrument values yield identical propensity scores, point identification requires the number of distinct propensity scores—not instrument values—to exceed the order of the MTE model.

Define  $\theta_d \equiv (\mu_d, \rho_{d1}, \dots, \rho_{dM})'$  for  $d = 0, 1$ , and let  $\theta \equiv (\theta'_1, \theta'_0)'$ . Then we have the following identification result:

**Theorem 2.2.1** (Identification). *Suppose Assumption 1 holds. Then  $\theta$  is point identified.*

*Proof.* Based on the model (2.1)–(2.2) and Assumption 1, we have

$$\begin{aligned}
\mathbb{E}[Y | D = 1, Z = z] &= \mathbb{E}[Y_1 | U \leq p(z), Z = z] \\
&= \mathbb{E}[Y_1 | U \leq p(z)] \\
&= \frac{1}{p(z)} \int_0^{p(z)} \mathbb{E}[Y_1 | U = u] du \\
&= \mu_1 + \sum_{m=1}^M \frac{\rho_{1m}}{p(z)} \int_0^{p(z)} h_m(u) du. \\
&= (\lambda_{10}(p(z)), \lambda_{11}(p(z)), \dots, \lambda_{1M}(p(z)))\theta_1. \tag{2.3}
\end{aligned}$$

The first line holds by equations (2.1), (2.2), and the normalization that  $\nu(z) = p(z)$ , which is jointly implied by Assumption 1.1 and 1.3. The second line holds by the exogeneity

condition in Assumption 1.2. The third line follows by the overlap condition in Assumption 1.4 and the normalization that  $U$  is uniformly distributed over  $[0, 1]$ . The fourth line holds by parametric restriction in Assumption 1.5. Likewise, we have

$$\mathbb{E}[Y \mid D = 0, Z = z] = \mu_0 + \sum_{m=1}^M \frac{\rho_{0m}}{1 - p(z)} \int_{p(z)}^1 h_m(u) du = (\lambda_{00}(p(z)), \lambda_{01}(p(z)), \dots, \lambda_{0M}(p(z)))\theta_0. \quad (2.4)$$

Taking  $z = z_0, z_1, \dots, z_K$  in equations (2.3) and (2.4) then yields two matrix equalities:

$$A_d \theta_d = \beta_d \quad \text{for } d = 0, 1, \quad (2.5)$$

where

$$A_d = \begin{bmatrix} \lambda_{d0}(p(z_0)) & \lambda_{d1}(p(z_0)) & \dots & \lambda_{dM}(p(z_0)) \\ \lambda_{d0}(p(z_1)) & \lambda_{d1}(p(z_1)) & \dots & \lambda_{dM}(p(z_1)) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{d0}(p(z_K)) & \lambda_{d1}(p(z_K)) & \dots & \lambda_{dM}(p(z_K)) \end{bmatrix}$$

and

$$\beta_d = (\mathbb{E}[Y \mid D = d, Z = z_0], \dots, \mathbb{E}[Y \mid D = d, Z = z_K])'.$$

Note that Assumption 1.6 and 1.7 guarantees that there exists a full-rank submatrix of  $A_d$  with  $\bar{K}$  rows and  $M + 1$  columns for each  $d = 0, 1$ . Therefore, both  $A_1$  and  $A_0$  have full column rank. Then  $\theta_1$  and  $\theta_0$  are point identified by equation (2.5).  $\square$

**Remark 2.2.1.** *This identification result closely aligns with the proof of Brinch et al. (2017, Proposition 1.ii), who shows that Assumption 1.7 is a necessary condition for identifying  $\theta$  in a MTE model satisfying Assumption 1.1–1.5. Building on their result, I add Assumption 1.6 to establish sufficient conditions for point identification.*

Based on the parametric restriction in Assumption 1.5, treatment effects can often be written as linear functions of the primitive parameter  $\theta$ . For example,  $\text{ATE} = \mu_1 - \mu_0$  and

$$\begin{aligned} \text{MTE}(u) &= \mathbb{E}[Y_1 - Y_0 \mid U = u] \\ &= c(u)' \theta_1 - c(u)' \theta_0, \end{aligned}$$

where  $c(u) = (1, h_1(u), \dots, h_M(u))$ . Moreover, suppose we are interested in the parameters that inform potential policy decisions such as average treatment effects on the (un)treated groups or policy-relevant treatment effects. In that case, the weight  $c$  is often unknown and depends on the underlying DGP (see Table 1 and 2 in Mogstad and Torgovitsky, 2018). If we write  $c = (c'_1, c'_0)'$  with  $c_d = (c_{d,0}, \dots, c_{d,M})'$  for  $d = 0, 1$  and denote  $h_0(u) \equiv 1$ , Table 2.1 summarizes the weights of various treatment effects under Assumption 1. In this paper, I consider inference on treatment effects of the form  $c'\theta$  where  $c_1 = -c_0$  since many causal effects of interest including those in Table 2.1 have symmetric weights:  $c_{0,m} = -c_{1,m}$  for all  $m = 1, \dots, M$ .

Table 2.1: Weights for Treatment Effects

Target parameter	Expression	Weights
ATE	$\mathbb{E}[Y_1 - Y_0]$	$c_{1,m} = \mathbb{1}[m = 0]$
MTE	$\mathbb{E}[Y_1 - Y_0 \mid U = u]$	$c_{1,m} = h_m(u)$
ATT	$\mathbb{E}[Y_1 - Y_0 \mid D = 1]$	$c_{1,m} = \frac{\mathbb{E}(\int_0^{p(Z)} h_m(u) du)}{\mathbb{P}(D=1)}$
ATU	$\mathbb{E}[Y_1 - Y_0 \mid D = 0]$	$c_{1,m} = \frac{\mathbb{E}(\int_p^1 h_m(u) du)}{\mathbb{P}(D=0)}$
LATE	$\mathbb{E}[Y_1 - Y_0 \mid p(z_0) < U < p(z_k)]$	$c_{1,m} = \frac{\int_{p(z_0)}^{p(z_k)} h_m(u) du}{p(z_k) - p(z_0)}$
Additive PRTE	PRTE with $p^\epsilon(z) = p(z) + \epsilon$	
Proportional PRTE	PRTE with $p^\epsilon(z) = (1 + \epsilon)p(z)$	$c_{1,m} = \frac{\mathbb{E}[\int_p^{p^\epsilon(Z)} h_m(u) du]}{\mathbb{E}[p^\epsilon(Z) - p(Z)]}$
Quota	PRTE with $p^\epsilon(z) = \min\{p(z), \epsilon\}$	

Define  $q(z_\ell) = \mathbb{P}(Z = z_\ell)$ , i.e., the probability mass function of the discrete IV. For the widely-used polynomial MTE model in empirical studies where  $h_m = u^m - \frac{1}{m+1}$ , the

weights of ATT, LATE, additive PRTE, and proportional PRTE become

$$\begin{aligned}
c_{1,m}^{\text{ATT}} &= \frac{1}{m+1} \left( \frac{\sum_{\ell=0}^K p(z_\ell)^{m+1} q(z_\ell)}{\sum_{\ell=0}^K p(z_\ell) q(z_\ell)} - 1 \right) \\
c_{1,m}^{\text{LATE}} &= \frac{1}{m+1} \left( \sum_{j=0}^m p(z_k)^j p(z_0)^{m-j} - 1 \right) \\
c_{1,m}^{\text{A-PRTE}} &= \frac{1}{m+1} \left( \sum_{\ell=0}^K q(z_\ell) \sum_{j=0}^m [p(z_\ell)^j (p(z_\ell) + \epsilon)^{m-j}] - 1 \right) \\
c_{1,m}^{\text{P-PRTE}} &= \frac{1}{m+1} \left[ \frac{(1+\epsilon)^{m+1} - 1}{\epsilon} \cdot \frac{\sum_{\ell=0}^K p(z_\ell)^{m+1} q(z_\ell)}{\sum_{\ell=0}^K p(z_\ell) q(z_\ell)} - 1 \right]
\end{aligned}$$

for  $m = 1, \dots, M$ , and the first element  $c_{1,0} = 1$  for all the above causal effects.

## 2.2.2 Weak identification

The identification strategy relies on the conditions in Assumption 1, particularly the existence of sufficiently many distinct propensity scores to point identify the structural parameter  $\theta$  through Assumption 1.7. However, variation in propensity scores is often limited. This limited variation may not be enough to guarantee correct asymptotic approximation in the construction of classical confidence sets. In section 2.3 and 2.4, I develop robust inference procedures for treatment effects of the form  $c'\theta$  without requiring the potentially restrictive Assumptions 1.7. My results have two implications. Along a sequence of DGPs:

- (1) When Assumption 1.7 holds but is close to fail, the parameters are point identified, and the robust confidence set achieves correct coverage for the causal parameter of interest;
- (2) When Assumption 1.7 fails, the parameters are partially identified, and the proposed robust confidence set maintains correct coverage of the causal parameter  $c'\theta$ , where  $\theta$  lies in the set defined by the linear system (2.5). Additionally, this uniform validity result does not rely on Assumption 1.6, though this assumption is typically satisfied under researchers' common specifications of the MTR functions.

### 2.2.3 Parameter space restriction

This section introduces the parameter space for the joint distribution of  $(Y, D, Z)$  for establishing uniform validity of the robust inference procedures without covariates. First, consider the usual i.i.d. sampling distribution of the observables.

**Assumption 2.** *The random vectors  $(Y_i, D_i, Z_i)$  for  $i = 1, \dots, n$  are i.i.d. with distribution  $F$ .*

Next, I introduce a set of regularity conditions to be imposed on the joint distribution  $F$ .

**Definition 2.2.1** (Parameter Space). *For some  $\delta, \zeta > 0$  and  $\epsilon \in (0, 1/2)$ , define the parameter space  $\mathcal{P}$  as the set of pairs  $(\theta, F)$  satisfying the following properties:*

1. *Equation (2.5) is satisfied with  $K \geq M$ , where  $\theta = (\theta'_1, \theta'_0)' \in \text{int}(\Theta) \subseteq \mathbb{R}^{2(M+1)}$  for some interior of a compact set  $\Theta$ ;*
2.  $\sup_{d=0,1} \sup_{z \in \text{supp}(Z)} \mathbb{E}_F[|Y|^{2+\delta} \mid D = d, Z = z] \leq \zeta$ ;
3.  $\epsilon \leq \inf_{z \in \text{supp}(Z)} \mathbb{P}_F(D = 1 \mid Z = z) \leq \sup_{z \in \text{supp}(Z)} \mathbb{P}_F(D = 1 \mid Z = z) \leq 1 - \epsilon$ ;
4.  $\epsilon \leq \inf_{z \in \text{supp}(Z)} \mathbb{P}_F(Z = z) \leq \sup_{z \in \text{supp}(Z)} \mathbb{P}_F(Z = z) \leq 1 - \epsilon$ ;
5.  $\epsilon \leq \inf_{d=0,1} \inf_{z \in \text{supp}(Z)} \text{var}_F(Y \mid D = d, Z = z)$ .

The first condition assumes the correct model specification derived from Assumptions 1.1-1.5, and the order of the MTE model does not exceed the support of the instrument.<sup>3</sup> The second condition is a mild restriction on the existence of moments of  $Y$  conditional on  $D$  and  $Z$ , which is essential for establishing the uniform version of the law of large numbers and central limit theorems. The third and fourth conditions, also known as the strong overlap conditions, require observing units for each value of the treatment and the instrument. Together with the final condition regarding sufficient variation in outcomes, these conditions rule out the singularity or near-singularity of the asymptotic variance of the moment equation (2.5). Since our parameter of interest has the form of  $c'\theta$ , let

---

<sup>3</sup> The condition  $K \geq M$  is necessary but not sufficient for point identification, particularly when Assumption 1.7 does not hold.

$\mathcal{P}_0 = \{(\lambda, F) : \lambda = c'\theta, (\theta, F) \in \mathcal{P}\}$  denote the null parameter space for this linear function, in which the weight  $c$  could possibly depend on underlying DGP  $F$ . It is important to emphasize that our parameter space does not require the relevance Assumptions 1.6 and 1.7, the latter of which is likely to fail. For a given  $\lambda \in \mathbb{R}$ , our goal is to develop uniformly valid tests for assessing the hypothesis  $H_0 : c'\theta = \lambda$  that is robust to the limited variation of propensity scores.

## 2.2.4 Notation and preliminaries

The asymptotic behavior of test statistics to be proposed will be unified by the asymptotic distribution of estimators on marginal distribution of instrument  $q(z_\ell) \equiv \mathbb{P}(Z = z_\ell)$ , propensity score  $p(z_\ell) = \mathbb{P}(D = 1 \mid Z = z_\ell)$ , and the expected outcome conditional on the treatment and instrument  $\beta_{d\ell} = \mathbb{E}[Y \mid D = d, Z = z_\ell]$  for  $d = 0, 1$  and  $\ell = 0, 1, \dots, K$ . Under the parameter space restriction outlined in section 2.2.3, we have the following sample-analog estimators for these quantities:

$$\begin{aligned}\hat{q}(z_\ell) &= \frac{1}{n} \sum_{i=1}^n \mathbb{1}[Z_i = z_\ell] \\ \hat{p}(z_\ell) &= \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}[D_i = 1, Z_i = z_\ell]}{\hat{q}(z_\ell)} \\ \hat{\beta}_{d\ell} &= \frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]}{\hat{q}(d, z_\ell)},\end{aligned}$$

where  $\hat{q}(1, z_\ell) = \hat{p}(z_\ell)\hat{q}(z_\ell)$  and  $\hat{q}(0, z_\ell) = (1 - \hat{p}(z_\ell))\hat{q}(z_\ell)$ .

Collecting these estimators into vectors, we have

$$\begin{aligned}\hat{p} &= (\hat{p}(z_\ell), \dots, \hat{p}(z_K))' \\ \hat{q} &= (\hat{q}(z_0), \dots, \hat{q}(z_K))' \\ \hat{\beta}_d &= (\hat{\beta}_{d0}, \dots, \hat{\beta}_{dK})' \text{ for } d = 0, 1,\end{aligned}$$

and  $\hat{\beta} = (\hat{\beta}'_1, \hat{\beta}'_0)'$ . Lemma A.1.1(a) provides the asymptotic distribution of the estimators  $\hat{p}$ ,  $\hat{q}$ , and  $\hat{\beta}$  under a drifting sequence in the parameter space  $\mathcal{P}$ . Moreover, Lemma A.1.1(b)

also shows that their asymptotic variances can be consistently estimated by

$$\begin{aligned}\hat{\Sigma}_p &= \text{diag} \left\{ \frac{\hat{p}(z_\ell)(1 - \hat{p}(z_\ell))}{\hat{q}(z_\ell)} : \ell = 0, 1, \dots, K \right\} \\ \hat{\Sigma}_q &= \{\hat{\Sigma}_q[i, j]\}_{i, j=0, 1, \dots, K} \\ \hat{\Sigma}_{\beta_d} &= \text{diag} \left\{ \frac{\hat{\sigma}_{d\ell}^2}{\hat{q}(d, z_\ell)} : \ell = 0, 1, \dots, K \right\} \\ \hat{\Sigma}_\beta &= \text{diag}\{\hat{\Sigma}_{\beta_1}, \hat{\Sigma}_{\beta_0}\},\end{aligned}$$

where

$$\begin{aligned}\hat{\sigma}_{d\ell}^2 &\equiv \frac{1}{n} \sum_{i=1}^n \frac{(Y_i - \hat{\beta}_{d\ell})^2 \mathbb{1}[D_i = d, Z_i = z_\ell]}{\hat{q}(d, z_\ell)} \\ \hat{\Sigma}_q[i, j] &\equiv \begin{cases} \hat{p}(z_i)(1 - \hat{p}(z_i)) & \text{if } i = j \\ -\hat{p}(z_i)\hat{p}(z_j) & \text{if } i \neq j. \end{cases}\end{aligned}$$

Throughout the paper, I use  $\partial_x f \in \mathbb{R}^k$  to denote the gradient of a scalar function  $f$  with respect to its argument  $x \in \mathbb{R}^k$ . If  $f$  maps to a vector in  $\mathbb{R}^l$ , then  $\partial_x f \in \mathbb{R}^{l \times k}$  denotes the Jacobian of  $f$  with respect to  $x \in \mathbb{R}^k$ . Let  $P_A = A(A'A)^{-1}A'$  denote the projection matrix onto the column spaces of matrix  $A$  and define  $M_A = I - P_A$  as the corresponding annihilator matrix.

### 2.3 Robust Inference in Linear MTE Models

In this section, I start with the simplest functional specification and impose linearity on the MTR function.

**Assumption 3** (Linear MTE model). *Assumption 1.5 holds with*

$$h_m(u) = u - \frac{1}{2}$$

and  $M = 1$ .

This assumption was first introduced by Olsen (1980) to characterize sample selection bias and later generalized by Brinch et al. (2017) to model MTE functions. This linear

MTE specification was recently adopted by Kowalski (2023) to extrapolate treatment effects between two experimental studies. The parameters  $\theta = (\mu_1, \rho_{11}, \mu_0, \rho_{01})$  can be identified using any pair of propensity scores  $(p(z_0), p(z_k))$  for  $k = 1, \dots, K$  that differ from each other. Point identification fails if and only if all propensity scores  $p(z_0), \dots, p(z_K)$  are identical, i.e., when there is no variation in propensity scores.

The linear MTE specification allows us to derive closed-form estimands for  $\theta$ , in which case equation (2.5) reduces to

$$\begin{pmatrix} \mu_1 \\ \rho_{11} \end{pmatrix} = \frac{1}{p(z_k) - p(z_0)} \begin{pmatrix} p(z_k) - 1 & -(p(z_0) - 1) \\ -2 & 2 \end{pmatrix} \begin{pmatrix} \mathbb{E}[Y | D = 1, Z = z_0] \\ \mathbb{E}[Y | D = 1, Z = z_k] \end{pmatrix}$$

and

$$\begin{pmatrix} \mu_0 \\ \rho_{01} \end{pmatrix} = \frac{1}{p(z_k) - p(z_0)} \begin{pmatrix} p(z_k) & -p(z_0) \\ -2 & 2 \end{pmatrix} \begin{pmatrix} \mathbb{E}[Y | D = 0, Z = z_0] \\ \mathbb{E}[Y | D = 0, Z = z_k] \end{pmatrix}.$$

Recall  $\beta_{dk} = \mathbb{E}[Y | D = d, Z = z_k]$ . Then the ATE and the slope of MTE can be written as

$$\begin{pmatrix} \mu_1 - \mu_0 \\ \rho_{11} - \rho_{01} \end{pmatrix} = \frac{1}{p(z_k) - p(z_0)} \begin{pmatrix} p(z_k) [\beta_{10} - \beta_{00}] - p(z_0) [\beta_{1k} - \beta_{0k}] + \beta_{1k} - \beta_{10} \\ 2(\beta_{00} - \beta_{10} + \beta_{1k} - \beta_{0k}) \end{pmatrix}.$$

Define the numerator of the right-hand-side equation as

$$\Delta_\mu(z_0, z_k) \equiv p(z_k) [\beta_{10} - \beta_{00}] - p(z_0) [\beta_{1k} - \beta_{0k}] + \beta_{1k} - \beta_{10}$$

and

$$\Delta_\rho(z_0, z_k) \equiv 2(\beta_{00} - \beta_{10} + \beta_{1k} - \beta_{0k}).$$

Since these quantities can be directly identified from data, we can express the treatment effects parameters  $\lambda = c'\theta$  with  $c = (c_\mu, c_\rho, -c_\mu, -c_\rho)'$  as the solution of the following moment function:

$$g_k(\lambda) = [p(z_k) - p(z_0)] \lambda - c_\mu \Delta_\mu(z_0, z_k) - c_\rho \Delta_\rho(z_0, z_k) = 0. \quad (2.6)$$

If the instrument  $Z$  is binary, then  $c'\theta$  is just identified by this linear moment restriction, otherwise, we can construct a vector of moment functions:

$$g(\lambda) = (g_1(\lambda), \dots, g_K(\lambda))' = 0_{K \times 1} \quad (2.7)$$

to improve the efficiency for inference on  $c'\theta$ . It is worth noting that other components of primitive parameter  $\theta$  do not enter into (2.7). As a result, their nonstandard asymptotic behavior (under weak identification) is irrelevant in this context. To this end, this linear moment function  $g(\lambda)$  will be used to construct asymptotically exact tests for treatment effects in linear MTE models.

### 2.3.1 Inference with known weights

In this section, I consider robust inference with a known weight  $c$ , which applies to inference on the ATE and the MTE. I have shown that  $g_k(\lambda)$  suffices to estimate  $\lambda$ , the causal effects of interest. Let  $\hat{g}_k(\lambda)$  be the sample analog of this moment function after plugging in the estimators  $\{\hat{p}(z_\ell), \hat{\beta}_{0\ell}, \hat{\beta}_{1\ell}\}_{\ell=0}^K$ , i.e.,

$$\hat{g}_k(\lambda) = [\hat{p}(z_k) - \hat{p}(z_0)] \lambda - c_\mu \hat{\Delta}_\mu(z_0, z_k) - c_\rho \hat{\Delta}_\rho(z_0, z_k)$$

where

$$\begin{aligned} \hat{\Delta}_\mu(z_0, z_k) &= \hat{p}(z_k)[\hat{\beta}_{10} - \hat{\beta}_{00}] - \hat{p}(z_0)[\hat{\beta}_{1k} - \hat{\beta}_{0k}] + \hat{\beta}_{1k} - \hat{\beta}_{10} \\ \hat{\Delta}_\rho(z_0, z_k) &= 2(\hat{\beta}_{00} - \hat{\beta}_{10} + \hat{\beta}_{1k} - \hat{\beta}_{0k}). \end{aligned}$$

Based on this moment condition, one might consider an Anderson-Rubin (AR) test statistic to assess the null hypothesis  $H_0 : c'\theta = \lambda$  as follows

$$\text{AR}_{n,k}(\lambda) = \left| \frac{\sqrt{n} \hat{g}_k(\lambda)}{\hat{s}_k(\lambda)} \right|^2 \quad (2.8)$$

where  $\hat{s}_k^2(\lambda)$  consistently estimates the asymptotic variance of the sample moment  $\hat{g}_k(\lambda)$ . To construct this variance estimator, note that an asymptotic linear expansion of  $\hat{g}_k(\lambda) - g_k(\lambda)$  gives

$$\hat{g}_k(\lambda) - g_k(\lambda) = \begin{pmatrix} \partial_{p'} g_k(\lambda) [\hat{p} - p] \\ + \partial_{\beta_1'} g_k(\lambda) [\hat{\beta}_1 - \beta_1] \\ + \partial_{\beta_0'} g_k(\lambda) [\hat{\beta}_0 - \beta_0] \end{pmatrix} + o_p(n^{-1/2}).$$

with coefficients  $\partial_p g_k(\lambda)$ ,  $\partial_{\beta_1} g_k(\lambda)$ , and  $\partial_{\beta_0} g_k(\lambda)$  defined as the gradient of  $g_k(\lambda)$  with respect to  $p$ ,  $\beta_1$ , and  $\beta_0$ , respectively. To estimate these coefficients, we have the sample

analog estimators below:

$$\begin{aligned}
\partial_p \hat{g}_k(\lambda) &= (-\lambda + (\hat{\beta}_{1k} - \hat{\beta}_{0k})c_\mu, 0, \dots, 0, \lambda - (\hat{\beta}_{10} - \hat{\beta}_{00})c_\mu, 0, \dots, 0)' \\
\partial_{\beta_1} \hat{g}_k(\lambda) &= ((1 - \hat{p}(z_k))c_\mu + 2c_\rho, 0, \dots, 0, -(1 - \hat{p}(z_0))c_\mu - 2c_\rho, 0, \dots, 0)' \\
\partial_{\beta_0} \hat{g}_k(\lambda) &= (\hat{p}(z_k)c_\mu - 2c_\rho, 0, \dots, 0, -\hat{p}(z_0)c_\mu + 2c_\rho, 0, \dots, 0)',
\end{aligned} \tag{2.9}$$

where nonzero elements appear on the first and the  $(k + 1)$ 'th elements of vectors.

This leads to a consistent variance estimator

$$\hat{s}_k^2(\lambda) = \partial_{p'} \hat{g}_k(\lambda) \hat{\Sigma}_p \partial_p \hat{g}_k(\lambda) + \partial_{\beta_1'} \hat{g}_k(\lambda) \hat{\Sigma}_{\beta_1} \partial_{\beta_1} \hat{g}_k(\lambda) + \partial_{\beta_0'} \hat{g}_k(\lambda) \hat{\Sigma}_{\beta_0} \partial_{\beta_0} \hat{g}_k(\lambda).$$

Let  $\alpha \in (0, 1)$  be the significant level. The next result establishes the uniform validity and asymptotic similarity of the AR test for  $H_0 : c'\theta = \lambda$  with  $\lambda \in \mathbb{R}$ .

**Proposition 2.3.1.** *Let Assumption 2 and 3 hold, and suppose that the weight  $c = (c_\mu, c_\rho, -c_\mu, -c_\rho)'$  is a nonzero fixed vector, then*

$$\begin{aligned}
&\liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F \left( AR_{n,k}(\lambda) > q_{\chi_1^2}(1 - \alpha) \right) \\
&= \limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F \left( AR_{n,k}(\lambda) > q_{\chi_1^2}(1 - \alpha) \right) = \alpha,
\end{aligned}$$

where  $q_{\chi_1^2}(1 - \alpha)$  denotes the  $(1 - \alpha)$ -quantile of the  $\chi_1^2$  distribution.

**Remark 2.3.1.** *The above result shows that the proposed AR test is valid and (uniformly) asymptotically similar in the sense of D. W. Andrews et al. (2020, eq.(2.6)). Constructing an asymptotically valid subvector test is already challenging in models with weakly identified nuisance parameters<sup>4</sup>, because the asymptotic distributions of many test statistics depend on those unknown nuisance parameters that cannot be consistently estimated (I. Andrews & Mikusheva, 2016b, p. 1595). However, I introduce a novel moment function that isolates the causal effects of interest while being free from  $\theta$ , thereby making the simple AR*

<sup>4</sup> Since  $c'\theta$  is the parameter of interest rather than  $\theta$  itself, the vector of primitive parameters  $\theta = (\mu_1, \rho_{11}, \mu_0, \rho_{01}) \in \mathbb{R}^4$  becomes the weakly identified nuisance parameter under limited variation of propensity scores. Although we can show some functions of  $\theta$  are strongly identified using the reparametrization technique proposed by Han and McCloskey (2019), two parameters  $(\rho_{11}, \rho_{01}) \in \mathbb{R}^2$  remain weakly identified even after this transformation (see Appendix A.8).

test feasible in this context. Unlike existing approaches to subvector inference with weak identification (see Appendix A.7.1), this method achieves asymptotic similarity, ensuring the inverted confidence set maintains  $1 - \alpha$  coverage asymptotically under both strong and weak identification.

With a binary instrument  $Z \in \{z_0, z_1\}$ , there exists a unique AR test statistic  $\text{AR}_{n,1}(\lambda)$  for robust inference on the causal parameter  $c'\theta$ . When the instrument is discrete with multiple values  $Z \in \{z_0, \dots, z_K\}$  where  $K > 1$ , we obtain multiple AR statistics  $\{\text{AR}_{n,k}(\lambda)\}_{k=1}^K$ . The informativeness of each statistic depends on the variation in propensity scores  $p(z_0) - p(z_k)$ . In this case, combining tests based on different propensity score pairs yields greater statistical power. Let

$$\hat{\pi} = (\hat{p}(z_1) - \hat{p}(z_0), \dots, \hat{p}(z_K) - \hat{p}(z_0))'$$

and

$$\hat{\gamma} = (c_\mu \hat{\Delta}_\mu(z_0, z_1) + c_\rho \hat{\Delta}_\rho(z_0, z_1), \dots, c_\mu \hat{\Delta}_\mu(z_0, z_K) + c_\rho \hat{\Delta}_\rho(z_0, z_K))'$$

Then consider valid inference based on a vector of linear moment functions:

$$\hat{g}(\lambda) = \hat{\pi}\lambda - \hat{\gamma} = (\hat{g}_1(\lambda), \dots, \hat{g}_K(\lambda))' \in \mathbb{R}^K.$$

Under strong identification where  $\hat{\pi}$  converges to a nonzero limit, an asymptotically efficient Wald statistic (which becomes the Lagrangian Multiplier statistic in this linear case) for testing  $H_0 : c'\theta = \lambda$  can be constructed below:

$$W_n(\lambda) = \frac{n\hat{g}(\lambda)' \hat{S}(\lambda)^{-1} \hat{\pi} \hat{\pi}' \hat{S}(\lambda)^{-1} \hat{g}(\lambda)}{\hat{\pi}' \hat{S}(\lambda)^{-1} \hat{\pi}},$$

where  $\hat{S}(\lambda)$  consistently estimates the asymptotic covariance matrix of  $\hat{\pi}\lambda - \hat{\gamma}$  under null hypothesis:

$$\hat{S}(\lambda) = \partial_p \hat{g}(\lambda) \hat{\Sigma}_p \partial_p' \hat{g}(\lambda) + \partial_{\beta_1} \hat{g}(\lambda) \hat{\Sigma}_{\beta_1} \partial_{\beta_1}' \hat{g}(\lambda) + \partial_{\beta_0} \hat{g}(\lambda) \hat{\Sigma}_{\beta_0} \partial_{\beta_0}' \hat{g}(\lambda).$$

Under weak identification, the propensity score differences  $\hat{\pi}$  may converge in probability to a zero vector, resulting in a nonstandard asymptotic distribution for  $W_n(\lambda)$ . To illustrate

this problem, consider a sequence of DGPs where  $\sqrt{n}\pi$  converges to a constant vector  $\pi_0$ . In this case,  $\sqrt{n}\hat{\pi}$  converges in distribution to a multivariate normal distribution with mean  $\pi_0$ , which cannot be consistently estimated from the data. Consequently, the asymptotic distribution of  $W_n(\lambda)$  contains the inestimable nuisance parameter  $\pi_0$ , making it infeasible to consistently estimate the unconditional quantiles of  $W_n(\lambda)$ 's limiting distribution. However, similar to the arguments of Moreira (2003) for linear IV models and I. Andrews and Mikusheva (2016b) for quasi-likelihood ratio tests in GMM models, the asymptotic distribution of  $W_n(\lambda)$  becomes independent of  $\pi_0$  when conditioned on its sufficient statistic, making it feasible to approximate the conditional distribution of  $W_n(\lambda)$ . Define

$$\hat{h}(\lambda) = \sqrt{n}\hat{\pi} - [\partial_p\pi] \hat{\Sigma}_p[\partial_{p'}\hat{g}(\lambda)] \hat{S}(\lambda)^{-1} \sqrt{n}\hat{g}(\lambda).$$

The statistic  $\hat{h}(\lambda)$  is asymptotically independent of  $\sqrt{n}\hat{g}(\lambda)$  and contains sufficient information about  $\pi_0$  such that the (asymptotic) distribution of  $W_n(\lambda)$  conditional on  $\hat{h}(\lambda)$  is free of  $\pi_0$ . To simulate this conditional distribution, I construct a counterpart of  $\sqrt{n}\hat{\pi}$ , denoted as  $\pi_s$ :

$$\pi_s = \hat{h}(\lambda) + [\partial_p\pi] \hat{\Sigma}_p[\partial_{p'}\hat{g}(\lambda)] \hat{S}(\lambda)^{-1/2} \eta^*$$

where  $\eta^* \sim \mathcal{N}(0_{K \times 1}, I_{K \times K})$  are simulated draws independent of data. A simulation counterpart of  $W_n(\lambda)$  can be obtained by replacing  $\sqrt{n}\hat{\pi}$  with  $\pi_s$  and  $\sqrt{n}\hat{g}(\lambda)$  with  $\hat{S}(\lambda)^{1/2}\eta^*$ :

$$W_n^*(\lambda) = \frac{(\eta^*)' \hat{S}(\lambda)^{-1/2} \pi_s \pi_s' \hat{S}(\lambda)^{-1/2} (\eta^*)}{\pi_s' \hat{S}(\lambda)^{-1} \pi_s}.$$

It follows that  $W_n^*(\lambda)$  has the same asymptotic distribution as  $W_n(\lambda)$  conditional on the realization of  $\hat{h}(\lambda)$ . Let  $\hat{q}_{W^*}(1 - \alpha)$  denote the  $(1 - \alpha)$ -quantile of the distribution of  $W_n^*(\lambda)$  conditional on the data. This critical value can be used to construct valid conditional Wald test:

**Theorem 2.3.1.** *Let Assumption 2 and 3 hold, and suppose the weight  $c = (c_\mu, c_\rho, -c_\mu, -c_\rho)'$  is a nonzero fixed vector, then*

$$\liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(W_n(\lambda) > \hat{q}_{W^*}(1 - \alpha)) = \limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(W_n(\lambda) > \hat{q}_{W^*}(1 - \alpha)) = \alpha.$$

**Remark 2.3.2.** *The statistic  $W_n(\lambda)$  can be interpreted as a linear combination of AR statistics in equation (2.8), with weight  $\hat{S}^{-1}\hat{\pi}$ . Under strong identification, this weighting vector can be shown to be optimal among all linear combinations of AR statistics, yielding the highest local asymptotic power.*

**Remark 2.3.3.** *In Lemma A.1.2, I show that the asymptotic variance of the moment function is uniformly positive definite as long as the weight  $c = (c_\mu, c_\rho, -c_\mu, -c_\rho)$  is nonzero. If we incorporate additional moment conditions induced by the difference of propensity scores of the form  $p(z_k) - p(z_j)$  for  $k, j \neq 0$ , then the asymptotic variance may become singular for some nonzero weight  $c$ . For example, if we set  $c_\mu = 0$  and  $c_\rho = 1$ , it follows that*

$$[\hat{p}(z_k) - \hat{p}(z_j)]\lambda - c_\rho \hat{\Delta}_\rho(z_j, z_k) = \hat{g}_k(\lambda) - \hat{g}_j(\lambda),$$

*implying that the moment condition constructed with the variation between  $z_k$  and  $z_j$  (on the left-hand side) is the difference of the moments constructed by using  $(z_k, z_0)$  and  $(z_j, z_0)$  (on the right-hand side).*

**Remark 2.3.4.** *An alternative statistic one could consider is the quasi-likelihood ratio statistic*

$$QLR_n(\lambda) = n \left[ \hat{g}(\lambda)' \hat{S}(\lambda)^{-1} \hat{g}(\lambda) - \inf_{\lambda \in \mathbb{R}} \hat{g}(\lambda)' \hat{S}(\lambda)^{-1} \hat{g}(\lambda) \right]$$

*and its corresponding conditional approach discussed in I. Andrews and Mikusheva (2016b). Here I focus on the conditional Wald approach because it is simpler to implement in practice and is first-order asymptotic equivalent to QLR test under strong identification. Furthermore, there is evidence that, in finite samples, it can yield shorter confidence intervals than the conditional likelihood ratio test and AR test (D. S. Lee et al., 2023; Van de Sijpe & Windmeijer, 2023).*

## 2.3.2 Inference with estimated weights

Sometimes the weight  $c$  needs to be estimated if researchers' interests focus on causal parameters such as ATT and LATE. In this case, the weight usually depends on the joint

distribution of  $(D, Z)$  as in Table 2.1, therefore I make the following assumption:

**Assumption 4.** *The weight  $c = c(p, q)$  is a function of propensity scores  $p = \{p(z_\ell)\}_{\ell=0}^K$  and marginal distribution of instrument  $q = \{q(z_\ell)\}_{\ell=0}^K$ , and this function is continuously differentiable.*

I plug a consistent and asymptotic normal estimator  $c(\hat{p}, \hat{q})$  into the construction of the moment function. This gives a modified sample moment function:

$$\hat{g}_k^\dagger(\lambda) = [\hat{p}(z_k) - \hat{p}(z_0)] \lambda - c_\mu(\hat{p}, \hat{q}) \hat{\Delta}_\mu(z_0, z_k) - c_\rho(\hat{p}, \hat{q}) \hat{\Delta}_\rho(z_0, z_k)$$

and

$$\hat{g}^\dagger(\lambda) = (\hat{g}_1^\dagger(\lambda), \dots, \hat{g}_K^\dagger(\lambda))' \in \mathbb{R}^K.$$

Since the weight  $c$  is estimated, it introduces sampling uncertainty into the asymptotic distribution of the moment function. Let

$$\Delta_\mu = (\Delta_\mu(z_0, z_1), \dots, \Delta_\mu(z_0, z_K))' \quad \text{and} \quad \Delta_\rho = (\Delta_\rho(z_0, z_1), \dots, \Delta_\rho(z_0, z_K))'$$

Define  $\hat{\Delta}_\mu$  and  $\hat{\Delta}_\rho$  as their corresponding estimators, where each entry is obtained by replacing  $\Delta_\mu(z_0, z_k)$  and  $\Delta_\rho(z_0, z_k)$  with their respective estimators  $\hat{\Delta}_\mu(z_0, z_k)$  and  $\hat{\Delta}_\rho(z_0, z_k)$ .

Applying the Delta method, we have the following first-order asymptotic expansion:

$$\hat{g}^\dagger(\lambda) - g(\lambda) = \begin{pmatrix} \partial_p g(\lambda)[\hat{p} - p] \\ + \partial_{\beta_1} g(\lambda)[\hat{\beta}_1 - \beta_1] \\ + \partial_{\beta_0} g(\lambda)[\hat{\beta}_0 - \beta_0] \\ - \Delta_\mu(\partial_{p'} c_\mu[\hat{p} - p] + \partial_{q'} c_\mu[\hat{q} - q]) \\ - \Delta_\rho(\partial_{p'} c_\rho[\hat{p} - p] + \partial_{q'} c_\rho[\hat{q} - q]) \end{pmatrix} + o_p(n^{-1/2}),$$

where  $\partial_p c$  and  $\partial_q c$  denote the partial derivative vectors of  $c$  with respect to vectors  $p$  and  $q$ , respectively, for  $c \in \{c_\mu(p, q), c_\rho(p, q)\}$ .

The asymptotic variance can be consistently estimated by

$$\begin{aligned} \hat{S}^\dagger(\lambda) &= \left( \partial_p \hat{g}^\dagger(\lambda) - \hat{\Delta}_\mu[\partial_{p'} \hat{c}_\mu] - \hat{\Delta}_\rho[\partial_{p'} \hat{c}_\rho] \right) \hat{\Sigma}_p \left( \partial_p \hat{g}^\dagger(\lambda) - \hat{\Delta}_\mu[\partial_{p'} \hat{c}_\mu] - \hat{\Delta}_\rho[\partial_{p'} \hat{c}_\rho] \right)' \\ &+ \left( \hat{\Delta}_\mu[\partial_{q'} \hat{c}_\mu] + \hat{\Delta}_\rho[\partial_{q'} \hat{c}_\rho] \right) \hat{\Sigma}_q \left( \hat{\Delta}_\mu[\partial_{q'} \hat{c}_\mu] + \hat{\Delta}_\rho[\partial_{q'} \hat{c}_\rho] \right)' \\ &+ \partial_{\beta_0} \hat{g}^\dagger(\lambda) \hat{\Sigma}_{\beta_0} \partial_{\beta_0}' \hat{g}^\dagger(\lambda) + \partial_{\beta_1} \hat{g}^\dagger(\lambda) \hat{\Sigma}_{\beta_1} \partial_{\beta_1}' \hat{g}^\dagger(\lambda), \end{aligned}$$

where  $\partial_x \hat{g}^\dagger(\lambda)$  is obtained from  $\partial_x g(\lambda)$  by replacing  $c$  with  $\hat{c}$  for  $x \in \{p, \beta_0, \beta_1\}$ , and  $\partial_x \hat{c}_\mu$  (or  $\partial_x \hat{c}_\rho$ ) denotes the sample analog of  $\partial_x c_\mu$  (or  $\partial_x c_\rho$ ) evaluated at  $(\hat{p}, \hat{q})$  for  $x \in \{p, q\}$ .

To achieve asymptotic independence from  $\sqrt{n} \hat{g}^\dagger(\lambda)$ , I modify the sufficient statistic  $\hat{h}(\lambda)$  to

$$\hat{h}^\dagger(\lambda) = \sqrt{n} \hat{\pi} - [\partial_p \pi] \hat{\Sigma}_p \left( \partial_{p'} \hat{g}^\dagger(\lambda) - \hat{\Delta}_\mu [\partial_{p'} \hat{c}_\mu] - \hat{\Delta}_\rho [\partial_{p'} \hat{c}_\rho] \right) \hat{S}^\dagger(\lambda)^{-1} \sqrt{n} \hat{g}^\dagger(\lambda).$$

The simulation counterpart of  $\hat{\pi}_s$  can then be constructed as

$$\pi_s^\dagger = \hat{h}^\dagger(\lambda) + [\partial_p \pi] \hat{\Sigma}_p \left( \partial_{p'} \hat{g}^\dagger(\lambda) - \hat{\Delta}_\mu [\partial_{p'} \hat{c}_\mu] - \hat{\Delta}_\rho [\partial_{p'} \hat{c}_\rho] \right) \hat{S}^\dagger(\lambda)^{-1/2} \eta^*,$$

where  $\eta^* \sim \mathcal{N}(0_{K \times 1}, I_{K \times K})$  are simulated draws independent of data. Under these modifications, the asymptotic distribution of the plug-in Wald statistic

$$W_n^\dagger(\lambda) = \frac{n \hat{g}^\dagger(\lambda) \hat{S}^\dagger(\lambda)^{-1} \hat{\pi} \hat{\pi}' \hat{S}^\dagger(\lambda)^{-1} \hat{g}^\dagger(\lambda)}{\hat{\pi}' \hat{S}^\dagger(\lambda)^{-1} \hat{\pi}}$$

has the same conditional distribution as

$$\frac{[\eta^*]' \hat{S}^\dagger(\lambda)^{-1/2} [\pi_s^\dagger] [\pi_s^\dagger]' \hat{S}^\dagger(\lambda)^{-1/2} [\eta^*]}{[\pi_s^\dagger]' \hat{S}^\dagger(\lambda)^{-1} [\pi_s^\dagger]}$$

when conditioning on  $\hat{h}^\dagger(\lambda)$  as sample size diverges. This leads to the following corollary regarding the uniform validity of the modified testing procedure:

**Corollary 2.3.1.** *Consider the weight  $c$  that depends on the data distribution  $F$ . Let Assumptions 2, 3 and 4 hold, and suppose that  $\inf_{(\theta, F) \in \mathcal{P}} \|c(p_F, q_F)\| > 0$ . Replacing  $(\hat{g}(\lambda), \hat{S}(\lambda), \pi_s)$  with their counterparts  $(\hat{g}^\dagger(\lambda), \hat{S}^\dagger(\lambda), \pi_s^\dagger)$  in Theorem 2.3.1 yields the same conclusion.*

**Remark 2.3.5.** *We can also modify the AR test described in Proposition 2.3.1 by replacing  $\hat{g}_k(\lambda)$  with  $\hat{g}_k^\dagger(\lambda)$  and replacing  $\hat{s}_k(\lambda)$  with the  $(k, k)$ -th diagonal element of  $\hat{S}^\dagger(\lambda)$  to make it asymptotically valid with the estimated weight.*

## 2.4 Robust Inference in Polynomial MTE Models

In this section, I relax the linearity Assumption 3 and extend the analysis to allow for any functional forms specified in Assumption 1.5. This includes the linear MTE model and the polynomial specification  $h_m(\cdot) = u^m - \frac{1}{m+1}$  that is commonly used in empirical applications.

For this broader class of models, constructing moment conditions that only involve the target causal parameter is less obvious. Therefore, I develop an alternative *improved projection* approach that builds on I. Andrews (2018) to achieve valid inference. When compared to the conditional Wald test in section 2.3, this new approach applies to a wider range of MTE models but is more computationally intensive and is potentially more conservative. On the theoretical front, I argue that the high-level conditions imposed by I. Andrews (2018) cannot be directly verified in the MTE setting. Therefore, I extend his sequential validity result (that based on high-level conditions) by establishing the uniform validity under primitive conditions on the parameter space outlined in Definition 2.2.1.

### 2.4.1 Improved projection inference

First, I describe how to adapt the improved projection test developed by I. Andrews (2018) for conducting inference on  $c'\theta$  using the following linear system of equations:

$$A\theta = \beta \tag{2.10}$$

where

$$A = \begin{pmatrix} A_1 & 0_{(K+1) \times (M+1)} \\ 0_{(K+1) \times (M+1)} & A_0 \end{pmatrix} \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_0 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_0 \end{pmatrix},$$

as defined below equation (2.5). Note that  $A$  is a matrix of transformed propensity scores and  $\beta$  is a vector of conditional expectations, both of which can be consistently estimated by their sample analogs  $\hat{A}$  and  $\hat{\beta}$  under appropriate assumptions. Since the matrix  $A$  captures the variation in propensity scores, it determines the strength of identification.

The conventional Wald test for conducting inference on  $c'\theta$  uses the first-order efficient

estimator obtained by minimizing the following minimum-distance objective function:

$$\hat{\theta}^{\text{eff}} \in \underset{\theta \in \Theta}{\operatorname{argmin}} n(\hat{A}\theta - \hat{\beta})' \hat{\Omega}(\theta)^{-1} (\hat{A}\theta - \hat{\beta})$$

where  $\hat{\Omega}(\theta)$  is a consistent estimator for the asymptotic variance of  $\sqrt{n}(\hat{A}\theta - \hat{\beta})$ , formally defined in equation (2.18). Then the classical Wald statistic for assessing  $H_0 : c'\theta = \lambda$  is

$$\text{Wald}_n(\lambda) = n(c'\hat{\theta}^{\text{eff}} - \lambda)' (c'(\hat{A}'\hat{\Omega}(\hat{\theta}^{\text{eff}})^{-1}\hat{A})^{-1}c)^{-1} (c'\hat{\theta}^{\text{eff}} - \lambda). \quad (2.11)$$

When Assumption 1.7 is nearly violated,  $\hat{A}$  may converge in probability to a matrix with deficient rank. For example, consider the following simple example on the linear MTE model with a binary IV.

**Example 2.4.1** (Linear MTE model with a binary IV). *Suppose  $K = M = 1$  and set  $h_1(u) = u - \frac{1}{2}$ . Along a sequence of DGPs  $\{F_n\}$ , let  $p_{F_n}(z_0) = p \in (0, 1)$  and  $p_{F_n}(z_1) = p + v_n \in (0, 1)$  with  $v_n \rightarrow 0$ . In such case, the probability limit of  $\hat{p}(z_0)$  and  $\hat{p}(z_1)$  are equal to  $p$ , and we have*

$$\hat{A} = \begin{pmatrix} 1 & \frac{1}{2}(\hat{p}(z_0) - 1) & & 0_{2 \times 2} \\ 1 & \frac{1}{2}(\hat{p}(z_1) - 1) & & 0_{2 \times 2} \\ & & 0_{2 \times 2} & 1 \\ & & & \frac{1}{2}\hat{p}(z_0) \\ & & & \frac{1}{2}\hat{p}(z_1) \end{pmatrix} \xrightarrow{p} \begin{pmatrix} 1 & \frac{1}{2}(p - 1) & & 0_{2 \times 2} \\ 1 & \frac{1}{2}(p - 1) & & 0_{2 \times 2} \\ & & 0_{2 \times 2} & 1 \\ & & & \frac{1}{2}p \\ & & & \frac{1}{2}p \end{pmatrix}$$

Note that the probability limit of  $\hat{A}$  becomes a singular matrix.

This singularity results in multiple minimizers of the limiting objective function when defining  $\hat{\theta}^{\text{eff}}$ , suggesting that the efficient estimator  $\hat{\theta}^{\text{eff}}$  may not exhibit the usual properties of consistency and asymptotic normality. To address this issue, one can derive inference results based on the moment condition:

$$m(\theta) \equiv A\theta - \beta = 0_{2(K+1) \times 1}$$

instead of using the estimator  $\hat{\theta}^{\text{eff}}$ . Let  $\hat{m}(\theta) \equiv \hat{A}\theta - \hat{\beta}$  denote the sample moment function. By substituting the true (or hypothesized) value  $\theta$  for  $\hat{\theta}^{\text{eff}}$  in  $\hat{\Omega}(\hat{\theta}^{\text{eff}})$  and plugging the

following first-order asymptotic expansion:

$$\sqrt{n}(\hat{\theta}^{\text{eff}} - \theta) = -(\hat{A}'\hat{\Omega}(\theta)^{-1}\hat{A})^{-1}\hat{A}'\hat{\Omega}(\theta)^{-1}\sqrt{n}(\hat{A}\theta - \hat{\beta}) + o_p(1)$$

into equation (2.11), this gives a locally equivalent Lagrangian Multiplier (LM) test statistic:

$$\text{LM}_n(\theta) \equiv n(\hat{A}\theta - \hat{\beta})'\hat{\Omega}(\theta)^{-1/2}P_{\hat{\Omega}(\theta)^{-1/2}\hat{A}(\hat{A}'\hat{\Omega}(\theta)^{-1}\hat{A})^{-1}c}\hat{\Omega}(\theta)^{-1/2}(\hat{A}\theta - \hat{\beta})$$

which does not require  $\hat{\theta}^{\text{eff}}$  to be consistent.

However, the singularity of  $\hat{A}'\hat{\Omega}(\theta)^{-1}\hat{A}$  persists under the projection operator in the LM statistic. Following the insights of Kleibergen (2005), one can orthogonalize columns of  $\hat{A} = (\hat{a}_1, \dots, \hat{a}_{2(M+1)})$  with respect to the variation in the moment condition  $\hat{m}(\theta)$  to obtain a new gradient estimator

$$\hat{D}(\theta) = (\hat{d}_1(\theta), \hat{d}_2(\theta), \dots, \hat{d}_{2(M+1)}(\theta))$$

where

$$\hat{d}_j(\theta) \equiv \hat{a}_j - \hat{\Gamma}_j(\theta)\hat{\Omega}(\theta)^{-1}(\hat{A}\theta - \hat{\beta}) \quad \text{for } j = 1, 2, \dots, 2(M+1).$$

Here  $\hat{\Gamma}_j(\theta)$  is a consistent estimator of the asymptotic covariance between  $\sqrt{n}\hat{m}(\theta)$  and the  $j$ -th column in  $\sqrt{n}(\hat{A} - A)$ , formally defined in equation (2.19).

By this orthogonalization,  $\sqrt{n}(\hat{D}(\theta) - A)$  is asymptotically independent of the moment condition  $\sqrt{n}\hat{m}(\theta)$  in large samples. Replacing  $\hat{A}$  with  $\hat{D}(\theta)$  in the LM statistic then leads to the ‘‘subvector’’ Robust LM (RLM) statistics for inference on  $c'\theta$ :

$$\text{RLM}_n(\theta) = n(\hat{A}\theta - \hat{\beta})'\hat{\Omega}(\theta)^{-1/2}P_{\hat{\Omega}(\theta)^{-1/2}\hat{D}(\theta)(\hat{D}(\theta)'\hat{\Omega}(\theta)^{-1}\hat{D}(\theta))^{-1}c}\hat{\Omega}(\theta)^{-1/2}(\hat{A}\theta - \hat{\beta}). \quad (2.12)$$

If  $\sqrt{n}A$  converges to a fixed matrix  $\mathcal{A}$  as in Kleibergen (2005, p. 1108) (implying  $A$  has a zero limit),  $\sqrt{n}\hat{D}(\theta)$  converges to a Gaussian matrix with mean  $\mathcal{A}$  that is asymptotically independent of the moment vector. Consequently, the projection matrix in the RLM statistic becomes asymptotically independent of the moments on both sides, implying that  $\text{RLM}_n(\theta)$  follows the standard  $\chi_1^2$  limiting distribution for a sequence of DGPs that induces a zero limit of  $\hat{A}$  at a rate  $n^{-1/2}$ .

Compared to the classical RLM statistic for full vector inference in Kleibergen (2005), this subvector RLM statistic (2.12) only attains power for deviations in the linear function  $\mathcal{C}'\theta$  rather than the full vector  $\theta$ . This feature has two important implications. On the one hand, when identification is sufficiently strong, the subvector RLM statistic is locally equivalent to the efficient subvector Wald statistic (2.11) when  $\theta$  approaches its true value; On the other hand, this equivalence may fail since the subvector RLM statistic cannot distinguish alternative values of  $\theta$  from the true value when they yield the same value of  $\mathcal{C}'\theta$ . To overcome this limitation, I. Andrews (2018) introduces a linear combination (LC) statistic that combines the RLM and AR statistics:

$$LC_n(\theta) = RLM_n(\theta) + a \cdot AR_n(\theta) \quad (2.13)$$

where

$$AR_n(\theta) = n(\hat{A}\theta - \hat{\beta})'\hat{\Omega}(\theta)^{-1}(\hat{A}\theta - \hat{\beta})$$

and  $a > 0$  is a tuning parameter on the weights attached to the AR statistic. By incorporating the AR term, the LC statistic diverges to infinity outside the  $n^{-1/2}$  neighborhood of the true parameter  $\theta$  under strong identification. This property guarantees power against any deviations from the true value. Moreover, when  $a$  is sufficiently small and the model is strongly identified, the projection test based on the LC statistic becomes approximately equivalent to the subvector Wald test.

## 2.4.2 Contributions and modifications

In a GMM model, I. Andrews (2018) shows that  $LC_n(\theta)$  converges to a mixture of two independent chi-squared distributions  $(1 + a)\chi_1^2 + a\chi_{2K+1}^2$  in a sequence of DGPs  $\{F_n\}_{n \geq 1}$  satisfying certain high-level conditions. As Andrews notes in his conclusion, the uniform validity of his result under more primitive conditions remains an open question. In this section, I make two key contributions: I show that his high-level conditions are not trivially satisfied in the MTE framework, and I establish the uniform validity of the LC test through a simple modification.

First, consider Assumption 4 of I. Andrews (2018). This assumption requires two convergence conditions by the existence of normalizing sequences: a sequence of full-rank matrices  $\{\Lambda_{1,n}\} \subseteq \mathbb{R}^{2(M+1) \times 2(M+1)}$  and a sequence of nonzero constants  $\{\Lambda_{2,n}\} \subseteq \mathbb{R}$ . Under these sequences, the normalized matrix  $\hat{D}(\theta)\Lambda_{1,n}$  must converge in distribution to a Gaussian matrix of full rank almost surely, and the normalized weight  $\Lambda'_{1,n}c\Lambda_{2,n}$  must converge to a nonzero vector.

The convergence part of this assumption can be verified straightforwardly in the context of Kleibergen (2005) as discussed above, who considers the sequence

$$\sqrt{n}A_{F_n} \rightarrow \mathcal{A} \in \mathbb{R}^{2(K+1) \times 2(M+1)} \quad (2.14)$$

in which case  $\Lambda_{1,n} = \sqrt{n}I$  and  $\Lambda_{2,n} = \frac{1}{\sqrt{n}}$ . More generally, Stock and Wright (2000) and Chaudhuri and Zivot (2011) consider a sequence

$$A_{F_n} = (0_{2(K+1) \times q}, A_{\text{full}}) + A_{\text{sing}, F_n} \quad \text{and} \quad \sqrt{n}A_{\text{sing}, F_n} \rightarrow \mathcal{A} \quad (2.15)$$

in which case  $A_{\text{full}}$  has full column rank, representing those parameters that are strongly identified. We can set  $\Lambda_{1,n} = \text{diag}\{\sqrt{n}I_q, I_{2(M+1)-q}\}$  and  $\Lambda_{2,n} = \frac{1}{\sqrt{n}}$  such that Assumption 4 still holds.

However, the sequences (2.14) and (2.15) are inappropriate for the MTE setup. To see this, consider again the example on a linear MTE model with a binary IV as discussed above:

**Example 2.4.2** (Linear MTE model with a binary IV). *Suppose  $K = M = 1$  and set  $h_1(u) = u - \frac{1}{2}$ . Along a sequence of DGPs  $\{F_n\}$ , let  $p_{F_n}(z_0) = p_{F_n}(z_1) = p \in (0, 1)$  for each  $n \geq 1$ . Since  $\bar{K} = |\{p_{F_n}(z_0), p_{F_n}(z_1)\}| = 1 < M + 1 = 2$ , Assumption 1.7 fails in this example. Note that the matrix  $A_{F_n}$  becomes*

$$A_{F_n} = \begin{pmatrix} 1 & \frac{1}{2}(p-1) & & \\ 1 & \frac{1}{2}(p-1) & & 0_{2 \times 2} \\ & & 0_{2 \times 2} & 1 & \frac{1}{2}p \\ & & & 1 & \frac{1}{2}p \end{pmatrix}$$

*In this case, neither (2.14) nor (2.15) holds here.*

More generally, a singular but nonzero limit of  $A_{F_n}$  would not satisfy conditions (2.14) or (2.15). Similar examples of weakly identified models that fail to meet these conditions are discussed in D. W. Andrews and Guggenberger (2017), where a singular value decomposition (SVD) technique is introduced to establish uniform validity of the classical RLM test for the *full* vector. To achieve the same goal of uniform validity, I generalize their SVD approach to address inference on parameter *functionals* in the MTE framework. My approach proceeds with a SVD of  $A_{F_n}$ :

$$A_{F_n} = C_{F_n} \underbrace{\left[ \begin{array}{cccc} \tau_{1,F_n} & 0 & \dots & 0 \\ 0 & \tau_{2,F_n} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tau_{2(M+1),F_n} \\ \underbrace{0_{2(K-M) \times 2(M+1)}}_{\Pi_{F_n}} \end{array} \right]}_{\Pi_{F_n}} B'_{F_n},$$

where  $C_{F_n} \in \mathbb{R}^{2(K+1) \times 2(K+1)}$  and  $B_{F_n} \in \mathbb{R}^{2(M+1) \times 2(M+1)}$  are orthogonal matrices, and  $\infty \geq \tau_{1,F_n} \geq \tau_{2,F_n} \geq \dots \geq \tau_{2(M+1),F_n} \geq 0$  are singular values of  $A_{F_n}$  in a descending order. I show that the following normalizing matrices establish convergence of  $\hat{D}(\theta)\Lambda_{1,n}$  and  $\Lambda'_{1,n}c\Lambda_{2,n}$ :

$$\Lambda_{1,n} = B_{F_n} \text{diag}\{(\tau_{1,F_n})^{-1}, \dots, (\tau_{q,F_n})^{-1}, \sqrt{n}, \dots, \sqrt{n}\}$$

$$\Lambda_{2,n} = \|\Lambda'_{1,n}c\|^{-1},$$

where  $q$  is the number of scaled singular values  $\{\sqrt{n}\tau_{j,F_n}\}_{j=1}^{2(M+1)}$  that diverge to infinity.

Despite the convergence of  $\hat{D}\Lambda_{1,n}$  under the constructed normalizing matrices, its asymptotic limit may be rank deficient. This issue prevents the use of the efficiently weighted projection matrix in the subvector RLM statistic. (A similar issue was also noticed by D. W. Andrews and Guggenberger (2017), who impose additional parameter space restrictions to avoid this problem). To address this problem, I modify the matrix  $\hat{D}$  by adding a small noise of size  $n^{-1/2}$ :

$$\tilde{D}(\theta) = \hat{D}(\theta) + \kappa n^{-1/2}\xi, \quad (2.16)$$

where  $\kappa > 0$  is a tuning parameter and  $\xi \in \mathbb{R}^{2(K+1) \times 2(M+1)}$  is a matrix of i.i.d. standard normal random variables independent of data. This perturbation ensures that  $\tilde{D}(\theta)\Lambda_{1,n}$  achieves full rank almost surely while having asymptotically negligible effects on the test statistic under strong identification. This perturbation technique draws inspiration from the AR/LM test<sup>5</sup>(D. W. Andrews, 2017). While AR/LM test focuses specifically on inference for a subset of parameters, my paper addresses a different problem of inference on a linear functional of parameters.

Define the new matrix under projection as follows:

$$\hat{Q}(\theta) = \hat{\Omega}(\theta)^{-1/2} \tilde{D}(\theta) (\tilde{D}(\theta)' \hat{\Omega}(\theta)^{-1} \tilde{D}(\theta))^{-1} c,$$

and the corresponding modified RLM (MRLM) statistic becomes

$$\text{MRLM}_n(\theta) = n(\hat{A}\theta - \hat{\beta})' \hat{\Omega}(\theta)^{-1/2} P_{\hat{Q}(\theta)} \hat{\Omega}(\theta)^{-1/2} (\hat{A}\theta - \hat{\beta}).$$

Then the modified LC (MLC) statistic is given by

$$\text{MLC}_n(\theta) = \text{MRLM}_n(\theta) + a \cdot \text{AR}_n(\theta). \quad (2.17)$$

In section 2.4.4, I establish the uniform validity of using (2.17) for conducting inference on the linear function  $c'\theta$  even if Assumption 1.7 might fail or be close to failing.

### 2.4.3 Implementation of the MLC test

In this section, I describe the implementation of the projection test based on the MLC statistic as below:

1. Construct the key quantities for the test statistics
  - (a) Construct the estimators of  $\hat{p}$  and  $\hat{\beta}$ , as well as their asymptotic variances estimators  $\hat{\Sigma}_p$  and  $\hat{\Sigma}_\beta$ , as in section 2.2.4. Plugging in the estimator  $\hat{p}$  into the matrix  $A$  then obtains the estimator  $\hat{A}$ .

---

<sup>5</sup> In addition to the modification in (2.16), D. W. Andrews (2017, eq. (7.11)) introduces two tuning parameters,  $(K_L^*, K_U^*)$ , to categorize identification strength. However, this categorization is not required for our inference procedure, and there is no clear guidance for choosing these parameters.

(b) Define the asymptotic variance estimator of the moment function

$$\hat{\Omega}(\theta) = H(\hat{p}, \theta) \hat{\Sigma}_p H(\hat{p}, \theta)' + \hat{\Sigma}_\beta \quad (2.18)$$

where

$$H(p, \theta) = \begin{bmatrix} \text{diag} \left\{ \sum_{m=0}^M \theta_{1m} \lambda'_{1m}(p(z_\ell)) : \ell = 0, 1, \dots, K \right\} \\ \text{diag} \left\{ \sum_{m=0}^M \theta_{0m} \lambda'_{0m}(p(z_\ell)) : \ell = 0, 1, \dots, K \right\} \end{bmatrix}.$$

(c) For each  $m = 0, 1, \dots, M$ , let

$$L_m(\hat{p}) = \text{diag}\{\lambda'_{1m}(\hat{p}(z_0)), \dots, \lambda'_{1m}(\hat{p}(z_K))\}$$

and

$$R_m(\hat{p}) = \text{diag}\{\lambda'_{0m}(\hat{p}(z_0)), \dots, \lambda'_{0m}(\hat{p}(z_K))\}.$$

For each  $j = 1, \dots, 2(M+1)$ , define a column vector  $\hat{d}_j(\theta)$  and the covariance estimator  $\hat{\Gamma}_j(\theta)$  as below

$$\begin{aligned} \hat{d}_j(\theta) &\equiv \hat{a}_j - \hat{\Gamma}_j(\theta) \hat{\Omega}(\theta)^{-1} (\hat{A}\theta - \hat{\beta}) \\ \hat{\Gamma}_j(\theta) &= M_j(\hat{p}) \hat{\Sigma}_p H(\hat{p}, \theta)' \end{aligned} \quad (2.19)$$

where

$$M_j(\hat{p}) = \begin{cases} \begin{bmatrix} L_{j-1}(\hat{p}) \\ 0_{(K+1) \times (K+1)} \end{bmatrix} & \text{if } j \leq M+1 \\ \begin{bmatrix} 0_{(K+1) \times (K+1)} \\ R_{j-M-2}(\hat{p}) \end{bmatrix} & \text{if } j > M+1 \end{cases}. \quad (2.20)$$

Let  $\hat{D}(\theta) = (\hat{d}_1(\theta), \dots, \hat{d}_{2(M+1)}(\theta))$  and  $\tilde{D}(\theta) = \hat{D}(\theta) + \kappa n^{-1/2} \xi$ , where  $\xi \in \mathbb{R}^{2(K+1) \times 2(M+1)}$  is a matrix of standard normal random variables independent of data.  $\kappa$  is a positive tuning parameter chosen as  $10^{-6}$  following D. W. Andrews (2017).

2. Given the estimators  $\hat{A}$ ,  $\hat{\beta}$ ,  $\hat{\Omega}(\theta)$ , and  $\tilde{D}(\theta)$  constructed from step 1, compute the modified LC statistics as in equation (2.17) for  $\theta \in \Theta$ . For the empirical application and simulation studies, I set the AR weight coefficient as 0.05, which is close to the

suggested weight  $a(\gamma)$  introduced by I. Andrews (2018, p. 343) by setting  $\alpha = 5\%$ ,  $\gamma = 10\%$ , and  $K = 15$  as in section 2.7.

3. Let  $\alpha \in (0, 1)$  be the significance level. The modified LC test rejects the null hypothesis  $H_0 : c'\theta = \lambda$  if the profiled test statistic over the linear manifold  $c'\theta = \lambda$  is larger than their respective critical values, i.e.,

$$\hat{\phi}_{\text{MLC}}(\lambda) = \mathbb{1} \left[ \inf_{c'\theta=\lambda} \text{MLC}_n(\theta) > q_{(1+a)\chi_1^2+a\chi_{2K+1}^2}(1-\alpha) \right]$$

where  $q_{(1+a)\chi_1^2+a\chi_{2K+1}^2}(1-\alpha)$  denotes the  $(1-\alpha)$ -quantile of the mixture chi-square distribution  $(1+a)\chi_1^2 + a\chi_{2K+1}^2$  with  $\chi_1^2 \perp \chi_{2K+1}^2$ . Additionally, define  $\hat{\phi}_{\text{MLC}}(\lambda) = 1$  for  $\lambda \notin \{c'\theta : \theta \in \Theta\}$ .

A robust confidence set can be obtained by inverting the test, that is,

$$\mathcal{C}_{\text{MLC}} = \{\lambda \in \mathbb{R} : \hat{\phi}_{\text{MLC}}(\lambda) = 0\}.$$

#### 2.4.4 Uniform validity

The following theorem establishes the uniform validity of the MLC test based on the profiling the MLC statistic.

**Theorem 2.4.1.** *Let Assumption 2 hold, and suppose that the weight  $c$  is a nonzero fixed vector. Then we have*

$$\limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{E}_F[\hat{\phi}_{\text{MLC}}(\lambda)] \leq \alpha.$$

By this theorem, the MLC test has an asymptotic size less than or equal to the nominal level  $\alpha \in (0, 1)$  for the parameter space  $\mathcal{P}_0$ . In other words, the MLC test is uniformly valid regardless of the model's identification status. In Appendix A.2, I further demonstrate that the MLC test is consistent for distant (or fixed) alternatives and has comparable power to the efficient Wald test under strong identification when the AR weight  $a$  is sufficiently small. These findings highlight the usefulness of the MLC test for researchers seeking robust inference against weak identification.

## 2.5 Incorporating Covariates

In this section, I assume a set of covariates  $W$  is included in the MTE model as follows:

$$Y_d = \mu_d(W) + V_d$$

$$D = \mathbb{1}[U \leq \nu(W, Z)],$$

where  $\mu_d(W) \equiv \mathbb{E}[Y_d \mid W]$  denotes the conditional expectation of potential outcomes. Similar to Theorem 2.2.1, we can identify the MTE function  $\mathbb{E}[Y_1 - Y_0 \mid F_{U|W}(U|W) = u, W = w]$  under the following assumptions.

**Assumption 5** (MTE model with covariates).

1.  $Z \perp\!\!\!\perp U \mid W$ .
2.  $\mathbb{E}[Y_d \mid Z, W, U] = \mathbb{E}[Y_d \mid W, U]$  and  $\mathbb{E}|Y_d| < \infty$  for  $d \in \{0, 1\}$ .
3.  $U \mid W = w$  is continuously distributed for all  $w \in \text{supp}(W)$ .
4.  $0 < \mathbb{P}(D = 1 \mid W = w, Z = z) < 1$  for all  $(w, z) \in \text{supp}(W, Z)$ .
5.  $\mathbb{E}[Y_d \mid F_{U|W}(U|W) = u, W = w] = \mu_d(w) + \sum_{m=1}^M \rho_{dm}(w)h_m(u)$  for some known continuous functions  $\{h_m(\cdot)\}_{m=1}^M$  for each  $d = 0, 1$  and  $u \in (0, 1)$ , where  $F_{U|W}(\cdot|w)$  denotes the distribution function of  $U$  conditional on  $W = w$ .
6.  $\{\lambda_{1m}(\cdot)\}_{m=0}^M$  and  $\{\lambda_{0m}(\cdot)\}_{m=0}^M$  are unisolvent on  $(0, 1)$ , where  $\lambda_{00}(\cdot) = \lambda_{10}(\cdot) \equiv 1$ , and

$$\lambda_{1m}(p) \equiv \frac{1}{p} \int_0^p h_m(u) du \quad \text{and} \quad \lambda_{0m}(p) \equiv \frac{1}{1-p} \int_p^1 h_m(u) du \quad \text{for } m = 1, \dots, M.$$

7.  $|\{\mathbb{P}(D = 1 \mid Z = z, W = w) : z \in \text{supp}(Z \mid W = w)\}| \geq M + 1$  for all  $w \in \text{supp}(W)$ .

Under the above assumption, we can normalize the distribution of  $U$  conditional on the random covariate  $W$  to be uniformly distributed over the unit interval (i.e.,  $F_{U|W}(u|W) = u$ ) and thus  $\nu(w, z) = p(w, z) \equiv \mathbb{P}(D = 1 \mid W = w, Z = z)$ . The structural quantities

$$\theta(w) = (\mu_1(w), \{\rho_{1m}(w)\}_{m=1}^M, \mu_0(w), \{\rho_{0m}(w)\}_{m=1}^M)$$

can be identified for all  $w \in \text{supp}(W)$  by the following separate regressions:

$$\begin{aligned}
\mathbb{E}[Y \mid D = 1, Z = z, W = w] &= \mu_1(w) + \sum_{m=1}^M \rho_{1m}(w) \int_0^{p(w,z)} \frac{h_m(u)}{p(w,z)} du \\
&= \mu_1(w) + \sum_{m=1}^M \rho_{1m}(w) \lambda_{1m}(p(w,z)) \\
\mathbb{E}[Y \mid D = 0, Z = z, W = w] &= \mu_0(w) + \sum_{m=1}^M \rho_{0m}(w) \int_{p(w,z)}^1 \frac{h_m(u)}{1-p(w,z)} du \\
&= \mu_0(w) + \sum_{m=1}^M \rho_{0m}(w) \lambda_{0m}(p(w,z)).
\end{aligned} \tag{2.21}$$

For some  $w \in \text{supp}(W)$ , the variation of  $\{p(w, z) : z \in \text{supp}(Z \mid W = w)\}$  can be weak in practice, especially when conditioning on certain groups of units who have very high (or low) probability of being treated. To address this problem, researchers commonly impose the additive separability condition as follows.

**Assumption 6** (Additive separability). *The MTR function is linear and additively separable in covariates and selection unobservable. That is, for  $d = 0, 1$ ,*

$$\mathbb{E}[Y_d \mid W, U] = \mu_d + W' \tau_d + \mathbb{E}[V_d \mid U].$$

This assumption eliminates heterogeneous effects of covariates  $W$  interacted with selection unobservable  $U$  on potential outcomes, i.e.,  $\rho_{dm}(w)$  does not vary with  $w$ . Up to a linear term  $W' \tau_d$ , the covariate  $W$  is mean independent of  $Y_d$  conditional on  $U$ . This enables point identification of MTEs on the unconditional support of propensity scores  $p(W, Z)$  by using variation from covariates in addition to exogenous variation from discrete instruments (Carneiro et al., 2011, Section I.B). Next, I show that the failure of Assumption 6 can lead to a biased estimator of treatment effects unless  $W$  is uncorrelated with the treatment and propensity score. Given this result, I consider Assumption 6 to be a strong assumption and implement inference in the empirical analysis by conditioning on covariates rather than relying on additive separability.

### 2.5.1 Bias from additive separability

I show the bias of treatment effect estimators for a specific class of DGP as follows:

$$\mathbb{E}[Y_d | W = w, U = u] = \mu_d + w'\tau_d + \rho_d(w)h(u), \quad (2.22)$$

where  $W \in \mathbb{R}^L$  is a vector of covariates,  $\rho_d(w) = \rho_d + w'\eta_d$  for  $d = 0, 1$ , and  $h(\cdot)$  is strictly increasing and integrates to zero over  $[0, 1]$ . This model can be regarded as the MTE model satisfying Assumption 5 with  $M = 1$  and  $\mu_d(w)$  being linear in  $w$ .<sup>6</sup> Note that the effects of  $W$  can vary across different values of selection unobservable  $U$ , which violates the additive separability Assumption 6. To obtain correct estimates on the model parameters, one should implement separate regressions as follows:

$$\begin{aligned} \mathbb{E}[Y | D = 1, W = w, P = p] &= \mu_1 + w'\tau_1 + \rho_1\lambda_1(p) + [w\lambda_1(p)]'\eta_1 \\ \mathbb{E}[Y | D = 0, W = w, P = p] &= \mu_0 + w'\tau_0 + \rho_0\lambda_0(p) + [w\lambda_0(p)]'\eta_0 \end{aligned} \quad (2.23)$$

where  $P = \mathbb{P}(D = 1 | Z, W)$  is the propensity score, and

$$\lambda_1(p) = \frac{1}{p} \int_0^p h(u)du \quad \text{and} \quad \lambda_0(p) = \frac{1}{1-p} \int_p^1 h(u)du.$$

However, suppose a researcher mistakenly imposes additive separability. This implies  $\rho_d(w)$  is considered as a constant in the model with  $\eta_0 = \eta_1 = 0_{L \times 1}$ . As a result, short regressions with omitted interaction terms  $W\lambda_d(P)$  will be implemented:

$$\begin{aligned} Y &= \tilde{\mu}_1 + W'\tilde{\tau}_1 + \tilde{\rho}_1\lambda_1(P) + Y^{\perp W, \lambda_1(P)|D=1} \quad \text{conditional on } D = 1 \\ Y &= \tilde{\mu}_0 + W'\tilde{\tau}_0 + \tilde{\rho}_0\lambda_0(P) + Y^{\perp W, \lambda_0(P)|D=0} \quad \text{conditional on } D = 0 \end{aligned} \quad (2.24)$$

where  $Y^{\perp X|D=d}$  denotes the OLS residual of  $Y$  from regressing  $Y$  on  $(1, X)$  conditional on subsamples  $D = d$ :

$$Y^{\perp X|D=d} \equiv Y - \tilde{X}'\mathbb{E}[\tilde{X}\tilde{X}' | D = d]^{-1}\mathbb{E}[\tilde{X}Y | D = d].$$

---

<sup>6</sup> Specification of this form under additive separability also appears in Kline and Walters (2019). The difference exists on the unobserved heterogeneity part  $\rho_d(w)h(u)$

where  $\tilde{X} = (1, X)'$ .

The next lemma compares the estimands from the correct specification (2.23) with the ones from the misspecified model (2.24). Note that the bias derived under the specific model class (2.22) indicates that the worst-case bias would worsen when extending to a broader model class where additive separability does not hold.

**Lemma 2.5.1.** *Under the DGP (2.22), the bias  $\tilde{\rho}_d - \rho_d$  equals*

$$\frac{\text{cov}((W'\lambda_d(P))^{\perp W|D=d}, \lambda_d(P)^{\perp W|D=d} \mid D = d)}{\text{var}(\lambda_d(P)^{\perp W|D=d} \mid D = d)} \eta_d \quad (2.25)$$

and the bias  $\tilde{\tau}_d - \tau_d$  equals

$$\mathbb{E}[(W^{\perp \lambda_d(P)|D=d})(W^{\perp \lambda_1(P)|D=d})' \mid D = d]^{-1} \mathbb{E}[(W^{\perp \lambda_d(P)|D=d})(W'\lambda_d(P)) \mid D = d] \eta_d \quad (2.26)$$

The bias formulas are direct consequences of Frisch–Waugh–Lovell (FWL) theorem. It shows that the degree of treatment effects heterogeneity of the covariate  $W$ , measured by  $\eta_d$ , has a nontrivial impact on the bias of estimates.

The bias formulas (2.25) and (2.26) simplify further under the following additional assumptions:

**Assumption 7.** *Under the DGP (2.22), the following conditions hold:*

1.  *$W$  is uncorrelated with the control functions of propensity score conditional on treatment status: For  $d = 0, 1$ ,*

$$\text{cov}(W, \lambda_d(P) \mid D = d) = 0_{L \times 1}$$

$$\text{cov}(WW', \lambda_d(P) \mid D = d) = 0_{L \times L}$$

$$\text{cov}(W, \lambda_d(P)^2 \mid D = d) = 0_{L \times 1}.$$

2.  *$W$  is uncorrelated with treatment:  $\mathbb{E}[W \mid D = 1] = \mathbb{E}[W \mid D = 0]$ .*

Assumption 7.1 posits that the covariate  $W$  and the control function  $\lambda_d(P)$  are uncorrelated up to their second moment, given the treatment or control groups. This assumption

is notably strong, suggesting that the covariate  $W$  cannot cause significant variation on the propensity scores within both groups. However, the bias of treatment effect estimands persists even under such stringent assumptions.

Only if researchers are willing to assume that  $W$  is completely “irrelevant” under an additional Assumption 7.2, the misspecified model (2.24) would yield correct estimates of average treatment effects and the slope of MTEs as the ones obtained from the true model (2.23).

I focus on the bias for ATE, conditional ATE, and slope of MTE curve<sup>7</sup>, which are defined as

$$\begin{aligned} \text{ATE} &= \mathbb{E}[Y_1 - Y_0] = \mu_1 - \mu_0 + \mathbb{E}[W'(\tau_1 - \tau_0)] \\ \text{CATE} &= \mathbb{E}[Y_1 - Y_0 \mid W = w] = \mu_1 - \mu_0 + w'(\tau_1 - \tau_0) \\ \text{Slope} &= \mathbb{E}[\rho_1(W) - \rho_0(W)] = \rho_1 - \rho_0 + \mathbb{E}[W'(\eta_1 - \eta_0)]. \end{aligned}$$

Under misspecified additive separability, the researcher may estimate these effects by

$$\begin{aligned} \widetilde{\text{ATE}} &= \tilde{\mu}_1 - \tilde{\mu}_0 + \mathbb{E}[W'(\tilde{\tau}_1 - \tilde{\tau}_0)] \\ \widetilde{\text{CATE}} &= \tilde{\mu}_1 - \tilde{\mu}_0 + w'(\tilde{\tau}_1 - \tilde{\tau}_0) \\ \widetilde{\text{Slope}} &= \tilde{\rho}_1 - \tilde{\rho}_0. \end{aligned}$$

The next theorem shows the difference between the causal parameters and their corresponding estimands under Assumption 7.

---

<sup>7</sup> The ATE and the slope of the MTE are of interest because the shape of the unconditional MTE,  $\mathbb{E}[Y_1 - Y_0 \mid U = u]$ , is uniquely determined by these quantities. This is also the commonly reported MTE curve evaluated at the mean value of covariates.

**Theorem 2.5.1.** *Under the model (2.22),*

1. *Suppose Assumption 7.1 holds, then*

$$\begin{aligned}\widetilde{ATE} - ATE &= (\mathbb{E}[W] - \mathbb{E}[W \mid D = 1])' \eta_1 \times \mathbb{E}[\lambda_1(P) \mid D = 1] \\ &\quad - (\mathbb{E}[W] - \mathbb{E}[W \mid D = 0])' \eta_0 \times \mathbb{E}[\lambda_0(P) \mid D = 0] \\ \widetilde{CATE} - CATE &= (w - \mathbb{E}[W \mid D = 1])' \eta_1 \times \mathbb{E}[\lambda_1(P) \mid D = 1] \\ &\quad - (w - \mathbb{E}[W \mid D = 0])' \eta_0 \times \mathbb{E}[\lambda_0(P) \mid D = 0] \\ \widetilde{Slope} - Slope &= (\mathbb{E}[W \mid D = 1] - \mathbb{E}[W \mid D = 0])' (\mathbb{P}(D = 0) \eta_1 + \mathbb{P}(D = 1) \eta_0).\end{aligned}$$

2. *Suppose Assumption 7.1 and 7.2 hold, then  $\widetilde{ATE} = ATE$  and  $\widetilde{Slope} = Slope$ .*

**Remark 2.5.1.** *While the estimation of the ATE may remain unbiased under both conditions outlined in Assumption 7, it is important to note that the bias on CATE does not necessarily disappear under such assumption. Consequently, researchers should be cautious about interpreting CATE estimates in short regressions (2.24) even if they have justified the validity of Assumption 7.*

From this theorem, it becomes evident that the bias on ATE and the slope of MTE are driven by two main factors: the magnitude of  $\eta_d$ , which represents the heterogeneity effects of  $W$ , and the extent to which the covariate  $W$  is unbalanced between the treatment and control groups. Therefore, the bias on those estimands can be quite significant if we omit heterogeneous effects of covariates that vary with unobserved heterogeneity  $U$ . In Appendix A.6, I provide a numerical example illustrating that such bias can even alter the sign of the ATE estimand when additive separability is mistakenly imposed.

## 2.5.2 Šidák-Bonferroni's correction

In this section, I implement the proposed identification-robust inference procedure conditional on a set of discrete covariates  $W$ . Assumption 2 is extended to incorporate additional covariates:

**Assumption 8.** *The random vectors  $(Y_i, D_i, Z_i, W_i)$  for  $i = 1, \dots, n$  are i.i.d. with distribution  $F$ , where  $W_i$  has finite support.*

Next I introduce a set of regularity conditions to be imposed on the distribution  $F$ . Let  $\theta = \{\theta(w) : w \in \text{supp}(W)\}$ . For some  $\delta, \zeta > 0$  and  $\epsilon \in (0, 1/2)$ , define the parameter space  $\mathcal{P}$  as the set of pairs  $(\theta, F)$  satisfying the following properties:

1. Equation (2.21) is satisfied with  $K \geq M$ , where  $\theta(w) \in \text{int}(\Theta) \subseteq \mathbb{R}^{2(M+1)}$  for some compact set  $\Theta$ , for all  $w \in \text{supp}(W)$ ,
2.  $\sup_{d=0,1} \sup_{(z,w) \in \text{supp}(Z,W)} \mathbb{E}_F[|Y|^{2+\delta} \mid D = d, Z = z, W = w] \leq \zeta$ ,
3.  $\epsilon \leq \inf_{(z,w) \in \text{supp}(Z,W)} \mathbb{P}_F(D = 1 \mid W = w, Z = z) \leq \sup_{(z,w) \in \text{supp}(Z,W)} \mathbb{P}_F(D = 1 \mid W = w, Z = z) \leq 1 - \epsilon$ ,
4.  $\epsilon \leq \inf_{(z,w) \in \text{supp}(Z,W)} \mathbb{P}_F(Z = z \mid W = w) \leq \sup_{(z,w) \in \text{supp}(Z,W)} \mathbb{P}_F(Z = z \mid W = w) \leq 1 - \epsilon$ ,
5.  $\epsilon \leq \inf_{w \in \text{supp}(W)} \mathbb{P}_F(W = w) \leq \sup_{w \in \text{supp}(W)} \mathbb{P}_F(W = w) \leq 1 - \epsilon$ ,
6.  $\epsilon \leq \inf_{d=0,1} \inf_{(z,w) \in \text{supp}(Z,W)} \text{var}_F(Y \mid D = d, Z = z, W = w)$ .

In particular, conditions 3, 4, and 5 together imply that

$$\text{supp}(D, Z, W) = \{0, 1\} \times \text{supp}(Z) \times \text{supp}(W)$$

and strong overlap holds.

For a fixed  $w \in \text{supp}(W)$ , let  $\mathcal{P}(w)$  be the projection set of  $\mathcal{P}$  onto  $(\theta(w), F_{Y,D,Z|W=w})$ . That is,  $(\theta(w), F_{Y,D,Z|W=w})$  belongs to  $\mathcal{P}(w)$  if there exists a DGP  $(\theta, F) \in \mathcal{P}$  that generates  $(\theta(w), F_{Y,D,Z|W=w})$ . Note that elements in  $\mathcal{P}(w)$  still satisfy conditions 1-6 but with a fixed  $w \in \text{supp}(W)$ . Define  $\mathcal{P}_0(w)$  as the space of the conditional causal effects  $\lambda(w) = c(w)' \theta(w)$ <sup>8</sup>:

$$\mathcal{P}_0(w) = \{(\lambda(w), F_{Y,D,Z|W=w}) : \lambda(w) = c(w)' \theta(w), (\theta(w), F_{Y,D,Z|W=w}) \in \mathcal{P}(w)\}.$$

Note that the weight  $c(w)$  can depend on the underlying distribution  $F_{Y,D,Z|W=w}$ , but I omit this dependence for the simplicity of notations.

---

<sup>8</sup> The weight  $c(w)$  depends on covariates  $w$  if it contains unknown propensity scores, e.g., ATT, LATE, and PRTE.

The space for the aggregated effects  $\lambda = \sum_{w \in \text{supp}(W)} q(w)\lambda(w)$  is denoted by

$$\mathcal{P}_0 = \left\{ (\lambda, F) : \lambda = \sum_{w \in \text{supp}(W)} q_F(w)c(w)'\theta(w), \quad (\theta, F) \in \mathcal{P} \right\}$$

where  $q_F(w)$  represents the marginal or conditional distribution of covariates, for example:

1. The marginal probability of covariates  $\mathbb{P}(W = w)$ , for ATE.
2. The weighted marginal probability:  $\mathbb{P}(W = w \mid D = d)$ , for ATT and ATU.
3. The weighted marginal probability:  $\mathbb{P}(W = w \mid p(W, z_0) < U < p(W, z_k))$ , for LATE.

The methods described in sections 2.3 and 2.4 have led to a confidence set  $\mathcal{C}(w)$  for the conditional effect  $\lambda(w)$  with uniformly valid coverage requirement as below:

**Definition 2.5.1.** For  $\omega \in \text{supp}(W)$ , a confidence set  $\mathcal{C}(w)$  for the conditional causal effects  $\lambda(w) = c(w)'\theta(w)$  is said to be uniformly valid with asymptotic level  $1 - \alpha$  if it depends on the samples  $\{(Y_i, D_i, Z_i) : W_i = w\}$  and satisfies

$$\liminf_{n \rightarrow \infty} \inf_{(\lambda(w), F_{Y,D,Z|W=w}) \in \mathcal{P}_0(w)} \mathbb{P}_F(\lambda(w) \in \mathcal{C}(w)) \geq 1 - \alpha.$$

The following lemma then shows that a valid confidence set for the unconditional effect  $\lambda$  can be obtained by combining the confidence sets  $\mathcal{C}(w)$  across the support of covariate  $W$  using an adjusted critical value.

**Lemma 2.5.2.** Let Assumption 8 hold. Suppose for each  $w \in \text{supp}(W)$ , there exists a uniformly valid confidence set  $\mathcal{C}(w)$  for  $\lambda(w)$  with asymptotic level  $(1 - \alpha)^{1/|\text{supp}(W)|}$ . Define a confidence set  $\mathcal{C}$  by taking a weighted average:

$$\mathcal{C} = \left\{ \lambda = \sum_{w \in \text{supp}(W)} q_F(w)\lambda(w) : \lambda(w) \in \mathcal{C}(w) \right\}.$$

Then we have

$$\liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(\lambda \in \mathcal{C}) \geq 1 - \alpha.$$

*Proof.* For any fixed  $(\lambda, F) \in \mathcal{P}_0$ , we have

$$\begin{aligned} \mathbb{P}_F(\lambda \in \mathcal{C}) &\geq \mathbb{P}_F\left(\bigcap_{w \in \text{supp}(W)} \{\lambda(w) \in \mathcal{C}(w)\}\right) \\ &= \prod_{w \in \text{supp}(W)} \mathbb{P}_F(\lambda(w) \in \mathcal{C}(w)) \end{aligned}$$

The second line uses the fact that  $\{\lambda(w) \in \mathcal{C}(w)\}$  is independent of  $\{\lambda(w') \in \mathcal{C}(w')\}$  if  $w \neq w'$  since confidence sets are estimated from independent samples following Definition 2.5.1.

Taking infimum on both sides obtains

$$\begin{aligned} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(\lambda \in \mathcal{C}) &= \inf_{(\lambda, F) \in \mathcal{P}_0} \prod_{w \in \text{supp}(W)} \mathbb{P}_F(\lambda(w) \in \mathcal{C}(w)) \\ &\geq \prod_{w \in \text{supp}(W)} \inf_{(\lambda(w), F_{Y, D, Z|W=w}) \in \mathcal{P}_0(w)} \mathbb{P}_F(\lambda(w) \in \mathcal{C}(w)) \end{aligned}$$

The second line holds since every element  $(\lambda, F) \in \mathcal{P}_0$  corresponds to a valid element of  $(\lambda(w), F_{Y, D, Z|W=w})$  contained in  $\mathcal{P}_0(w)$  for each  $w \in \text{supp}(W)$ . Let sample size  $n$  go to infinity and then it follows that

$$\liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(\lambda \in \mathcal{C}) \geq \prod_{w \in \text{supp}(W)} \liminf_{n \rightarrow \infty} \inf_{(\lambda(w), F_{Y, D, Z|W=w}) \in \mathcal{P}_0(w)} \mathbb{P}_F(\lambda(w) \in \mathcal{C}(w)) \geq 1 - \alpha.$$

So the desired result is proved.  $\square$

Consider a unknown  $q_F(w)$  that needs to be estimated. Denote  $\mathcal{C}_q$  the valid confidence set of  $q_F = \{q_F(w) : w \in \text{supp}(W)\}$  with asymptotic level  $1 - \alpha_1$  ( $0 < \alpha_1 < \alpha$ ), which satisfies

$$\liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(q_F \in \mathcal{C}_q) \geq 1 - \alpha_1.$$

One can construct such confidence set by inverting the Wald test based on the sample analog estimator of  $q_F$ .

Let  $\mathcal{C}(w)$  be a uniformly valid confidence set with asymptotic level  $(1 - \alpha_2)^{1/|\text{supp}(W)|}$  for  $\lambda(w)$  as defined in Definition 2.5.1. Then we can construct the confidence set  $\mathcal{C}$  for the

unconditional causal effects as below:

$$\mathcal{C} = \left\{ \lambda = \sum_{w \in \text{supp}(W)} q(w)\lambda(w) : \lambda(w) \in \mathcal{C}(w), q \in \mathcal{C}_q \right\}.$$

**Lemma 2.5.3.** *Let Assumption 8 hold. Suppose for each  $w \in \text{supp}(W)$ , there exists a uniformly valid confidence set  $\mathcal{C}(w)$  for  $\lambda(w)$  with asymptotic level  $(1 - \alpha_2)^{1/|\text{supp}(W)|}$ , and there exists a uniformly valid confidence set for  $q_F$  with asymptotic level  $1 - \alpha_1$ , where  $\alpha_1 = \alpha - \alpha_2$ . Then we have*

$$\liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(\lambda \in \mathcal{C}) \geq 1 - \alpha.$$

*Proof.* For a fixed pair  $(\lambda, F) \in \mathcal{P}_0$ , we have

$$\begin{aligned} \mathbb{P}_F(\lambda \in \mathcal{C}) &= \mathbb{P}_F \left( \sum_{w \in \text{supp}(W)} q(w)\lambda(w) \in \mathcal{C} \right) \\ &\geq \mathbb{P}_F \left( q \in \hat{\mathcal{C}}_q, \lambda(w) \in \mathcal{C}(w) \text{ for each } w \in \text{supp}(W) \right) \\ &\geq \mathbb{P}_F \left( q \in \hat{\mathcal{C}}_q \right) + \prod_{w \in \text{supp}(W)} \mathbb{P}_F(\lambda(w) \in \mathcal{C}(w)) - 1 \end{aligned}$$

Following the same arguments in Lemma 2.5.2, we have

$$\begin{aligned} \liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(\lambda \in \mathcal{C}) &\geq \liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F(q \in \mathcal{C}_q) \\ &\quad + \prod_{w \in \text{supp}(W)} \liminf_{n \rightarrow \infty} \inf_{(\lambda(w), F_{Y,D,Z|W=w}) \in \mathcal{P}_0(w)} \mathbb{P}_F(\lambda(w) \in \mathcal{C}(w)) - 1 \\ &= 1 - \alpha_1 - \alpha_2 \\ &= 1 - \alpha. \end{aligned}$$

So the desired conclusion is proved.  $\square$

In practice, we usually can achieve precise estimation of  $q(z)$  if the sample size conditional on covariates is considerably large. In this case, the standard errors from estimating  $q(z)$  are often small, and therefore the corresponding impact can usually be ignored.

### 2.5.3 Discussion of inference under additive separability

In this section, I briefly describe how to implement the proposed robust inference procedures if researchers still want to maintain additive separability. Following the existing literature discussed in Appendix A.7.2, additional functional restrictions should be imposed on both stages, such as the additive separability (Assumption 6) and a distributional assumption on selection unobservable  $U$ . To fix ideas, consider a model with covariates  $W \in \mathbb{R}^L$  and a discrete instrument  $Z$  below:

$$\begin{aligned} Y_d &= \mu_d + W'\tau_d + V_d \\ D &= \mathbb{1}[U \leq W'\pi + \delta(Z)] \\ U &\perp\!\!\!\perp (W, Z) \quad \text{and} \quad U \sim F_U(\cdot) \end{aligned}$$

where  $F_U$  denotes the CDF for a known distribution (e.g., logit or normal specifications). Additionally, assume the additive separable structure holds:

$$\mathbb{E}[V_d \mid W = w, F_U(U) = u] = \mathbb{E}[V_d \mid F_U(U) = u]$$

for which I maintain the parametric specification of control functions:

$$\mathbb{E}[V_d \mid F_U(U) = u] = \sum_{m=1}^M \rho_{dm} h_m(u).$$

The separate regression approach gives the following observable implications for identification of the structural parameters:

$$\begin{aligned} \mathbb{E}[Y \mid D = 1, W = w, Z = z] &= \mu_1 + w'\tau_1 + \sum_{m=1}^M \rho_{1m} \lambda_{1m}(w'\pi + \delta(z)) \\ &\equiv \mu_1 + w'\tau_1 + \lambda_1(w'\pi + \delta(z))' \rho_1 \\ \mathbb{E}[Y \mid D = 0, W = w, Z = z] &= \mu_0 + w'\tau_0 + \sum_{m=1}^M \rho_{0m} \lambda_{0m}(w'\pi + \delta(z)) \\ &\equiv \mu_0 + w'\tau_0 + \lambda_0(w'\pi + \delta(z))' \rho_0 \end{aligned} \tag{2.27}$$

where  $\rho_d = (\rho_{d1}, \dots, \rho_{dM})'$  and  $\lambda_d(t) = (\lambda_{d1}(t), \dots, \lambda_{dM}(t))'$  for  $t \in \mathbb{R}$ , in which we let

$$\lambda_{1m}(t) \equiv \frac{1}{F_U(t)} \int_0^{F_U(t)} h_m(u) du$$

$$\lambda_{0m}(t) \equiv \frac{1}{1 - F_U(t)} \int_{F_U(t)}^1 h_m(u) du.$$

With this additively separable structure, the instrument  $Z$  is not necessarily required for point identification of the structural parameters  $\mu$ ,  $\tau$ , and  $\rho$  (Pan et al., 2024). Consider the following example:

**Example 2.5.1.** *In a Normal MTE model where  $(U, V_d) \sim \mathcal{N}(0, \Sigma_d)$ , and*

$$\Sigma_d = \begin{pmatrix} 1 & \rho_d \\ \rho_d & \sigma_d^2 \end{pmatrix},$$

*we have*

$$\mathbb{E}[V_d \mid F_U(U) = u] = \rho_d \Phi^{-1}(u),$$

*where  $\Phi^{-1}$  denotes the inverse of standard normal CDF. So this implies  $M = 1$  with  $h_1(u) = \Phi^{-1}(u)$  in the equation (2.27), where  $\lambda_d(\cdot)$  denotes the inverse Mills' ratio:*

$$\lambda_1(t) = -\frac{\phi(u)}{\Phi(u)} \quad \text{and} \quad \lambda_0(t) = \frac{\phi(u)}{1 - \Phi(u)}.$$

*In this model, the structural parameters are not identified when both  $\pi = 0$  and  $\delta(z) \equiv \delta$ . If we only have  $\delta(z) \equiv \delta$  for all  $z \in \text{supp}(Z)$  (i.e., there is no exogenous variation from IV at all), we can still point identify all parameters given that  $\text{rank}(1, W, \lambda_d(W'\pi + \delta)) = 2 + L$  almost surely. This point identification is a joint consequence of the additively separable structure and nonlinearity of the transformation  $\lambda_d(\cdot)$ . It also illustrates how the variation in covariates  $W$  can help identify the endogenous coefficients  $(\rho_1, \rho_0)$  without instruments, even if  $W$  does not satisfy the exclusion restriction.*

From this example, weak identification under additive separability depends jointly on the coefficients of covariates and instruments in the first stage. Even if instruments are weak

and cannot generate much variation, the structural coefficients might still be strongly identified if covariates strongly influence individuals' treatment decisions (that  $\pi$  is sufficiently distant from zero).

Recall that the parameters of interest are linear functionals of the MTE function:

$$\text{MTE}(w, u) = \mu_1 - \mu_0 + w'(\tau_1 - \tau_0) + \sum_{m=1}^M (\rho_{1m} - \rho_{0m})h_m(u).$$

For example, consider the PRTE:

$$\begin{aligned} \frac{\mathbb{E}[Y^\epsilon - Y]}{\mathbb{E}[D^\epsilon - D]} &= \mathbb{E}_{Z,W} \left[ \int_{p(W,Z)}^{p^\epsilon(W,Z)} \text{MTE}(W, u) du \right] / \mathbb{E}[D^\epsilon - D] \\ &= \mu_1 - \mu_0 + \mathbb{E}[W]'(\tau_1 - \tau_0) + \sum_{m=1}^M \frac{\rho_{1m} - \rho_{0m}}{\mathbb{E}[D^\epsilon - D]} \times \mathbb{E} \left[ \int_{p(W,Z)}^{p^\epsilon(W,Z)} h_m(u) du \right] \end{aligned}$$

where

$$\mathbb{E}[D^\epsilon - D] = \epsilon \quad \text{for additive PRTE}$$

$$\mathbb{E}[D^\epsilon - D] = \epsilon \cdot \mathbb{P}(D = 1) \quad \text{for proportional PRTE}$$

$$p(w, z) = F_U(w'\pi + \delta(z)).$$

To proceed with inference on the target parameter, I suggest the following step:

1. Estimate the first stage and obtain the estimated coefficients  $\hat{\pi}, \hat{\delta}(\cdot)$ .
2. Fix the parameters of  $\rho_d = \{\rho_{dm}\}_{m=1}^M$  and estimate  $\tau_d$  in separate regressions (2.27) by regressing  $Y_i - \lambda_d(W_i'\hat{\pi} + \hat{\delta}(Z_i))'\rho_d$  on  $(1, W_i')$ . Under the assumption that  $W$  has full rank (or  $\lambda_{\min}(\text{var}(W)) > \epsilon$  uniformly over the class of valid DGPs), we can obtain consistent estimators of  $\mu_d$  and  $\tau_d$  under a fixed value of  $\rho_d$ . Denote the estimators as  $\hat{\mu}_d(\rho_d)$  and  $\hat{\tau}_d(\rho_d)$ , respectively.
3. Following the previous step, we can construct the moment condition of  $(\rho_0, \rho_1)$  as below:

$$\hat{m}_1(\rho_1) = \sum_{i=1}^n f_1(W_i, Z_i) D_i \left( Y_i - \hat{\mu}_1(\rho_1) - W_i' \hat{\tau}_1(\rho_1) - \lambda_1(W_i' \hat{\pi} + \hat{\delta}(Z_i))' \rho_1 \right)$$

$$\hat{m}_0(\rho_0) = \sum_{i=1}^n f_0(W_i, Z_i) (1 - D_i) \left( Y_i - \hat{\mu}_0(\rho_0) - W_i' \hat{\tau}_0(\rho_0) - \lambda_0(W_i' \hat{\pi} + \hat{\delta}(Z_i))' \rho_0 \right),$$

where  $f_d(w, z) \in \mathbb{R}^p$  is a measurable function of  $(w, z)$  with  $p \geq M$  for  $d = 0, 1$ . The functions  $\{f_d(w, z)\}_{d=0,1}$  are chosen to avoid the issue discussed in Appendix A.8, such that the asymptotic variance of the moment functions is non-singular for all  $\rho_d \in \mathbb{R}^M$  for  $d = 0, 1$ .

4. With these moment conditions, we can conduct inference on the target parameter<sup>9</sup> by applying the improved projection method from section 2.4.

## 2.6 Monte Carlo Simulation

In this section, I compare the finite-sample performance of the proposed MLC test with the classical Wald test using simulated data from a simple quadratic MTE model with a three-valued instrument.<sup>10</sup> Consider a DGP specified as follows: For each  $d = 0, 1$ ,

$$\begin{aligned} Y_d &= \mu_d + V_d \\ D &= \mathbb{1}[U \leq p(Z)] \\ V_d &= \rho_{d1} \left( U - \frac{1}{2} \right) + \rho_{d2} \left( U^2 - \frac{1}{3} \right) + e_d, \end{aligned}$$

where  $Z$  is uniformly distributed over three points  $\{z_0, z_1, z_2\}$  and is independent of  $(U, e_1, e_0)$ . The error terms  $(e_1, e_0)$  follow a joint normal distribution with zero mean and covariance matrix  $\Sigma_e = 0.5 \cdot I_{2 \times 2}$ . For simplicity, I set  $\mu_1 = \mu_0 = 0$  and impose strong endogeneity by specifying

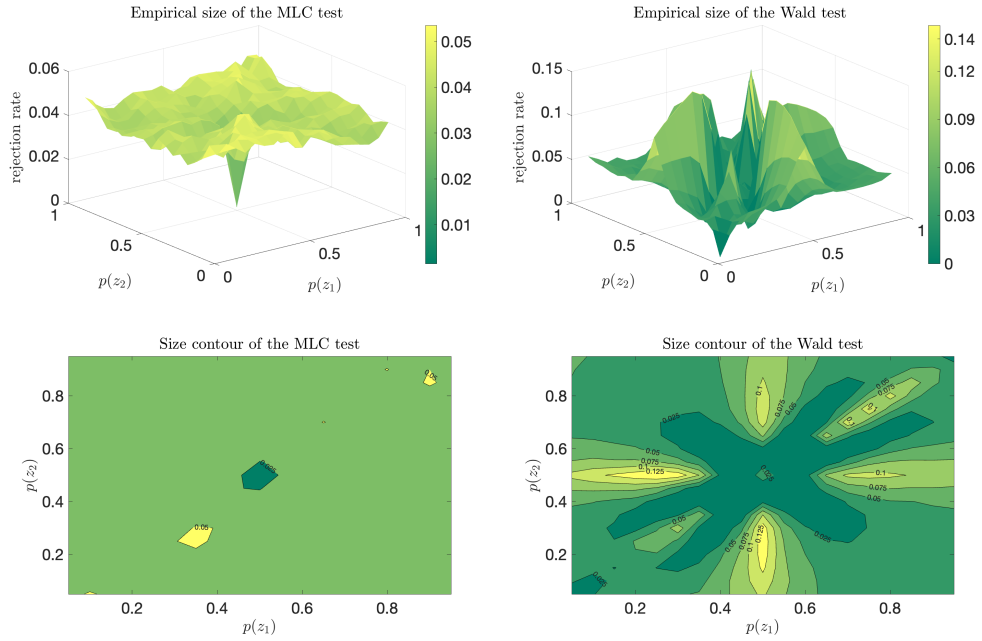
$$\rho_{11} = \rho_{12} = -5 \quad \text{and} \quad \rho_{01} = \rho_{02} = 5.$$

Regarding the specification of the propensity score  $p(z)$  for  $z \in \{z_0, z_1, z_2\}$ , I fix  $p(z_0) = 0.5$  and let  $(p(z_1), p(z_2))$  vary across  $(0.05, 0.95)^2$  to generate a variety of degrees/directions of weak identification. By drawing 2,000 i.i.d. simulation samples, I implement the Wald test and the MLC test with AR weight  $a = 0.05$  to assess the null hypothesis on testing ATE  $H_0 : \mu_1 - \mu_0 = 0$ . The average null rejection rates based on 5% significance level are

<sup>9</sup> In the target parameter,  $\mu_d$  and  $\tau_d$  are replaced with their corresponding estimator  $\hat{\mu}_d(\rho_d)$  and  $\hat{\tau}_d(\rho_d)$ , respectively. Other consistently estimable quantities are replaced with their corresponding estimators.

<sup>10</sup> The power comparison between conditional Wald tests, MLC tests, and Wald tests in a linear MTE model can be found in Appendix A.4.

displayed in Figure 2.2.



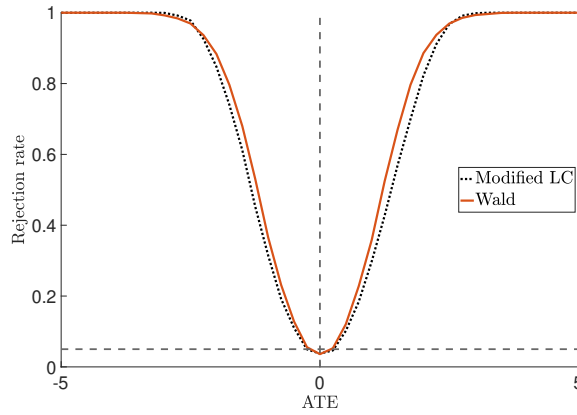
Note: The plots show empirical rejection rates for testing the null hypothesis  $H_0 : ATE = 0$  at the 5% significance level. The upper panel presents three-dimensional surfaces of rejection rates, while the lower panel shows the corresponding contour plots. Each plot varies  $p(z_1)$  and  $p(z_2)$  across  $(0.05, 0.95)$  with  $p(z_0)$  fixed at 0.5. The contour lines correspond to rejection rates of 2.5%, 5%, 7.5%, 10%, and 12.5%, with regions between these lines representing intermediate rejection rates. Darker regions indicate lower rejection rates.

FIGURE 2.2: 3D Plot of Empirical Size for MLC and Wald Tests

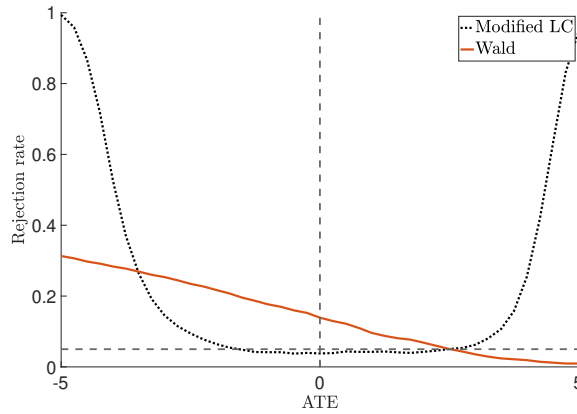
The null rejection rates reveal distinct patterns in tests performance across propensity score combinations. The conventional Wald test exhibits under-rejection when the three propensity scores are similar, indicating trivial power in this weakly identified scenario to be shown below. When the propensity scores cluster at two points (i.e., when either  $p(z_1)$  or  $p(z_2)$  equals  $p(z_0) = 0.5$ ), the ATE parameter becomes partially identified, and the Wald test's rejection rates reach to 14%, substantially exceeding the nominal 5% significance level. This demonstrates the Wald test's invalidity under identification failure. In contrast, the proposed MLC test maintains proper size control across all propensity score configurations, demonstrating its robustness to weak identification.

Figure 2.3 compares the power of the MLC and Wald tests across different identification

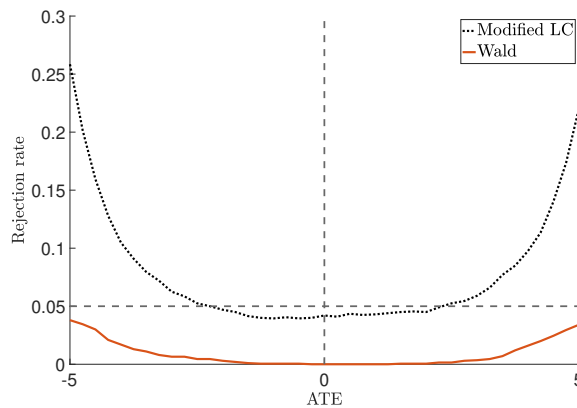
(a)  $p(z) = [0.2, 0.5, 0.8]$



(b)  $p(z) = [0.2, 0.5, 0.5]$



(c)  $p(z) = [0.4, 0.5, 0.6]$



Note: Testing ATE at values on  $[-5, 5]$  with the true effects fixed at zero. The significance level is set at 5%. The sample size equals 2,000 and the average rejection rates are computed with 2,000 independent Monte-Carlo simulations.

FIGURE 2.3: Power Curves of the MLC and Wald Test

scenarios. In Panel (a), where propensity scores are well-separated (strong identification), the Wald test achieves asymptotic efficiency, and the MLC test delivers comparable rejection power. In Panel (b), where the instrument takes only two distinct values (partial identification), the Wald test exhibits both size distortion at the null and power loss at alternatives, while the MLC test maintains correct size and achieves full power against distant alternatives outside the identified set. In Panel (c), where propensity scores are clustered together (weak identification), the Wald test’s rejection rates fall consistently below the nominal level, indicating negligible power, while the MLC test retains nontrivial power at distant alternatives.

## ***2.7 Empirical Application to Misdemeanor Prosecution***

In this section, I revisit the empirical analysis by Agan et al. (2023), who examined the causal effects of misdemeanor prosecution on defendants’ subsequent criminal activity. Based on the quasi-randomized assignment of nonviolent misdemeanor cases to assistant district attorneys (ADAs), their LATE and MTE estimates demonstrate that nonprosecution leads to a large reduction in the likelihood of defendants’ future criminal involvement over the next two years. To study the policy effects of increasing nonprosecution, they analyzed the impact of imposing a presumption of misdemeanor nonprosecution by relying on an existing policy change issued by a new district attorney. In this section, I use their data to answer some policy-relevant questions concerning the exogenous change on ADA nonprosecution rates. My approach does not require the additional data or information related to an actual policy that is already implemented, but instead extrapolates causal effects to address this issue.

Similar on their analysis, I use the MTE framework to extrapolate treatment effects outside compliers. However, my goal is to analyze causal effects of implementing several counterfactual policies on ADA leniency, which were not explored in their empirical studies. To adapt their data into the framework considered in this paper, I implement the proposed inference procedure conditional on each court and use ADAs’ identity as the discrete in-

strument since ADAs are randomly assigned conditional on courts and time. To avoid bias introduced by many instruments (see the references in Mikusheva and Sun, 2022 for further discussions), I combine ADAs into a total of 15 groups for each court, excluding the smaller courts (BRI, CHE, and CHA) due to insufficient ADA counts.

Let the treatment  $D$  be the nonprosecution status of defendants (that takes value 1 if the defendant is not prosecuted) and the outcome  $Y$  be the indicator of subsequent criminal complaints within two years postarrest. To validate the strong overlap condition on propensity scores, I focus on ADAs that handle more than 50 cases and have nonprosecution rate at least 0.025 within each court. The summary statistics of ADAs' nonprosecution rates conditional on each court are provided in Table 2.2 below. From this table we can see that the range of propensity scores varies from 0.21 to 0.50. When employing a high-order polynomial model on the control functions such as the cubic specification in Figure IV of Agan et al. (2023) in this setup, the finite-sample performance of confidence sets may be too poor to guarantee valid coverage. Specifically, in Appendix A.5, I show the evidence of weak instruments in cubic and quartic MTE models, while the classical  $F$  test (with threshold 10) does not deliver the same conclusion. The reason is that  $F$  statistic aims to detect deviations from the null where all propensity scores are equal to each other. However, this is not sufficient to strongly identify MTE models with flexible structure that require three or more propensity score to be well separated from each other (see Figure 2.1).

For all the counterfactual experiments, I consider an exogenous change to the ADA nonprosecution rates. Let  $p^\epsilon(z)$  be the counterfactual propensity score of not prosecuting defendants, where  $\epsilon$  is a nonnegative scalar (or vector) denoting the deviation from status quo. I compare the induced outcome  $Y^\epsilon$ , defined as

$$Y^\epsilon = Y_1 \mathbb{1}[U \leq p^\epsilon(Z)] + Y_0 \mathbb{1}[U > p^\epsilon(Z)]$$

with the observed outcome  $Y$  in the data.

Similar to the policy invariance assumption imposed by J. J. Heckman and Vytlačil (2005), suppose that this counterfactual change on propensity score does not shift the dis-

Table 2.2: Summary Statistics of Nonprosecution Rates across Courts

Court	Number of ADA	Nonprosecution Rate			Sample Size
		Min	Mean	Max	
SBO	16	0.04	0.24	0.54	3,921
EBOS	22	0.08	0.32	0.53	7,566
WROX	43	0.05	0.37	0.55	8,905
BMC	65	0.04	0.17	0.4	9,593
ROX	66	0.06	0.16	0.33	13,333
DOR	73	0.04	0.13	0.43	13,523
CHA	4	0.09	0.24	0.42	362
CHE	12	0.07	0.16	0.3	872
BRI	5	0.14	0.29	0.48	885
Total	262	0.04	0.22	0.55	58,960

tribution of potential outcomes and unobserved cost  $U$  of being selected into treatment, therefore omitting any “general equilibrium” effects that may arise after the change of prosecution rates. I consider three different comparisons between the counterfactual outcome  $Y^\epsilon$  and the observed outcome  $Y$ :

1. Non-normalized policy effects:

$$\alpha(\epsilon) \equiv \mathbb{E}[Y^\epsilon - Y]. \quad (2.28)$$

This criterion follows Heckman and Vytlacil (2001). The term "non-normalized" distinguishes these effects from the policy-relevant treatment effects discussed below.

2. Policy-relevant treatment effects (PRTE)

$$\bar{\alpha}(\epsilon) \equiv \frac{\mathbb{E}[Y^\epsilon - Y]}{\mathbb{E}[p^\epsilon(Z) - p(Z)]}. \quad (2.29)$$

This criterion follows J. J. Heckman and Vytlacil (2005). This quantity can be interpreted as the treatment effect for units shifted into treatment via the counterfactual policy. Note that the additive/proportional PRTEs suggested by Carneiro et al. (2010, 2011) and Mogstad and Torgovitsky (2018) are special cases of  $\bar{\alpha}(\epsilon)$  for particular choices of  $p^\epsilon$ , as described in equations (2.31) and (2.32).

### 3. Marginal policy-relevant treatment effects (MPRTE)

$$\alpha_+(0) \equiv \lim_{\epsilon \searrow 0} \frac{\mathbb{E}[Y^\epsilon - Y]}{\mathbb{E}[p^\epsilon(Z) - p(Z)]}. \quad (2.30)$$

This criterion follows Carneiro et al. (2010). When  $p^\epsilon(z) \geq p(z)$ , this parameter can be interpreted as the average treatment effects for marginal defendants at the edge of being prosecuted. The equivalence between average MTE and MPRTE  $\alpha_+(0)$  has been established by Carneiro et al. (2010).

Next, I discuss the counterfactual propensity scores of interests and present the confidence sets for these policy effects.

#### 2.7.1 Marginal policy relevant treatment effects

First, suppose policymakers increase the ADAs’ nonprosecution rates up to a positive constant  $\epsilon$  or to a proportion  $\epsilon$ , where  $\epsilon > 0$ . Consider the additive or proportional marginal PRTE  $\alpha_+(0)$  under the new policy:

$$\text{Additive change: } p^\epsilon(z) = p(z) + \epsilon, \quad (2.31)$$

$$\text{Proportional change: } p^\epsilon(z) = (1 + \epsilon)p(z) \quad (2.32)$$

respectively.

Table 2.3 collects the 95% and 90% confidence sets on marginal policy effects  $\alpha_+(0)$  for the additive leniency increase in equation (2.31) by letting  $\epsilon$  approach zero. The “unconditional” confidence sets are obtained without controlling for court identities, whereas the “average” confidence sets are weighted average of court-specific confidence sets weighted by the size of courts. The differences between these two sets of results highlight the importance of adjusting for court identity as a confounding variable. Incorporating court identity generally increases the uncertainty of causal effect predictions.

Comparing average Wald confidence sets with average MLC confidence sets, the results show that weak identification is a potential concern for models with polynomial orders higher than cubic. In such cases, robust confidence sets may not be informative about the sign of causal effects due to the inability to estimate a highly flexible model by using limited

variation of propensity scores. Although the ranges of both confidence sets are similar under the quadratic specification, Wald confidence sets remain negatively significant, whereas MLC confidence sets lose such significance due to weak identification. In Appendix A.5 (Table A.2), I evaluate the strength of identification for the marginal PRTEs based on the additive change in (2.31). The findings further confirm the presence of weak identification for MTE models with cubic or quartic orders.

Table 2.3: 95% (top) and 90% (bottom) Confidence Sets for Additive MPRTE  $\alpha_+(0)$

	MTE polynomial	Wald	MLC
Uncond.	Linear	[-0.19, -0.11]	[-0.17, -0.13]
Uncond.	Quadratic	[-0.19, -0.11]	[-0.18, -0.13]
Uncond.	Cubic	[-0.17, -0.06]	[-0.15, -0.08]
Uncond.	Quartic	[-0.18, -0.05]	[-0.17, -0.06]
Average	Linear	[-0.24, -0.03]	[-0.23, -0.03]
Average	Quadratic	[-0.29, -0.01]	[-0.33, 0.09]
Average	Cubic	[-0.32, 0.03]	[-0.79, 0.59]
Average	Quartic	[-0.39, 0.01]	[-0.76, 0.61]
Uncond.	Linear	[-0.18, -0.12]	[-0.16, -0.14]
Uncond.	Quadratic	[-0.19, -0.12]	[-0.17, -0.14]
Uncond.	Cubic	[-0.17, -0.07]	[-0.14, -0.09]
Uncond.	Quartic	[-0.17, -0.06]	[-0.13, -0.08]
Average	Linear	[-0.22, -0.04]	[-0.21, -0.06]
Average	Quadratic	[-0.27, -0.04]	[-0.28, 0.01]
Average	Cubic	[-0.29, -0.00]	[-0.60, 0.34]
Average	Quartic	[-0.35, -0.02]	[-0.63, 0.38]

### 2.7.2 Quota

Instead of mandating all ADAs to increase their nonprosecution rates simultaneously, it might be more convenient for policymakers to set up a lower and upper bound for ADA nonprosecution rates. For example, let  $\epsilon = (\underline{\epsilon}, \bar{\epsilon}) \in [0, 1]^2$  with  $\underline{\epsilon} \leq \bar{\epsilon}$  denoting the lower and upper bounds of ADAs' nonprosecution rate, respectively. Then the counterfactual propensity score becomes

$$p^\epsilon(z) = \min(\max\{\underline{\epsilon}, p(z)\}, \bar{\epsilon})$$

Since this counterfactual propensity score is non-smooth in  $p(z)$  which may invalidate the inferential results, I employ a smooth approximation as follows:

$$p^\epsilon(z; \phi) = -\frac{1}{\phi} \log \left( \frac{1}{e^{\phi p(z)} + e^{\phi \underline{\epsilon}}} + \frac{1}{e^{\phi \bar{\epsilon}}} \right).$$

One can show that  $p^\epsilon(z; \phi) \rightarrow p^\epsilon(z)$  as  $\phi \rightarrow \infty$ . In practice, I set  $\phi = 30$  to approximate the counterfactual propensity score  $p^\epsilon(z)$ . I analyze the (non-)normalized policy effects when policymakers implement a lower bound  $\underline{\epsilon} \in [0.05, 0.3]$  while maintaining  $\bar{\epsilon} = 1$  for each court. For notational consistency, I use  $\epsilon$  to denote the value of this lower bound instead of  $\underline{\epsilon}$ .

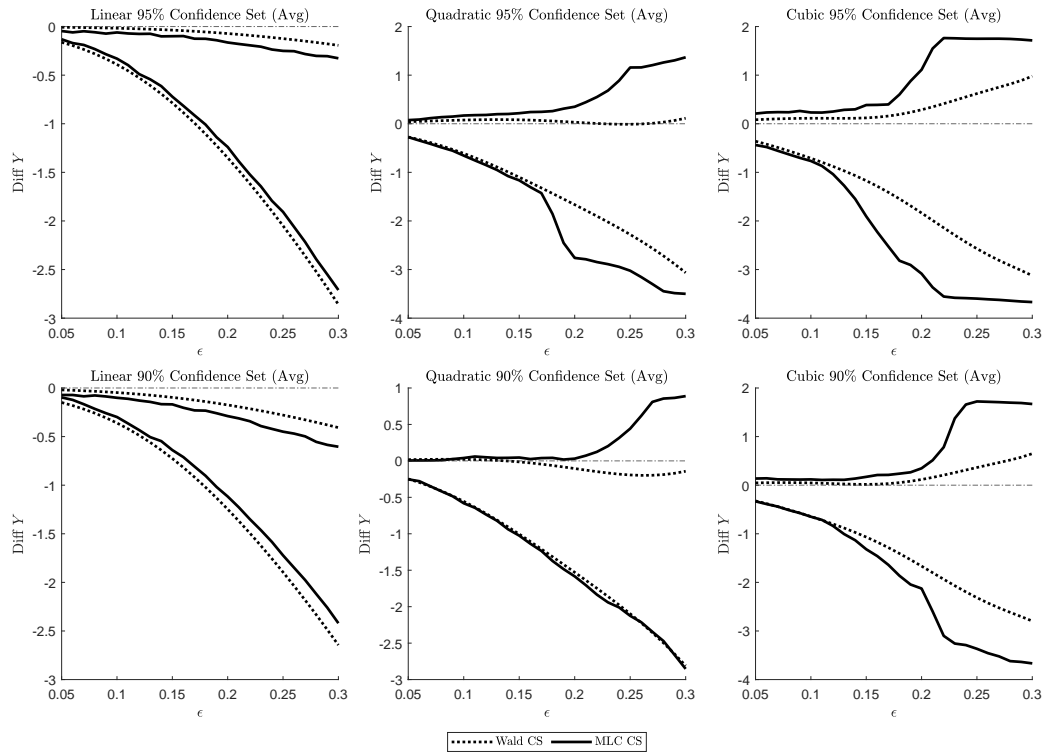


FIGURE 2.4: 95% (top) and 90% (bottom) Average Confidence Sets for Non-normalized Policy Effects  $\alpha(\epsilon)$  when Setting a Lower Bound  $\underline{\epsilon}$  for Nonprosecution Rate

Figure 2.4 displays the average confidence sets for the non-normalized policy effects  $\alpha(\epsilon)$  of implementing a lower bound on nonprosecution rates. Under linear extrapolation, the

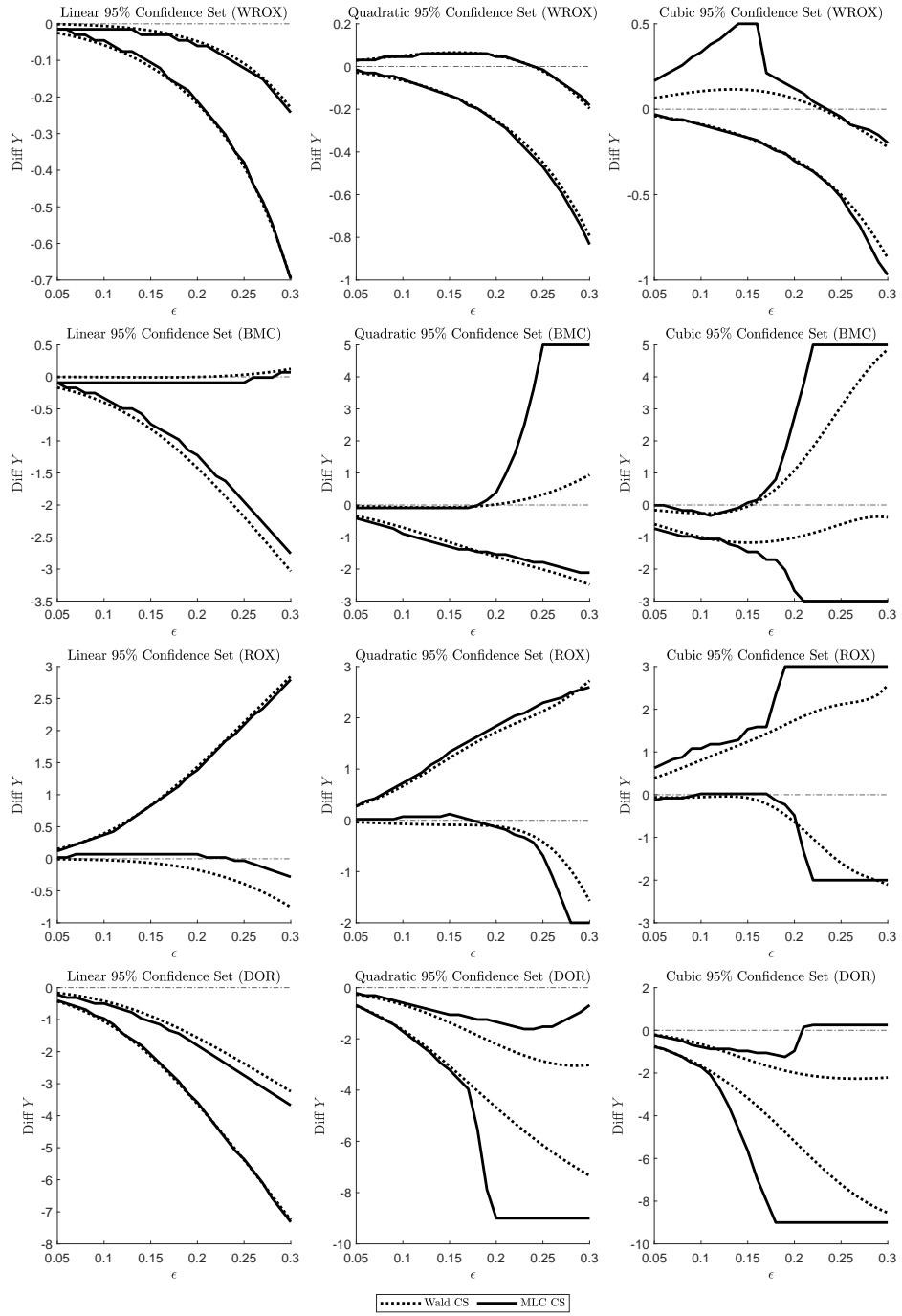


FIGURE 2.5: 95% Court-specific Confidence Sets for Non-normalized Policy Effects  $\alpha(\epsilon)$  when Setting a Lower Bound  $\underline{\epsilon}$  for Nonprosecution Rate

results suggest that imposing a lower quota on nonprosecution is beneficial. However, this evidence becomes less conclusive with quadratic and cubic specifications as the lower bound  $\epsilon$  increases, reflecting limitations of available exogenous variation at higher extrapolation levels. The Wald confidence sets appear "spuriously precise" by failing to account for this limited IV variation. Consequently, policy recommendations derived from the robust MLC approach may differ from those based on the classical Wald approach. For instance, based on the 90% confidence sets from the quadratic MTE model, policymakers concerned with robustness to weak IVs may prefer setting the minimum nonprosecution rate below 20%, since the reductions in recidivism become insignificant above this threshold under robust confidence sets.

Figure 2.5 shows the court-specific 95% confidence sets for the four largest courts in the sample. Unlike other courts, court ROX exhibits increases in recidivism from imposing the quota on nonprosecution rates. For courts WROX and BMC, the implied thresholds of  $\epsilon$  for a significant crime reduction suggested by the MLC confidence sets are similar to those under the Wald confidence sets, indicating that these thresholds are robust to weak identification. These thresholds also inform heterogeneous policy recommendations. For instance, a significant reduction in recidivism is observed in court WROX if policymakers set the quota at  $\epsilon > 0.23$ , whereas for court BMC, such significant effects are only observed if  $\epsilon < 0.15$ .

## **2.8 Conclusion**

In this paper, I propose two robust inference procedures for a class of causal effects identified by MTEs with discrete instruments. In a linear MTE model, I introduce a conditional Wald test, which is simple and asymptotically similar regardless of identification strength. In a more generic polynomial MTE model, I propose a modified linear combination test that achieves uniform validity against weak identification and has satisfactory power properties under strong identification. Finally, I use the proposed methods to investigate the counterfactual effects of policy changes on ADAs' leniency in the study of misdemeanor

prosecution by Agan et al. (2023).

There are several avenues for future research. First, while this paper focuses on discrete variation from instruments, extending the weak IV analysis to a semiparametric MTE model with continuous propensity scores would be empirically relevant. In addition, the MLC test requires the choice of tuning parameter on the weight of AR statistics, which trades off the power of test under different identification strengths (see the discussion in Appendix A.2). It would be interesting to investigate the optimal choice of this tuning parameter following the regret analysis of I. Andrews (2016). Finally, extending robust inference techniques to accommodate many weak instruments would be valuable, especially considering the typically large number of judges involved in empirical studies (Jochmans, 2023).

### 3. Marginal Homogeneity Tests with Panel Data

In this essay, we study the problem of testing marginal homogeneity with the panel data. This essay is a joint work with Federico Bugni and Jackson Bunting.

#### 3.1 Introduction

This paper considers a hypothesis testing problem for panel data,  $\{\{X_{i,t}\}_{t=1}^T\}_{i=1}^n$ . We assume that the data are independent and identically distributed (i.i.d.) across units  $i = 1, \dots, n$ , but allow for arbitrary dependence across time. For each period  $t = 1, \dots, T$ , let  $F_t$  denote the common marginal cumulative distribution function (CDF) of  $X_{i,t}$ . We say that the data generating process satisfies *marginal homogeneity* if these marginal distributions are homogeneous or time-invariant, i.e.,

$$F_1 = F_2 = \dots = F_T. \quad (3.1)$$

In this paper, we propose several tests for the hypothesis of marginal homogeneity and investigate their properties.

Marginal homogeneity is relevant in economic models such as dynamic discrete games. In this case, researchers often observe  $X_{i,t} = (A_{i,t}, S_{i,t})$ , representing action and state variables for units  $i = 1, \dots, n$  (individuals, firms, etc.) over  $t = 1, \dots, T$  periods, respectively. In this context, a standard approach is to assume that the conditional choice probabilities (i.e.,  $P(A_{i,t} = a \mid S_{i,t} = s)$ ) and state transition probabilities (i.e.,  $P(S_{i,t+1} = s' \mid A_{i,t} = a, S_{i,t} = s)$ ) are homogeneous, and posit a structural model for them. Under standard assumptions, these objects yield a homogeneous structural model for  $P(X_{i,t+1} = x' \mid X_{i,t} = x) = f_\theta(x', x)$  with parameter  $\theta$ . This posited structure forms the basis of inference on  $\theta$  in dynamic discrete choice games. Notably, this inference does not invoke the marginal homogeneity hypothesis in (3.1). However, if this condition holds, it provides valuable efficiency gains in the estimation of  $\theta$ . To see this, note that marginal homogeneity in this context implies that  $X_{i,t}$  is in a steady state with a marginal CDF  $F$ , which yields the following

structural equation:

$$dF(x') = \int_{x \in \mathcal{X}} f_{\theta}(x', x) \times dF(x) \quad \text{for all } x' \in \mathcal{X} \quad (3.2)$$

Imposing (3.2) in the estimation of  $\theta$  can deliver substantial efficiency gains relative to the standard method that does not impose it. In this sense, the marginal homogeneity hypothesis in (3.1) is a source of efficiency gains in the structural estimation of dynamic discrete choice games.

Beyond dynamic discrete games, marginal homogeneity tests have been applied to financial data. For example, Ditzhaus and Gaigall (2022) (see also references therein) tests for possible dependence between two stock market indices. In terms of our notation, we can express their data as  $\{\{X_{i,t}\}_{t=1}^2\}_{i=1}^n$ , where  $\{X_{i,1}\}_{i=1}^n$  represents the monthly returns of the first stock market (e.g., Nikkei 225 Stock Average) and  $\{X_{i,2}\}_{i=1}^n$  represents the monthly returns of the second market (e.g., Dow Jones Industrial Average). Invoking classical models for stock prices, Ditzhaus and Gaigall, 2022 posit monthly returns to be i.i.d. across  $i = 1, \dots, n$  (i.e., months), while the interconnectedness of global financial markets implies that there may be dependence across  $t = 1, 2$  (i.e., stock markets). In this context, the marginal homogeneity hypothesis in (3.1) implies that the various stock markets have equally distributed returns. Finally, we note that their paper focuses on pairs of stock markets (i.e.,  $T = 2$ ), whereas we consider applications with  $T > 2$ .

An inherent feature of the preceding examples is that the data are likely to exhibit dependence across time periods. In dynamic discrete games, actions and states depend on their past values. In fact, this is precisely what gives the discrete game its dynamic nature. In the application presented in Ditzhaus and Gaigall (2022), stock returns across the globe are likely to be interrelated. Beyond these two examples, dependence over time is common when dealing with panel data. For this reason, the classical goodness-of-fit testing literature focused on independent samples (e.g., Lehmann & Romano, 2022, Section 17.2.1), and its generalization to  $T$ -sample problems, does not apply.

This paper studies the  $T$ -sample testing problem with possibly dependent data. Namely,

we implement our tests by comparing a studentized or non-studentized  $T$ -sample version of the Cramér-von Mises statistic with a suitable critical value. We consider three possible methods to construct the critical value: asymptotic approximations, the bootstrap, and time permutations. We show that the first two methods lead to asymptotically exact hypothesis tests, with or without studentization. Results for the permutation test are more nuanced: the permutation test based on a non-studentized statistic is asymptotically exact when  $T = 2$ , but is asymptotically invalid when  $T > 2$ . Once studentized, the permutation test is always asymptotically exact. Finally, under a time-exchangeability assumption, the permutation test is exact in finite samples, both with or without studentization. On the other hand, relative to the non-studentized case, the asymptotic analysis of the studentized statistics requires an additional assumption: the variance-covariate matrix used in the studentization must be non-singular, an assumption that can fail in practice. See the related discussion in Section 3.2.2 and our empirical application in Section 3.6.

For independent cross-sectional data, the marginal homogeneity hypothesis in (3.1) becomes the standard equality-of-distribution hypothesis for  $T$ -sample data, which has been thoroughly studied in the literature. In such case, Lehmann and Romano (2022, Theorem 17.2.1) shows that permutation tests of homogeneity are finite-sample exact. Chung and Romano (2013) explores the behavior of permutation tests with studentized test statistics. Relatedly, Bugni and Horowitz (2021) studies the application of permutation tests to functional cross-sectional data. Relatively speaking, the test for marginal homogeneity hypothesis in (3.1) with panel data (i.e., allowing for time dependence) has received less attention. Gaigall (2020) is among the first papers to test for marginal homogeneity in panel data with  $T = 2$ . In subsequent work, Ditzhaus and Gaigall (2022) broadened the analysis with  $T = 2$  to paired functional data. These papers show that permutation tests are asymptotically valid with two periods of panel data. Neither of these papers considers the case with  $T > 2$ , which is common in economic applications. To our knowledge, our paper is the first one to consider testing for marginal homogeneity in panel data models, allowing for  $T > 2$ . In this respect, one remarkable finding of our paper reveals that the

asymptotic validity of the non-studentized permutation test breaks down when we consider  $T > 2$ .

A related testing problem in dynamic discrete choice games is concerned with evaluating the homogeneity of the state transition probabilities, i.e.,  $P(X_{i,t+1} = x' \mid X_{i,t} = x) = P(X_{i,t'+1} = x' \mid X_{i,t'} = x)$  for all  $t, t' < T - 1$ . See Otsu et al. (2016) and Bugni et al. (2024) for recent contributions on this topic. The homogeneity of state transition probabilities and the marginal homogeneity in (3.1) are non-nested hypotheses, and so our contribution is complementary to but distinct from these references.

In other related work, Pauly et al. (2015) and Friedrich et al. (2017) investigate the validity of permutation tests to evaluate the presence of treatment effects in experiments under factorial designs. There are important differences between these papers and ours. The first key difference is in the class of data permutations used to implement the tests. Pauly et al. (2015) and Friedrich et al. (2017) generate their test by permuting observations in both units and time indices. In contrast, we generate our test by permuting the time index of our observations. We show that the classes of distributions under which the two types of permutation tests are finite-sample valid are non-nested; see Lemma B.1.2. Another difference is that Pauly et al. (2015) and Friedrich et al. (2017) focus on studentized statistics, while we consider both studentized and non-studentized statistics. In this respect, it is relevant to point out that analyzing studentized statistics requires additional assumptions compared to non-studentized ones; see discussion in Section 3.2.2. Finally, our Monte Carlo simulations suggest that our permutation test appears more powerful than theirs in finite samples. This is related to the fact that our permutation test only considers time index permutations, which provide a better contrast to detect departures from the marginal homogeneity hypothesis in (3.1).

The remainder of the paper is organized as follows. Section 3.2 introduces the hypothesis test problem in greater detail. Section 3.3 contains our main theoretical results. Section 3.4 discusses the power of the proposed inference methods. In Section 3.5, we evaluate the finite-sample performance of these tests via Monte Carlo simulations. Section 3.6 considers

an empirical application based on Igami and Yang (2016). Section 3.7 concludes. The paper’s appendix collects all of the proofs and several auxiliary results.

### 3.2 The Hypothesis Testing Problem

As the introduction explains, this paper considers a hypothesis-testing problem for panel data with  $n$  units and  $T$  time periods. Inspired by the typical application in economics, we consider an asymptotic framework in which  $n$  grows and  $T$  remains fixed. We denote the data by  $\mathbf{X}_n = \{\{X_{i,t}\}_{t=1}^T\}_{i=1}^n$ . As explained earlier, we allow the data to be arbitrarily dependent across time  $t = 1, \dots, T$ , and assume i.i.d. across units  $i = 1, \dots, n$ . We formalize this assumption next.

**Assumption 9.** *For all  $t = 1, \dots, T$ ,  $\{X_{i,t}\}_{i=1}^n$  are i.i.d. with marginal CDF  $F_t$ .*

Our goal is to test whether the marginal homogeneity hypothesis in (3.1) holds in the data, i.e.,

$$H_0 : (3.1) \text{ holds} \quad \text{vs.} \quad H_1 : (3.1) \text{ does not hold.} \quad (3.3)$$

We propose implementing this hypothesis test by rejecting  $H_0$  in (3.3) whenever a test statistic exceeds a suitable critical value. That is, for any significant level of  $\alpha \in (0, 1)$ , we propose

$$\phi_n(\alpha) = 1 \{S_n > c_n(1 - \alpha)\}, \quad (3.4)$$

where  $\phi_n(\alpha)$  indicates the test function,  $S_n$  denotes the test statistic, and  $c_n(1 - \alpha)$  indicates the critical value. In the remainder of this section, we describe the test statistic (Section 3.2.1) and establish its asymptotic distribution under the null hypothesis of marginal homogeneity (Section 3.2.2). With these results in place, Section 3.3 provides three inference methods, each based on a different type of critical values.

#### 3.2.1 Test statistics

We propose implementing our test using the Cramér-von Mises (CvM) statistic, given by the sample-weighted sum of squared differences of the empirical CDFs for all consecutive periods. For simplicity, we evaluate these differences on a finite number of user-defined

points on the real line  $\mathcal{U}_K = \{u_1, u_2, \dots, u_K\}$  with  $u_0 := -\infty < u_1 < u_2 < \dots < u_K$ . For reasons that will be explained soon, we refer to this as the *non-studentized* CvM statistic, given by

$$S_n = n \sum_{k=1}^K \sum_{t=1}^{T-1} [\hat{F}_t(u_k) - \hat{F}_{t+1}(u_k)]^2 \hat{P}(u_k), \quad (3.5)$$

where  $\hat{F}_t$  is the empirical CDF in period  $t = 1, \dots, T$ , and  $\hat{P}(u_k)$  is the empirical analog of the aggregate probability in the interval  $(u_{k-1}, u_k]$ , i.e.,  $k = 1, \dots, K$ ,

$$\hat{P}(u_k) = \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n 1(u_{k-1} < X_{i,t} \leq u_k). \quad (3.6)$$

It is easy to see that (3.5) can be reexpressed as follows

$$S_n = \hat{Z}' \hat{Z},$$

where  $\hat{Z} \in \mathbb{R}^{(T-1)K}$  and, for all  $(t, k) \in \{1, \dots, T-1\} \times \{1, \dots, K\}$ ,

$$\hat{Z}_{(t-1)K+k} = \sqrt{n \hat{P}(u_k)} [\hat{F}_t(u_k) - \hat{F}_{t+1}(u_k)]. \quad (3.7)$$

Under  $H_0$  in (3.3), our formal arguments (see the proof of Theorem 3.2.1) reveal that  $\hat{Z}$  is asymptotically distributed according to  $N(\mathbf{0}, \Sigma_Z)$ , where for each  $(t, k), (\tilde{t}, \tilde{k}) \in \{1, \dots, T-1\} \times \{1, \dots, K\}$ ,

$$\begin{aligned} \Sigma_Z[(t-1)K+k, (\tilde{t}-1)K+\tilde{k}] &= \\ & \sqrt{P(u_k)P(u_{\tilde{k}})} \times \text{cov}[1(X_{i,t} \leq u_k) - 1(X_{i,t+1} \leq u_k), 1(X_{i,\tilde{t}} \leq u_{\tilde{k}}) - 1(X_{i,\tilde{t}+1} \leq u_{\tilde{k}})], \end{aligned} \quad (3.8)$$

and  $P(u_k)$  is the aggregate probability in the interval  $(u_{k-1}, u_k]$ ,

$$P(u_k) = \frac{1}{T} \sum_{t=1}^T [F_t(u_k) - F_t(u_{k-1})]. \quad (3.9)$$

As a corollary,  $S_n$  has a generalized chi-squared asymptotic distribution under  $H_0$  in (3.3), with weights determined by the eigenvalues of  $\Sigma_Z$ . The dependence of the limiting distribution on  $\Sigma_Z$  explains why we refer to  $S_n$  as the *non-studentized* CvM statistic.

If  $\Sigma_Z$  is a non-singular matrix, it is natural to also consider a studentized version of the CvM statistic. To this end, we consider the *studentized* CvM statistic, given by

$$\bar{S}_n = \hat{Z}' \hat{\Sigma}_Z^- \hat{Z}, \quad (3.10)$$

where  $\hat{\Sigma}_Z^-$  is the generalized inverse of the empirical analog of  $\Sigma_Z$ , denoted by  $\hat{\Sigma}_Z$ . For each  $(t, k), (\tilde{t}, \tilde{k}) \in \{1, \dots, T-1\} \times \{1, \dots, K\}$ , we define  $\hat{\Sigma}_Z$  as follows:

$$\begin{aligned} \hat{\Sigma}_Z[(t-1)K + k, (\tilde{t}-1)K + \tilde{k}] \equiv \\ \sqrt{\hat{P}(u_k)\hat{P}(u_{\tilde{k}})} \times \frac{1}{n} \sum_{i=1}^n \left[ \begin{array}{c} \left(1(X_{i,t} \leq u_k) - \hat{F}_t(u_k) - 1(X_{i,t+1} \leq u_k) + \hat{F}_{t+1}(u_k)\right) \times \\ \left(1(X_{i,\tilde{t}} \leq u_{\tilde{k}}) - \hat{F}_{\tilde{t}}(u_{\tilde{k}}) - 1(X_{i,\tilde{t}+1} \leq u_{\tilde{k}}) + \hat{F}_{\tilde{t}+1}(u_{\tilde{k}})\right) \end{array} \right] \end{aligned} \quad (3.11)$$

for each  $(t, k), (\tilde{t}, \tilde{k}) \in \{1, \dots, T-1\} \times \{1, \dots, K\}$ . If  $H_0$  in (3.3) holds and  $\Sigma_Z$  is a non-singular,  $\bar{S}_n$  has a chi-squared asymptotic distribution with  $(T-1)K$  degrees of freedom. The lack of dependence of this limiting distribution on  $\Sigma_Z$  justifies referring to  $\bar{S}_n$  as the *studentized* CvM statistic.

**Remark 3.2.1** (On the choice of test-statistics). *The test statistics in (3.5) and (3.10) are CvM-type statistics evaluated over a finite set of points  $\mathcal{U}_K$ . In principle, our analysis extends to other types of test statistics evaluated over these points, such as the Kolmogorov-Smirnov statistic. It is also worth reiterating that  $\mathcal{U}_K$  can be arbitrarily chosen by the researcher. In this respect, the most important simplifying aspect in (3.5) and (3.10) is that we use a finite number of points rather than an infinite number of points, such as a continuum.*

*There are several reasons to prefer a finite set  $\mathcal{U}_K$  over a continuum. First and foremost, it leads to simpler asymptotic analysis regarding the studentization of test statistics. Second, for applications in which the data are discrete and with finite support  $S_X$ , one can set  $\mathcal{U}_K = S_X \setminus \{\max S_X\}$  without any loss of information (note that equality of marginal distributions for all points in  $S_X \setminus \{\max S_X\}$  is equivalent to  $H_0$ ). Many empirical applications, including the one in Section 3.6, feature discrete data with finite support. Third,*

the extension to the continuum case could be implemented along the lines of the Khmaladze (2016) transformation. See also the related work by Chung and Olivares (2021). Having said this, this extension requires new and considerably more complicated arguments, which we consider out of the scope of our paper.

### 3.2.2 Asymptotic distribution under the null hypothesis

In this section, we derive the asymptotic distribution of the non-studentized CvM statistic  $S_n$  in (3.5) and the studentized version  $\bar{S}_n$  in (3.10). Our characterization of the asymptotic distribution of the studentized CvM statistic relies on the following assumption.

**Assumption 10.**  $\Sigma_Z$  is positive definite.

We note that Assumption 10 is required for our asymptotic analysis of the studentized CvM statistic and is not necessary in the case of the non-studentized version. For a suitable choice of  $\mathcal{U}_K$ , Assumption 10 is widely applicable to many data-generating processes. We now describe scenarios in which this assumption does not hold. First, note that Assumption 10 would fail if  $\mathcal{U}_K$  includes any point that is “irrelevant” with respect to the support of  $\mathbf{X}_n$ , i.e.,  $\hat{P}(u_k) = 0$  for some  $k = 1, \dots, K$ . An example of this occurs for any  $u_k \in \mathcal{U}_K$  that lies below the support of  $\mathbf{X}_n$ . In this case, one can always restore the validity of Assumption 10 by removing all of these irrelevant points. Second, one should never include  $u_k \in \mathcal{U}_K$  that equals or exceeds the support of  $\mathbf{X}_n$ . Doing this would result in  $1(X_{i,t} \leq u_k) = 1(X_{i,t+1} \leq u_k) = 1$ , leading to  $1(X_{i,t} \leq u_k) - 1(X_{i,t+1} \leq u_k) = 0$ , rendering  $\Sigma_Z$  singular. Finally,  $\Sigma_Z$  would be singular if there is no full communication between some states. For example, consider a Markov chain for  $X_{i,t} \in \{1, 2, 3, 4\}$  with  $P(X_{i,t+1} = x' | X_{i,t} = x) = \Pi[x, x']$  for all  $x, x' \in \{1, 2, 3, 4\}$ , where for any  $\rho_1, \rho_2, \rho_3, \rho_4 \in (0, 1)$ ,

$$\Pi = \begin{bmatrix} \rho_1 & 1 - \rho_1 & 0 & 0 \\ 1 - \rho_2 & \rho_2 & 0 & 0 \\ 0 & 0 & \rho_3 & 1 - \rho_3 \\ 0 & 0 & 1 - \rho_4 & \rho_4 \end{bmatrix}$$

This transition matrix implies no communication between the first and last two states. Then,  $1(X_{i,t} \leq u_k) - 1(X_{i,t+1} \leq u_k) = 0$  for any  $u_k \in [2, 3)$ , producing a singular  $\Sigma_Z$ .

The next result establishes the asymptotic distribution of the non-studentized and studentized CvM statistics under the marginal homogeneity hypothesis in (3.3).

**Theorem 3.2.1.** *Let Assumption 9 hold.*

(a) *Under  $H_0$  in (3.3),*

$$S_n \xrightarrow{d} S \equiv \sum_{j=1}^{(T-1)K} \lambda_j \zeta_j^2, \quad (3.12)$$

where  $\{\lambda_j\}_{j=1}^{(T-1)K}$  are the eigenvalues of  $\Sigma_Z$  in (3.8) and  $\{\zeta_j\}_{j=1}^{(T-1)K}$  are i.i.d.  $N(0, 1)$ .

(b) *Under Assumption 10 and  $H_0$  in (3.3),*

$$\bar{S}_n \xrightarrow{d} \chi_{(T-1)K}^2, \quad (3.13)$$

where  $\chi_{(T-1)K}^2$  denotes the chi-squared distribution with  $(T-1)K$  degrees of freedom.

Theorem 3.2.1 shows that under the marginal homogeneity hypothesis, the non-studentized CvM statistic in (3.5) converges to a generalized chi-square distribution with the weights determined by the eigenvalues of  $\Sigma_Z$ . Since  $\Sigma_Z$  is not necessarily positive definite, some eigenvalues may be zero, leading to reduced degrees of freedom. When  $\Sigma_Z$  is positive definite, then the limiting distribution of studentized CvM statistic in (3.10) is chi-square distributed. These results are the basis of the critical values proposed in the following section.

### **3.3 Critical Values and Validity of Inference**

This section describes three critical values for the CvM test statistics proposed in Section 3.2.1. Each critical value gives rise to a different hypothesis test according to equation (3.4). We formally study the validity of each one of these methods.

#### **3.3.1 Asymptotic approximation**

In this section, we propose a hypothesis test for (3.3) by approximating the quantiles of the asymptotic distribution in Theorem 3.2.1. To this end, we now introduce some notations. For any  $\ell \in \mathbb{R}^{(T-1)K}$ , let  $S(\ell) \geq 0$  denote a random variable with the generalized

chi-square distribution of weights equal to  $\ell$ , unit vector of degrees of freedom, zero vector of non-centrality parameters, and no constant or normal terms. Also, for any  $(x, \ell) \in \mathbb{R} \times \mathbb{R}^{(T-1)K}$ , let  $G(x, \ell)$  denote the CDF of  $S(\ell)$  evaluated at  $x \in \mathbb{R}$ . This function can be numerically computed with arbitrary accuracy by simulating its empirical distribution.

For the non-studentized CvM statistic, we propose

$$c_n^A(1 - \alpha) = \inf \{x \in \mathbb{R} : G(x, \hat{\lambda}_n) \geq 1 - \alpha\}, \quad (3.14)$$

where  $\hat{\lambda}_n$  denotes the eigenvalues of  $\hat{\Sigma}_Z$ , and the following hypothesis test:

$$\phi_n^A(\alpha) = 1 \{S_n > c_n^A(1 - \alpha)\}. \quad (3.15)$$

For the studentized CvM statistic, we propose the following hypothesis test:

$$\bar{\phi}_n^A(\alpha) = 1 \{\bar{S}_n > \bar{c}_n^A(1 - \alpha)\}, \quad (3.16)$$

where  $\bar{c}_n^A(1 - \alpha)$  equals the  $(1 - \alpha)$ -quantile of the (standard) chi-squared distribution with  $(T - 1)K$  degrees of freedom.

The next result shows that the hypothesis tests in (3.15) and (3.16) are asymptotically valid.

**Theorem 3.3.1.** *Let Assumption 9 and  $H_0$  in (3.3) hold, and let  $\alpha \in (0, 1)$ .*

- (a)  $\lim_{n \rightarrow \infty} E_P[\phi_n^A(\alpha)] \leq \alpha$ . *Furthermore, the inequality becomes an equality if  $\Sigma_Z \neq 0_{(T-1)K \times (T-1)K}$ .*
- (b) *Under Assumption 10,  $\lim_{n \rightarrow \infty} E_P[\bar{\phi}_n^A(\alpha)] = \alpha$ .*

### 3.3.2 Bootstrap

This section proposes a hypothesis test for (3.3) via the bootstrap. To this end, we repeatedly resample the data with replacement across units  $i = 1, \dots, n$  to construct a bootstrap sample, denoted by  $\{X_i^*\}_{i=1}^n$ . For each bootstrap sample, the bootstrap analog of the non-studentized and studentized CvM statistics are given by

$$S_n^* = (\hat{Z}^*)'(\hat{Z}^*) \quad \text{and} \quad \bar{S}_n^* = (\hat{Z}^*)'\hat{\Sigma}_Z^-(\hat{Z}^*). \quad (3.17)$$

where, for all  $(t, k) \in \{1, \dots, T-1\} \times \{1, \dots, K\}$ ,

$$\hat{Z}_{(t-1)K+k}^* \equiv \sqrt{\hat{P}(u_k)} \frac{1}{\sqrt{n}} \sum_{i=1}^n ([1(X_{i,t}^* \leq u_k) - \hat{F}_t(u_k)] - [1(X_{i,t+1}^* \leq u_k) - \hat{F}_{t+1}(u_k)]). \quad (3.18)$$

**Remark 3.3.1.** *One could also define  $\bar{S}_n^*$  in (3.17) with  $\hat{\Sigma}_Z$  replaced by its bootstrap analog. Our main text omits this option for brevity, but we include it in our Monte Carlo simulations.*

By repeating the bootstrap sampling sufficiently many times, we can approximate the conditional distributions  $P(S_n^* \leq x | \mathbf{X}_n)$  and  $P(\bar{S}_n^* \leq x | \mathbf{X}_n)$  with arbitrary accuracy.

For the non-studentized CvM statistic, we propose

$$c_n^B(1 - \alpha) = \inf \{x : P(S_n^* \leq x | \mathbf{X}_n) \geq 1 - \alpha\}, \quad (3.19)$$

and the following hypothesis test:

$$\phi_n^B(\alpha) = 1 \{S_n > c_n^B(1 - \alpha)\}. \quad (3.20)$$

For the studentized CvM statistic, we propose

$$c_n^B(1 - \alpha) = \inf \{x : P(\bar{S}_n^* \leq x | \mathbf{X}_n) \geq 1 - \alpha\}, \quad (3.21)$$

and the following hypothesis test:

$$\bar{\phi}_n^B(\alpha) = 1 \{\bar{S}_n > \bar{c}_n^B(1 - \alpha)\}. \quad (3.22)$$

The next result shows that the hypothesis tests in (3.19) and (3.21) are asymptotically valid.

**Theorem 3.3.2.** *Let Assumption 9 and  $H_0$  in (3.3) hold, and let  $\alpha \in (0, 1)$ .*

(a)  $\lim_{n \rightarrow \infty} E_P[\phi_n^B(\alpha)] \leq \alpha$ . *Furthermore, the inequality becomes an equality if  $\Sigma_Z \neq$*

$$0_{(T-1)K \times (T-1)K}.$$

(b) *Under Assumption 10,  $\lim_{n \rightarrow \infty} E_P[\bar{\phi}_n^B(\alpha)] = \alpha$ .*

### 3.3.3 Permutations

In this section, we propose a hypothesis test for (3.3) by random permutations of the data. Our permutations are motivated by the marginal homogeneity hypothesis in (3.1), and they consist of randomly permuting the time index  $t = 1, \dots, T$  for each unit  $i = 1, \dots, n$ .

These tests require the following notation. Let  $\mathcal{M}$  denote the set of all permutations of the indices  $\{1, \dots, T\}$ , and  $\mathcal{M}^n$  is defined as the set of all possible permutations of the time indices over  $n$  observations. A typical element of  $\mathcal{M}^n$  is given by  $\pi^n \equiv \{\{\pi_i(1), \dots, \pi_i(T)\}\}_{i=1}^n$ , where  $\pi_i(t)$  denotes the permuted time index of the observation  $X_{i,t}$ , and  $\{\pi_i(1), \dots, \pi_i(T)\}$  is an arbitrary time permutation that belongs to the set  $\mathcal{M}$ . In other words, the permuted version of the data  $\mathbf{X}_n = \{\{X_{i,t}\}_{t=1}^T\}_{i=1}^n$  can be written as  $\mathbf{X}_n^\pi \equiv \{\{X_{i,\pi_i(t)}\}_{t=1}^T\}_{i=1}^n$ .

For each permutation  $\pi^n \in \mathcal{M}^n$ , the permutation analog of the non-studentized and studentized CvM statistics are given by

$$S_n^\pi = (\hat{Z}^\pi)'(\hat{Z}^\pi) \quad \text{and} \quad \bar{S}_n^\pi = (\hat{Z}^\pi)' \hat{\Sigma}_{Z^\pi}^- (\hat{Z}^\pi), \quad (3.23)$$

where, for all  $(t, k) \in \{1, \dots, T-1\} \times \{1, \dots, K\}$ ,

$$\hat{Z}_{(t-1)K+k}^\pi \equiv \sqrt{n \hat{P}^\pi(x)} (\hat{F}_t^\pi(u_k) - \hat{F}_{t+1}^\pi(u_k)) : (t, k) \in \{1, \dots, T-1\} \times \{1, \dots, K\}, \quad (3.24)$$

and, for all  $(t, k), (\tilde{t}, \tilde{k}) \in \{1, \dots, T-1\} \times \{1, \dots, K\}$ ,

$$\begin{aligned} \hat{\Sigma}_{Z^\pi}[(t-1)K+k, (\tilde{t}-1)K+\tilde{k}] &\equiv \sqrt{\hat{P}^\pi(u_k) \hat{P}^\pi(u_{\tilde{k}})} \\ &\times \frac{1}{n} \sum_{i=1}^n \left[ \begin{aligned} &\left( \mathbf{1}(X_{i,\pi_i(t)} \leq u_k) - \hat{F}_t^\pi(u_k) - \mathbf{1}(X_{i,\pi_i(t+1)} \leq u_k) + \hat{F}_{t+1}^\pi(u_k) \right) \times \\ &\left( \mathbf{1}(X_{i,\pi_i(\tilde{t})} \leq u_{\tilde{k}}) - \hat{F}_{\tilde{t}}^\pi(u_{\tilde{k}}) - \mathbf{1}(X_{i,\pi_i(\tilde{t}+1)} \leq u_{\tilde{k}}) + \hat{F}_{\tilde{t}+1}^\pi(u_{\tilde{k}}) \right) \end{aligned} \right] \end{aligned} \quad (3.25)$$

where  $\hat{P}^\pi(u_k) \equiv \frac{1}{nT} \sum_{t=1}^T \sum_{i=1}^n \mathbf{1}(u_{k-1} < X_{i,\pi_i(t)} \leq u_k)$  and  $\hat{F}_t^\pi(u_k) \equiv \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_{i,\pi_i(t)} \leq u_k)$ .

These permutation test statistics can be used to construct permutation-based tests along the lines of Lehmann and Romano (2022, Section 17.2.1). For the non-studentized

CvM statistic, we propose a critical value  $c_n^\pi(1 - \alpha)$ , which is the  $(1 - \alpha)$ -quantile over all possible permutations of the non-studentized CvM statistics (denoted by  $\{S_n^\pi : \pi^n \in \mathcal{M}^n\}$ ).

The corresponding hypothesis test is defined as:

$$\phi_n^\pi(\alpha) = 1 \{S_n > c_n^\pi(1 - \alpha)\}. \quad (3.26)$$

For the studentized CvM statistic, we propose the analogous object but for the studentized CvM statistic. That is,  $\bar{c}_n^\pi(1 - \alpha)$  equals to the  $(1 - \alpha)$ -quantile over all the possible permutations for the studentized CvM statistics (given by  $\{\bar{S}_n^\pi : \pi^n \in \mathcal{M}^n\}$ ). The corresponding hypothesis test is given by:

$$\bar{\phi}_n^\pi(\alpha) = 1 \{\bar{S}_n > \bar{c}_n^\pi(1 - \alpha)\}. \quad (3.27)$$

**Remark 3.3.2.** *The tests in (3.26) and (3.27) are the non-random versions of the standard permutation tests described in Lehmann and Romano (2022, Section 17.2.1). The key difference between the non-random and the random versions lies in the handling of ties between the test statistic and the critical value: the non-random version does not reject, while the random version rejects with a specific probability. While being more conservative, the non-random version is preferred because it involves a simpler decision rule similar to the one used in previous tests.*

The next result studies the asymptotic validity of the hypothesis tests in (3.26) and (3.27).

**Theorem 3.3.3.** *Let Assumption 9 and  $H_0$  in (3.3) hold, and let  $\alpha \in (0, 1)$ .*

- (a) *For  $T = 2$ ,  $\lim_{n \rightarrow \infty} E_P[\phi_n^\pi(\alpha)] \leq \alpha$ . Furthermore, the inequality becomes an equality if  $\Sigma_Z \neq 0_{(T-1)K \times (T-1)K}$ .*
- (b) *When  $T > 2$ , the non-studentized permutation test in (3.26) is invalid. That is, it is possible to have  $\liminf_{n \rightarrow \infty} E_P[\phi_n^\pi(\alpha)] > \alpha$  for some distribution  $P$ .*
- (c) *Under Assumption 10,  $\lim_{n \rightarrow \infty} E_P[\bar{\phi}_n^\pi(\alpha)] = \alpha$ .*

We first describe the result for the non-studentized test in (3.26). Part (a) shows that the non-studentized test is asymptotically valid for  $T = 2$ . This result aligns with the analysis

in Gaigall (2020) and Ditzhaus and Gaigall (2022). Part (b) reveals that the previous result fails for  $T > 2$ . This finding is apparently new in the literature and empirically relevant, as many economic applications involve panel data with more than two periods. To gain intuition about (a) and (b), it is useful to compare the variance-covariance of the  $i$ th observation, i.e.,  $\{X_{i,t}\}_{t=1}^T$ , before and after permutations. If the variance-covariance remains the same, then the permutation test can be shown to be asymptotically valid. See the proof of Theorem 3.3.3 for a formal justification. When  $T = 2$ , the  $i$ th observation is  $(X_{i,1}, X_{i,2})$  before the permutation and a mixture of  $(X_{i,1}, X_{i,2})$  and  $(X_{i,2}, X_{i,1})$  after the permutation. Under  $H_0$  in (3.3), these two random vectors share the same variance-covariance matrix. When  $T > 2$ , this equivalence breaks down. For example, when  $T = 3$ , the  $i$ th observation is  $(X_{i,1}, X_{i,2}, X_{i,3})$  before the permutation and a mixture of  $(X_{i,1}, X_{i,2}, X_{i,3})$ ,  $(X_{i,1}, X_{i,3}, X_{i,2})$ ,  $(X_{i,2}, X_{i,1}, X_{i,3})$ ,  $(X_{i,2}, X_{i,3}, X_{i,1})$ ,  $(X_{i,3}, X_{i,1}, X_{i,2})$ , and  $(X_{i,3}, X_{i,2}, X_{i,1})$  after the permutation. The variance-covariance matrix of the former and the latter differ even under  $H_0$  in (3.3).

Finally, part (c) of Theorem 3.3.3 shows that the studentized test in (3.27) is asymptotically valid for any  $T \geq 2$ . This result is in line with several studies in the literature that prove the asymptotic validity of permutation tests for suitably normalized test statistics, e.g., Chung and Romano (2013), Chung and Olivares (2021), DiCiccio and Romano (2017), and Janssen (1997). This result is also related to the earlier discussion regarding variance-covariance matrices of the  $i$ th observation before and after permutations. Once the data is studentized, the variance-covariance matrix of the  $i$ th observation is asymptotically invariant to permutations. By the previous paragraph's logic, the permutation test is asymptotically valid for the studentized test statistic.

An important advantage of the permutation tests over the ones described in previous subsections lies in their finite-sample validity under an important class of distributions that satisfy  $H_0$  in (3.3). To explain this clearly, let  $\Omega_{\text{TE}}$  denote the set of distributions that satisfy time exchangeability, i.e., for any  $i = 1, \dots, n$ ,  $\{X_{i,t}\}_{t=1}^T$  has the same distribution as  $\{X_{i,\pi(t)}\}_{t=1}^T$  for each  $\pi \in \mathcal{M}$ . The next result describes the finite-sample validity of our

test under suitable conditions.

**Theorem 3.3.4.** *Let Assumption 9 hold. Then,*

(a)  $P \in \Omega_{\text{TE}}$  implies that  $P$  satisfies  $H_0$  in (3.3). However, the converse does not hold.

(b) For each  $P \in \Omega_{\text{TE}}$ , our permutation tests are finite-sample valid, i.e.,

$$E_P[\phi_n^\pi(\alpha)] \leq \alpha \quad \text{and} \quad E_P[\bar{\phi}_n^\pi(\alpha)] \leq \alpha. \quad (3.28)$$

Theorem 3.3.4 implies that our permutation tests are finite-sample valid under suitable conditions. Part (a) says that a time-exchangeable distribution satisfies the null hypothesis  $H_0$  in (3.3) under our maintained Assumption 9. Under such distribution, the permutation tests provide finite-sample size control. We stress that size control is not exact (i.e., the inequalities in (3.28) might be strict) only because we are using a non-randomized permutation test; see Remark 3.3.2. That is, if we replaced our permutation test with its random version, these would enjoy exact size control (i.e., both inequalities in (3.28) would hold with equality).

The combination of Theorems 3.3.3 and 3.3.4 justifies the use of studentized permutation test in (3.27) (and also the non-studentized one in (3.26) when  $T = 2$ ). Theorem 3.3.3 indicates that this test is asymptotically exact, and Theorem 3.3.4 shows that it is finite-sample valid for an important class of distributions in  $\Omega_{\text{TE}}$ . As already mentioned, the finite-sample validity makes the permutation test an exceptionally attractive inference method compared to those discussed in previous subsections.

**Remark 3.3.3.** *It is worth noting that our class of time permutations differs from the class of “all permutations” that uniformly permute both time periods and units. The latter has been used in the previous literature, such as Friedrich et al. (2017). Naturally, our time permutations form a strict subset of the class of “all permutations”. On the flip side, under our maintained Assumption 9, the class of distributions that satisfy exchangeability over “all permutations” is more restrictive than the set of distributions that satisfy time exchangeability only; see Lemma B.1.2. In other words, if we used a permutation test based*

on “all permutations”, we could only demonstrate Theorem 3.3.4 for a substantially smaller class of distributions than  $\Omega_{\text{TE}}$ .

### 3.4 Power Analysis

This section briefly describes the power properties of the various hypothesis-testing procedures considered in this paper. Given the results in Section 3.3, we restrict attention to the hypothesis tests that are asymptotically valid under Assumptions 910:

- The asymptotic approximation-based test, both non-studentized and studentized.
- The bootstrap-based test, both non-studentized and studentized.
- The studentized permutation-based test.
- The non-studentized permutation-based test for  $T = 2$ .

As explained in Section 3.2.1, our CvM test statistic is defined to detect differences in marginal CDFs at any point in  $\mathcal{U}_K$ . We thus focus our power analysis on the following subset of  $H_1$ :

$$\bar{H}_1 : F_t(u) \neq F_r(u) \quad \text{for some } t, r = 1, \dots, T \text{ and } u \in \mathcal{U}_K. \quad (3.29)$$

For all fixed hypotheses in  $\bar{H}_1$ , it is not hard to see that the CvM test statistic in (3.5) and (3.10) diverges. At the same time, one can establish that the critical values described throughout this paper remain bounded in probability. For this reason, all asymptotically valid tests are consistent against any fixed hypothesis in  $\bar{H}_1$ .

We now compare the local power properties of these tests. We consider local alternative hypotheses under  $\bar{H}_1$ , which are sequences of DGPs whose marginal CDFs  $\{F_{n,t} : t = 1, \dots, T\}$  satisfy  $\{\sqrt{n}(F_{n,t}(u_k) - F_{n,r}(u_k)) : t, r = 1, \dots, T, u \in \mathcal{U}_K\} \rightarrow c \in \mathbb{R}^{T^2K}$  with  $c'c > 0$ . Under these sequences of distributions, we can repeat the arguments used to prove Theorem 3.2.1 to establish the asymptotic distribution of the CvM test statistic in (3.5) and (3.10). It is not hard to see that these become the non-central versions of the asymptotic distributions in (3.12) and (3.13) under  $H_0$ . Moreover, under these local alternatives, the critical values considered throughout this paper can be shown not to change their asymptotic behavior. As a corollary, the asymptotically valid tests based on the non-

studentized statistics share the same local power properties, and the same holds for the asymptotically valid tests based on the studentized statistics.

### 3.5 Monte Carlo Simulations

This section investigates the finite-sample performance of the various tests for marginal homogeneity tests proposed in this paper. To this end, we repeatedly simulate independent panel datasets  $\mathbf{X}_n = \{\{X_{i,t}\}_{t=1}^T\}_{i=1}^n$  where, for each  $i = 1, \dots, n$  and  $t = 1, \dots, T$ ,

$$X_{i,t} = (\varepsilon_{i,t} + \xi_i)\alpha_t$$

with

$$\begin{pmatrix} \{\varepsilon_{i,t}\}_{t=1}^T \\ \xi_i \end{pmatrix} \sim N\left(0_{(T+1) \times 1}, \begin{bmatrix} \Xi & 0_{T \times 1} \\ 0_{1 \times T} & 1 \end{bmatrix}\right),$$

where  $\{\alpha_t\}_{t=1}^T$  is a sequence of constants and  $\Xi$  is a constant positive-definite matrix. By definition,  $\xi_i$  is a random effect and  $\{\varepsilon_{i,t}\}_{t=1}^T$  are transient shocks with variance-covariate matrix  $\Xi$ . We consider two specifications for  $\Xi$ :

- WN:  $\Xi = I_{T \times T}$ , i.e.,  $\{\varepsilon_{i,t}\}_{t=1}^T$  is a white noise process with zero mean and unit variance. The distribution of  $\mathbf{X}_n$  is time-exchangeable under this specification.
- AR(1): for all  $t_1, t_2 = 1, \dots, T$  with  $t_1 \neq t_2$ ,  $\Xi[t_1, t_1] = 1$  and  $\Xi[t_1, t_2] = (-0.9)^{|t_1 - t_2|}$ , i.e.,  $\{\varepsilon_{i,t}\}_{t=1}^T$  is an AR(1) process zero mean, variance one, and correlation coefficient  $-0.9$ . In this case, the distribution of  $\mathbf{X}_n$  is not time-exchangeable.

We consider two options for  $\{\alpha_t\}_{t=1}^T$ , which determines whether the data satisfies marginal homogeneity or not. For simulations under  $H_0$  in (3.3), we use  $\alpha_t = 1$  for  $t = 1, \dots, T$ . For simulations under  $H_1$  in (3.3), we set  $\alpha_t = \sqrt{1 + (t-1)/2}$  for  $t = 1, \dots, T$ .

To compute the CvM statistics we set  $\mathcal{U}_K$  equal to the 1/6, 2/6, 3/6, 4/6, and 5/6 empirical quantiles of  $\{\{X_{i,t}\}_{t=1}^T\}_{i=1}^n$  (thus,  $K = 5$ ). For each dataset, we implemented the following tests with a significance level of  $\alpha = 5\%$ :

- AA: Asymptotic approximation tests in Section 3.3.1, non-studentized as in (3.15) and studentized as in (3.16).

- BS: Bootstrap tests in Section 3.3.2, non-studentized as in (3.20) and studentized as in (3.22).
- BS2: Studentized bootstrap tests based on a bootstrapped covariance estimator, as described in Remark 3.3.1.
- PT: Permutation tests in Section 3.3.3, non-studentized as in (3.26) and studentized as in (3.27).
- PT2: Permutation tests based on the class of “all permutations” (i.e., time periods and units), as described in Remark 3.3.3. These can be non-studentized and studentized.

We consider simulations with  $n \in \{30, 60, 120, 240, 480\}$  units and  $T \in \{2, 3\}$  periods. The results shown in the tables are obtained from  $S = 5,000$  independent panel data draws based on the design described.

Table 3.1 describes the rejection rates of the various tests under marginal homogeneity, i.e.,  $H_0$  in (3.3). When  $T = 2$ , all our proposed hypothesis tests (AA, BS, PT, studentized or not) provide exact size control when  $n$  is sufficiently large. These results are consistent with our asymptotic analysis. This conclusion also seems to apply to the BS2 and the studentized PT2 tests. On the other hand, the non-studentized PT2 test does not control size. This is consistent with our discussion in Lemma B.1.2, where we argue that the permutation class used to implement our PT tests is qualitatively different from the one used for the PT2 tests. Our simulations also allow us to examine how our asymptotically valid methods perform when  $n$  is relatively small. In this respect, one interesting finding is that the non-studentized versions of the AA and BS tests outperform their corresponding studentized ones when  $n$  is small. On the other hand, the PT test performs equally well with and without studentization for small  $n$ . The results for  $T = 3$  are qualitatively similar to those for  $T = 2$ , except for the PT test. For  $T > 2$ , our formal results show that the studentized PT test is asymptotically valid, but the non-studentized PT test is not. This is clearly evidenced in our simulations with AR(1) shocks, where the non-studentized PT test exhibits rejection rates close to 13%. However, if data is time-exchangeable as with the WN shocks, then we observe that PT test is finite-sample valid regardless of studentization

even if  $T = 3$ .

Table 3.1: Empirical rejection rates (in %) under  $H_0$  in (3.3) for  $\alpha = 5\%$  based on  $S = 5,000$  i.i.d. panel datasets

$T$	$\Xi$	Test type	Critical value	$n = 30$	$n = 60$	$n = 120$	$n = 240$	$n = 480$
$T = 2$	WN	Non-stud.	AA	5.48	5.54	5.08	4.76	5.16
			BS	5.60	5.36	5.14	4.78	5.06
			PT	4.54	4.94	4.82	4.62	5.08
			PT2	1.02	0.80	0.76	0.96	1.00
		Studentized	AA	14.34	8.54	6.94	6.08	4.96
			BS	14.24	8.56	6.98	6.00	4.86
			BS2	2.56	4.66	5.02	5.30	4.54
			PT	5.06	5.08	5.20	5.20	4.44
	AR(1)	Non-stud.	PT2	4.50	5.02	5.16	5.28	4.44
			AA	5.12	6.02	5.12	5.24	5.22
			BS	5.36	5.90	5.12	5.22	5.10
			PT	4.16	5.32	4.88	5.00	5.06
		Studentized	PT2	3.46	4.58	4.32	4.62	4.54
			AA	14.00	8.78	6.88	6.30	5.58
			BS	14.18	8.96	6.98	6.40	5.52
			BS2	3.58	4.86	5.16	5.38	5.08
$T = 3$	WN	Non-stud.	PT	5.12	5.36	5.16	5.34	5.00
			PT2	5.10	5.32	5.40	5.42	5.16
			AA	5.68	4.80	5.46	5.16	5.56
			BS	5.84	4.74	5.38	5.22	5.52
		Studentized	PT	5.30	4.66	5.38	5.00	5.38
			PT2	0.88	0.72	0.60	0.94	1.10
			AA	35.82	16.74	9.88	7.04	5.80
			BS	36.00	16.78	9.90	6.98	5.84
	AR(1)	Non-stud.	BS2	0.70	3.28	4.66	4.82	4.92
			PT	5.02	5.08	4.98	4.84	4.88
			PT2	3.78	4.52	4.82	4.66	4.86
			AA	6.22	6.18	5.76	5.20	5.64
		Studentized	BS	6.08	6.24	5.74	5.42	5.38
			PT	12.70	13.80	13.88	13.68	12.96
			PT2	6.14	7.30	6.94	6.44	6.84
			AA	33.44	17.06	10.38	7.26	5.96
Studentized	BS	33.94	16.92	10.12	7.22	5.80		
	BS2	1.78	3.14	4.26	4.80	4.98		
	PT	4.62	5.26	5.08	5.14	5.12		
	PT2	3.36	4.54	4.84	4.86	5.02		

Table 3.2 explores the performance of the same test for data configurations that do not satisfy marginal homogeneity. To make the comparison fair, we focus on asymptotically valid inference methods. For  $T = 2$ , this includes studentized and non-studentized versions of the AA, BS, and PT tests, and studentized BS2. Our results indicate that studentized

tests are considerably more powerful than their corresponding non-studentized versions. The main difference in the case of  $T = 3$ , is that we now have to eliminate the non-studentized PT test. With this exception, the results with  $T = 3$  are qualitatively similar to those for  $T = 2$ .

Table 3.2: Empirical rejection rates (in %) under  $H_1$  in (3.3) based on  $\alpha_t = \sqrt{1 + (t - 1)/2}$  for  $t = 1, \dots, T$ , with  $\alpha = 5\%$  based on  $S = 5,000$  i.i.d. panel datasets

$T$	$\Xi$	Test type	Critical value	$n = 30$	$n = 60$	$n = 120$	$n = 240$	$n = 480$
$T = 2$	WN	Non-stud.	AA	8.04	11.06	16.22	32.12	65.34
			BS	8.06	11.00	16.34	32.12	65.08
			PT	6.66	10.20	15.78	31.68	64.98
			PT2	1.30	2.14	3.44	7.64	28.86
		Studentized	AA	19.88	18.74	27.12	46.40	78.66
			BS	20.04	18.70	27.14	46.50	78.86
			BS2	4.92	10.96	22.38	43.90	77.68
			PT	8.32	11.96	22.60	43.94	77.72
	AR(1)	Non-stud.	PT2	7.66	11.74	22.42	43.82	77.84
			AA	6.50	7.64	9.54	18.14	39.56
			BS	6.58	7.76	9.22	18.14	39.70
			PT	5.36	7.08	8.90	17.80	39.40
		Studentized	PT2	4.48	6.10	7.90	15.68	36.88
			AA	18.82	17.60	22.52	40.96	72.34
			BS	18.96	17.76	22.50	40.90	72.24
			BS2	5.16	10.96	18.26	38.30	71.30
$T = 3$	WN	Non-stud.	PT	7.48	11.48	18.48	38.36	71.06
			PT2	7.46	11.52	18.52	38.44	71.22
			AA	6.98	8.50	12.70	27.70	72.42
			BS	7.16	8.54	12.64	27.78	72.56
		Studentized	PT	6.32	8.14	12.18	27.26	72.44
			PT2	1.00	1.34	1.58	3.78	18.86
			AA	51.14	46.10	62.68	89.28	99.66
			BS	51.22	45.94	62.56	89.20	99.68
	AR(1)	Non-stud.	BS2	1.86	16.14	47.42	85.40	99.60
			PT	11.30	21.28	48.82	85.74	99.60
			PT2	9.20	19.60	47.88	85.52	99.54
			AA	6.38	7.16	7.56	10.44	21.30
		Studentized	BS	6.34	7.10	7.64	10.58	21.28
			PT	14.26	16.64	19.74	28.10	65.38
			PT2	6.42	8.42	9.38	12.98	28.18
			AA	62.84	72.54	94.60	99.94	100.00
Studentized	BS	63.34	72.50	94.60	99.94	100.00		
	BS2	6.24	33.40	86.92	99.88	100.00		
	PT	16.30	45.90	89.34	99.88	100.00		
	PT2	13.00	43.08	88.52	99.88	100.00		

Tables 3.1 and 3.2 offer interesting conclusions regarding the finite sample properties of

the asymptotically valid tests. When the sample size is relatively small, the non-studentized tests appear to provide better size control than their studentized counterparts, though at the expense of lower power. As the sample size increases, the studentized tests improve their size control. As a result, the studentized tests seem to be a better option for larger sample sizes: they offer adequate size control and relatively higher power compared to their non-studentized versions.

### **3.6 Empirical Application**

This section applies our marginal homogeneity tests to the state variable in the dynamic discrete choice game in Igami and Yang (2016). In this paper, the authors develop and estimate a dynamic entry model oligopoly game among Canada’s five main hamburger chains: A&W, Burger King, Harvey’s, McDonald’s, and Wendy’s. They use yearly data from 400 geographical markets<sup>1</sup> located in seven major Canadian cities between 1970 and 2004, i.e.,  $n = 400$  and  $T = 35$ . For each market-year pair  $(i, t)$ , they observe the number of stores for each chain, population, and income.

The state variable used in Igami and Yang (2016) is a discrete categorical variable whose value represents the number of stores of each chain, population, and income of a given market-year pair. We now briefly explain its construction, and defer to their Igami and Yang (2016, Section 4.1) for details. The paper restricts the number of stores per chain to three, and divides population and income into quartiles.

For each period  $t = 1, \dots, T = 35$  and market  $i = 1, \dots, n = 400$ ,  $X_{i,t}$  is uniquely determined by the number of stores (up to three) for each chain  $j = 1, 2, 3, 4, 5$ ,  $N_{i,t,j} \in \{0, 1, 2, 3\}$ , the population quartile  $P_{i,t} \in \{1, 2, 3, 4\}$ , and the income quartile  $I_{i,t} \in \{1, 2, 3, 4\}$ . So,  $X_{i,t} = 1$  indicates that  $N_{i,t,j} = 0$  for all  $j = 1, 2, 3, 4, 5$ ,  $P_{i,t} = 1$ , and  $I_{i,t} = 1$ ,  $X_{i,t} = 2$  indicates that  $N_{i,t,j} = 0$  for all  $j = 1, 2, 3, 4, 5$ ,  $P_{i,t} = 1$ , and  $I_{i,t} = 2$ , and so on, until  $X_{i,t} = 16,384$  indicates that  $N_{i,t,j} = 3$  for all  $j = 1, 2, 3, 4, 5$ ,  $P_{i,t} = 4$ , and  $I_{i,t} = 4$ . While  $X_{i,t}$  could take up to  $4^7 = 16,384$  possible values, it only takes 467 distinct values in the

---

<sup>1</sup> The paper defines a market as a cluster of stores located within a 0.5-mile radius at any point of their sample period. Markets in downtown areas are omitted as these experience a different nature of competition.

entire dataset.

The data spans an extensive period in which the Canadian fast-food industry grew considerably. As Igami and Yang (2016, Section 3.2) reports, the average number of shops per market was less than 0.5 during 1970's to approximately 1.8 in the early 2000's. We now provide further evidence about the evolution of this industry. Figure 3.1 shows the average number of stores per market over time desegregated by chain. This figure shows that the average number of stores per market has increased across all chains, with the most significant increase observed for McDonald's compared to the other chains. Figure 3.2 shows the average number of competitors per market over time. This figure reveals that the frequency of empty markets decreased steadily between 1970 and 2000, while the frequencies of monopoly, duopoly, and triopoly or more steadily increased over the same period. In contrast, during 2000-2004, the frequency of each market type has remained relatively stable. This evidence suggests that the Canadian fast-food industry has been evolving between 1970 and the early 2000's, and may have reached a steady state during the last years of the sample. The hypothesis tests developed in this paper can be used to evaluate whether the state distribution is homogeneous over any of the periods in the sample.

As explained in Section 3.1, the marginal homogeneity of the state variable can be a source for efficiency gains in the estimation of dynamic discrete games. With this motivation in mind, we now apply our marginal homogeneity tests to our panel data of the state variable. Given the large number of values that the variable takes, Assumption 10 does not hold over the sample period. For this reason, we only consider non-studentized tests. We implement our tests for two subsets of periods. Since we consider panel data with  $T > 2$ , the non-studentized asymptotic approximation and bootstrap tests are valid, but the permutation test is not. Table 3.3 presents the results of our hypothesis tests for two subsets of sample periods. First, we consider a subset of our data every five years, i.e., 1970, 1975, 1980, 1985, 1990, 1995, and 2000. In this case, our tests strongly reject the hypothesis of marginal homogeneity. This result is expected, as it is consistent with the

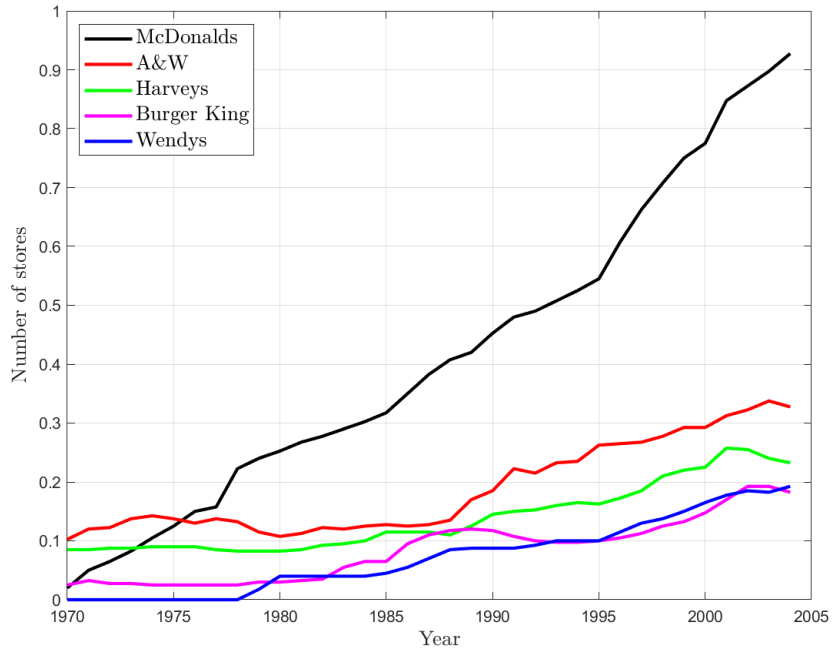


FIGURE 3.1: Average Number of Stores per Market over Time for Each Chain

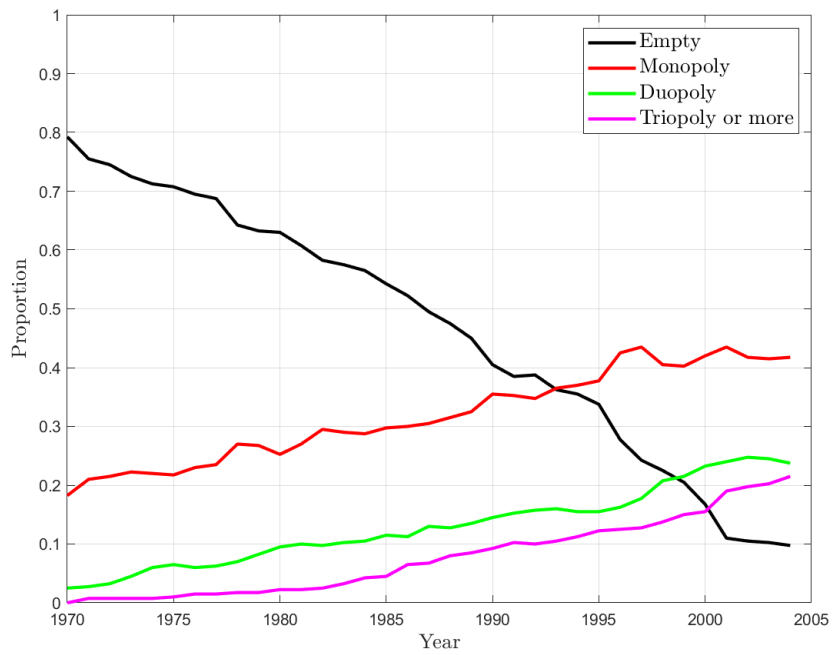


FIGURE 3.2: Average Type of Market over Time

informal discussion in the previous paragraph regarding the growth of the Canadian fast-food industry between 1970 and 2000. Second, we repeat the analysis for the last four years in the sample period, i.e., 2001, 2002, 2003, and 2004. In this case, our tests do not reject the hypothesis of marginal homogeneity. These results suggest that the Canadian fast-food industry may be in a steady state in the latter part of the sample period.

Table 3.3: Marginal Homogeneity Tests for Igami and Yang (2016)

Note: Marginal homogeneity tests are applied to Igami and Yang (2016) with  $\alpha = 5\%$ . The Pre-2000 sample refers to the subsample with  $t \in \{1970, 75, 80, 90, 95, 00\}$ . The Post-2000 sample refers to the subsample with  $t \in \{2001, 02, 03, 04\}$ .

Sample	Test type	Test statistic	Critical Value with $\alpha = 5\%$		
			Asy. approx.	Bootstrap	Permutation
Pre-2000	Non-stud.	10.76	0.88	0.82	2.35
Post-2000	Non-stud.	0.08	0.13	0.13	0.23

### 3.7 Conclusions

This paper proposes hypothesis tests to evaluate whether panel data satisfies the hypothesis of marginal homogeneity. As we argue in the paper, marginal homogeneity is a relevant property in economic settings such as dynamic discrete games.

Our asymptotic framework for panel data considers a diverging number of units  $n$  and a fixed number of periods  $T$ . We implement our tests by comparing a studentized or non-studentized  $T$ -sample version of the Cramér-von Mises statistic with a suitable critical value. Relative to the non-studentized case, the asymptotic analysis of the studentized statistics requires an additional assumption: the variance-covariate matrix used in the studentization must be non-singular. It is relevant to note that this condition can fail in practice. In fact, it failed in our empirical application.

We investigate three methods to construct the critical value: asymptotic approximations, the bootstrap, and time permutations. We prove that the asymptotic approximation and bootstrap tests are asymptotically valid, regardless of whether we use studentized or non-studentized test statistics. The permutation test based on a non-studentized statistic

is asymptotically exact when  $T = 2$ , but is asymptotically invalid when  $T > 2$ . In contrast, the permutation test based on a studentized statistic is always asymptotically exact. Finally, under a time-exchangeability assumption, the permutation test is valid in finite samples, both with and without studentization.

We also study the power of the various methods. The asymptotically valid tests we consider are consistent and have non-trivial asymptotic power under suitable local alternatives. Moreover, the asymptotically valid tests based on the non-studentized statistics share the same local power properties, and the same holds for the asymptotically valid tests based on the studentized statistics.

Our Monte Carlo simulations investigate the finite sample behavior of our tests. The non-studentized tests exhibit better finite-sample size control than their studentized counterparts, though this comes at the cost of lower power. Finally, we apply our test to the state variable of the dynamic oligopoly model of the Canadian fast-food industry in Igami and Yang (2016). Our findings suggest that the industry evolved between 1970 and 2000, and appears to have reached a steady state since then.

## 4. A General Approach to Relaxing Unconfoundedness

This essay proposes a general approach to relax unconfoundedness that incorporates several previous relaxations as special cases, and we move beyond averages by providing sharp bounds for a large class of parameters, which were previously unknown even under special cases. This is a joint work with Matthew Masten and Alexandre Poirier.

### 4.1 Introduction

A large literature studies the identification and estimation of treatment effects when a binary treatment is randomly assigned conditional on covariates. This assumption is called unconfoundedness, conditionally independent treatment assignment, or ignorability, among other terms. With observational data it is often considered very strong, however, so a corresponding literature has developed to relax this assumption. These papers use a variety of different classes of relaxations of unconfoundedness. That is, there are different ways of formalizing the idea that treatment is “almost” randomly assigned, given the covariates. This variation raises a question: How do these different relaxations compare to each other? This question is important because empirical researchers are often concerned that the number of robustness checks they must consider is constantly growing; if some of these checks are related, however, then that relationship can potentially be used to simplify the overall analysis. Moreover, mathematically related analyses do not necessarily provide “independent” evidence of robustness, a second motivation for better understanding the relationships between different relaxations of an assumption.

With that aim, this paper makes two main contributions. First, we define a general class of relaxations, which includes several previous approaches as special cases. Second, we derive closed form, analytical identification results for treatment effects under this general class of relaxations. This paper therefore unifies several disparate identification results in the literature. In doing so, we also provide a variety of new identification results, because we study an extensive list of parameters, including quantile treatment effects (QTEs) and the distribution of treatment effects (DTEs), whereas most existing papers focus solely on

average-type treatment effects. These new results were previously unknown even for the specific types of relaxations that have been considered before. We give a precise discussion of how our results compare to the previous literature in the next subsection.

In section 4.2 we set up the baseline treatment effects model and define the target parameters we study. We define our general class of relaxations at the start of section 4.3. We show how this class relates to previous relaxations in sections 4.3.1 and 4.3.2. In section 4.4 we derive general analytical identification results for marginal cdfs of potential outcomes and monotonic functionals of those cdfs. We apply those results in section 4.5 to obtain analytical bounds on various treatment effect parameters. We conclude in section 4.6.

## Related literature

A vast literature studies unconfoundedness; we do not attempt a comprehensive review here. Instead we discuss the most closely related prior work. Nonparametric relaxations of unconfoundedness were pioneered by Paul Rosenbaum’s work (see his 2002 or 2017 books for a survey, for example). His work focuses on sensitivity analysis within the context of finite sample randomization inference (c.f., chapter 5 of Imbens and Rubin 2015). Much of the subsequent literature has instead focused on large population level identification analysis. In particular, inspired by Rosenbaum’s approach, Tan (2006) proposed the *marginal sensitivity model* (MSM), a specific nonparametric relaxation of unconfoundedness (which we review in section 4.3). Given this relaxation, Tan showed that bounds on parameters of interest can be characterized as the solutions to optimization problems with infinitely many constraints, but did not provide any formal results, proofs, or closed form expressions for these bounds. Zhao et al. (2019) derived non-sharp bounds on the average potential outcome  $\mathbb{E}(Y_x)$  and the average treatment effect (ATE) under the MSM, but also did not derive closed form expressions for these bounds. Dorn and Guo (2023) strengthened that result by deriving sharp bounds on  $\mathbb{E}(Y_x)$ , ATE, and the average effect of treatment on the treated (ATT) under the MSM, but again without closed form expressions. Dorn et al. (2024) subsequently

refined that result by obtaining closed form expressions for sharp bounds on  $\mathbb{E}(Y_x)$  and ATE under the MSM, in addition to developing the concept of double-validity and double sharpness. Tan (2024) gives alternative sharp bound expressions for the ATE under the MSM. Kallus and Zhou (2018) studied policy learning under the MSM, which is related to identification of the average weighted welfare (what they call the “policy value”), but they do not derive population bounds on this parameter.

This existing literature on the MSM largely focuses on average potential outcomes  $\mathbb{E}(Y_x)$  or the ATE. Our paper provides the first sharp bounds on a wide variety of target parameters under the MSM, including the quantile treatment effect (QTE), the quantile treatment effect on the treated (QTT), the distribution of treatment effects (DTE), and the average weighted welfare (AWW). Moreover, for many of the parameters we study, our bounds are closed form. The existence of closed form expressions simplifies the construction of estimation and inference procedures, and also allows us to analytically examine how the bounds depend on the distribution of the observed data, and thus which features of the data lead results to be robust.

Masten and Poirier (2016, 2018a) proposed an alternative relaxation of unconfoundedness called *conditional c-dependence*, and derived closed form sharp bounds on a variety of treatment effect parameters under this relaxation, including  $\mathbb{E}(Y_x)$ , ATE, ATT, the QTE, and the DTE (in Masten and Poirier 2020). In the current paper we extend these identification results to a class of parameters that also includes the average weighted welfare (AWW), weighted average treatment effects, and to quantiles of the distribution of conditional average treatment effects (QCATE). Those earlier papers also restricted attention to continuous or binary outcomes whereas our new results apply for any distribution of the outcome, including mixed continuous-discrete distributions. We also show how the conditional *c-dependence* relaxation is related to the marginal sensitivity model.

While our general class of relaxations includes several previously proposed relaxations of unconfoundedness, there are alternative relaxations where it is not yet clear if they can be accommodated by our class. This includes Bonvini and Kennedy (2022), who derive closed-

form sharp bounds on ATE under a mixture-style relaxation, and Huang and Pimentel (2025), who derive closed-form non-sharp bounds on ATT under an assumption about how much unobserved variables can affect the variance of odds ratios similar to those that arise in the MSM; in appendix A.6 they derive sharp but non-closed form bounds on the ATT under the same relaxation. It also includes Ding and VanderWeele (2016) and VanderWeele and Ding (2017), who derive closed-form non-sharp bounds on the causal relative risk under assumptions about relative risks involving latent confounders; Sjölander (2024) derives closed-form sharp bounds under the same relaxation.

There are several related papers that provide general methods for deriving bounds. Dorn and Yap (2024) show how to derive analytical expressions for sharp bounds on parameters that can be written as certain weighted averages of outcomes under a restriction on a generalized likelihood ratio. Like us, they show that their class of relaxations includes several previous relaxations (such as the MSM of Tan 2006 and conditional  $c$ -dependence of Masten and Poirier 2018a). Whereas we only study relaxations of unconfoundedness, they also show how to use their results to do sensitivity analysis for instrumental variables and regression discontinuity designs. Their analysis of unconfoundedness, however, focuses on average potential outcomes and ATE, whereas we also study parameters like the QTE and DTE. A large literature in econometrics has studied how to derive identified sets for a variety of parameters under a variety of assumptions when all observed variables are discretely distributed; see, for example, Torgovitsky (2019) and Gu et al. (2024), and the references therein. Duarte (2024) uses similar ideas to numerically compute identified sets for a variety of sensitivity analyses when all variables are discretely distributed. Rambachan et al. (2023) provide a variety of general sensitivity analyses for binary outcomes. In contrast to this literature, we obtain analytical sharp bounds without any restriction on the distribution of the outcome variable, which allows the outcome to be continuously distributed or even mixed continuous-discretely distributed.

Several prior papers also discuss the relationship between various relaxations of unconfoundedness. Masten and Poirier (2023) discuss mean independence, quantile independence

assumptions, and a weaker version of quantile independence that they call  $\mathcal{U}$ -independence. Their focus is on interpreting relaxations in terms of treatment selection models, rather than providing identification results for a broad class of relaxations. Zhao et al. (2019, section 7.2) discuss the relationship between the MSM and Rosenbaum’s sensitivity model. For binary outcomes, Rambachan et al. (2023, appendix D) relate their relaxation to the MSM, Rosenbaum’s sensitivity model, conditional  $c$ -dependence, and an approach called Tukey’s factorization.

**Notation**

For random a variable  $A$  and a random vector  $B$ , we let  $F_{A|B}(a | b) := \mathbb{P}(A \leq a | B = b)$  denote the conditional cdf. For  $\tau \in (0, 1)$ , we let  $Q_{A|B}(\tau | b) := \inf\{a \in \mathbb{R} : F_{A|B}(a | b) \geq \tau\}$  denote the left-inverse of this cdf, that is, its conditional quantile function.

**4.2 Setup and Target Parameters**

We are interested in the causal impact of a binary treatment variable  $X \in \{0, 1\}$  on an outcome variable  $Y$ . Let  $(Y_1, Y_0)$  be potential outcomes under treatment and no treatment respectively. Denote the realized outcome by

$$Y = XY_1 + (1 - X)Y_0.$$

Let  $W$  be a vector of covariates with support  $\text{supp}(W) \subseteq \mathbb{R}^{dw}$ . We use  $p_{x|w}$  to denote  $\mathbb{P}(X = x | W = w)$ .  $p_{1|w}$  thus denotes the propensity score. We assume realizations of  $(Y, X, W)$  are observed by the researcher. Our identification analysis abstracts from sampling uncertainty and assumes the joint distribution of  $(Y, X, W)$  is known.

Throughout the paper we maintain the following assumption, which is a strict overlap assumption. It is also sometimes called strict positivity.

**Assumption 11.** *There exists  $\epsilon > 0$  such that  $p_{1|w} \in [\epsilon, 1 - \epsilon]$  for all  $w \in \text{supp}(W)$ .*

With observational data, a commonly imposed assumption is unconfoundedness. It is also called selection on observables, ignorability, or conditional independence. This assumption states that potential outcomes are independent of treatment given covariates  $W$ . This

conditional independence is either imposed jointly on both potential outcomes, or on each potential outcome separately:

$$Y_x \perp\!\!\!\perp X \mid W \text{ for } x \in \{0, 1\} \quad \text{or} \quad (Y_1, Y_0) \perp\!\!\!\perp X \mid W. \quad (4.1)$$

Under Assumption 11 and either version of unconfoundedness given in equation (4.1), it is well known that the pair of distribution functions  $(F_{Y_1|W}, F_{Y_0|W})$  is point-identified via

$$F_{Y_x|W}(y \mid w) = \mathbb{P}(Y \leq y \mid X = x, W = w)$$

for  $x \in \{0, 1\}$ . This point-identification implies that many parameters that summarize aspects of the distribution of  $(Y_1, Y_0, X, W)$  are point-identified. Specifically, parameters that can be expressed as functionals of  $F_{Y_1|W}$ ,  $F_{Y_0|W}$ , and the known distribution of observables  $(Y, X, W)$ , are point-identified. For example, we can point-identify the Conditional Average Treatment Effect (CATE) as defined by  $\text{CATE}(w) := \mathbb{E}[Y_1 - Y_0 \mid W = w]$  because it can be written as

$$\mathbb{E}[Y_1 - Y_0 \mid W = w] = \int y_1 dF_{Y_1|W}(y_1 \mid w) - \int y_0 dF_{Y_0|W}(y_0 \mid w).$$

However, parameters that depend on other aspects of the distribution of potential outcomes may only be partially identified. For example, consider the distribution function of  $Y_1 - Y_0$ , the unit-level treatment effect:

$$F_{Y_1 - Y_0}(z) = \int \mathbb{1}(y_1 - y_0 \leq z) dF_{Y_1, Y_0}(y_1, y_0).$$

This parameter depends on the structure of the dependence between the two potential outcomes, which is unknown from either version of the unconfoundedness assumption. As discussed in Fan and Park (2010),  $F_{Y_1 - Y_0}(z)$  is partially identified and sharp bounds can be recovered in terms of the distribution of  $(Y, X, W)$ .

To help classify treatment effect parameters, consider the decomposition

$$F_{Y_1, Y_0|X, W}(y_1, y_0 \mid x, w) = C_{1,0|X, W}\left(F_{Y_1|X, W}(y_1 \mid x, w), F_{Y_0|X, W}(y_0 \mid x, w) \mid x, w\right),$$

where  $C_{1,0|X, W}(\cdot, \cdot \mid x, w)$  is a copula that characterizes the dependence between  $Y_1$  and  $Y_0$  conditional on  $(X, W) = (x, w)$ . By Sklar's Theorem (Sklar 1959), such a copula exists.

Given that  $F_{Y,X,W}$  is known, we consider treatment effect parameters that can be written as a function of  $(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W})$ . We denote these parameters through the functional

$$\theta(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W}), \quad (4.2)$$

and we give several examples below. The dependence of  $\theta$  on some of its arguments is suppressed if the functional is constant with respect to them. We will sometimes denote these parameters as functionals of  $(F_{Y_1|W}, F_{Y_0|W}, F_{Y,X,W})$  rather than  $(F_{Y_1|X,W}, F_{Y_0|X,W}, F_{Y,X,W})$ . These two formulations are equivalent due to the relationship

$$F_{Y_x|W}(y | w) = F_{Y|X,W}(y | x, w)p_{x|w} + F_{Y_x|X,W}(y | 1 - x, w)p_{1-x|w} \quad (4.3)$$

which holds for all  $(y, x, w) \in \mathbb{R} \times \{0, 1\} \times \text{supp}(W)$ .

Next we consider eleven example target parameters. Our results give sharp bounds for all eleven parameters under our general class of assumptions, including the marginal sensitivity model as a special case. For many of these parameters we also obtain analytical, closed form expressions for the bound functions.

**Example 4.2.1** (Conditional Quantile Treatment Effect). *For quantile index  $\tau \in (0, 1)$  and covariate value  $w \in \text{supp}(W)$ , the conditional quantile treatment effect (CQTE) can be written as*

$$CQTE(\tau | w) := Q_{Y_1|W}(\tau | w) - Q_{Y_0|W}(\tau | w) = \theta_{CQTE}(F_{Y_1|W}, F_{Y_0|W}; \tau, w)$$

where  $\theta_{CQTE}(F_{Y_1|W}, F_{Y_0|W}; \tau, w) := F_{Y_1|W}^{-1}(\tau | w) - F_{Y_0|W}^{-1}(\tau | w)$  where  $F_{Y_x|W}^{-1}(\tau | w) = \inf\{y \in \mathbb{R} : F_{Y_x|W}(y | w) \geq \tau\}$  is the left-inverse of  $F_{Y_x|W}(\cdot | w)$ , or its conditional quantile function.

**Example 4.2.2** (Conditional Average Treatment Effect). *For covariate value  $w \in \text{supp}(W)$ , the conditional average treatment effect (CATE) can be written as*

$$CATE(w) := \mathbb{E}[Y_1 - Y_0 | W = w] = \theta_{CATE}(F_{Y_1|W}, F_{Y_0|W}; w)$$

where  $\theta_{CATE}(F_{Y_1|W}, F_{Y_0|W}; w) := \int y_1 dF_{Y_1|W}(y_1 | w) - \int y_0 dF_{Y_0|W}(y_0 | w)$ .

**Example 4.2.3** (Average Treatment Effect). Denote the average treatment effect (ATE) as

$$ATE := \mathbb{E}[Y_1 - Y_0] = \theta_{ATE}(F_{Y_1|W}, F_{Y_0|W}, F_W)$$

where  $\theta_{ATE}(F_{Y_1|W}, F_{Y_0|W}, F_W) := \int (\int y_1 dF_{Y_1|W}(y_1 | w) - \int y_0 dF_{Y_0|W}(y_0 | w)) dF_W(w)$ . We can also consider weighted average treatment effects of the kind

$$WATE(\omega) := \mathbb{E}[\omega(W)(Y_1 - Y_0)]$$

for an identified function  $\omega : \text{supp}(W) \rightarrow \mathbb{R}_{\geq 0}$ . The ATE is a special case where  $\omega(w) = 1$ .

**Example 4.2.4** (Average Treatment Effect on the Treated). Denote the average treatment effect on the treated (ATT) as

$$ATT := \mathbb{E}[Y_1 - Y_0 | X = 1] = \theta_{ATT}(F_{Y_0|X,W}, F_{Y,X,W})$$

where  $\theta_{ATT}(F_{Y_0|X,W}, F_{Y,X,W}) := \mathbb{E}[Y | X = 1] - \int \int y_0 dF_{Y_0|X,W}(y_0 | 1, w) dF_{W|X}(w | 1)$ .

**Example 4.2.5** (Quantile Treatment Effect). For  $\tau \in (0, 1)$ , denote the quantile treatment effect (QTE) as

$$QTE(\tau) := Q_{Y_1}(\tau) - Q_{Y_0}(\tau) = \theta_{QTE}(F_{Y_1|W}, F_{Y_0|W}, F_W; \tau)$$

where  $\theta_{QTE}(F_{Y_1|W}, F_{Y_0|W}, F_W; \tau) := F_{Y_1}^{-1}(\tau) - F_{Y_0}^{-1}(\tau)$  and  $F_{Y_x}(\cdot) = \int F_{Y_x|W}(\cdot | w) dF_W(w)$  for  $x \in \{0, 1\}$ .

**Example 4.2.6** (Quantile Treatment Effect on the Treated). For  $\tau \in (0, 1)$ , denote the quantile treatment effect on the treated (QTT) as

$$QTT(\tau) := Q_{Y_1|X}(\tau | 1) - Q_{Y_0|X}(\tau | 1) = \theta_{QTT}(F_{Y_0|X,W}, F_{Y,X,W}; \tau)$$

where  $\theta_{QTT}(F_{Y_0|X,W}, F_{Y,X,W}; \tau) := Q_{Y|X}(\tau | 1) - F_{Y_0|X}^{-1}(\tau | 1)$  with

$$F_{Y_0|X}(\cdot | 1) = \int F_{Y_0|X,W}(\cdot | 1, w) dF_{W|X}(w | 1).$$

**Example 4.2.7** (CATE Distribution). For  $\tau \in (0, 1)$ , denote the quantile of the CATE (QCATE) as

$$QCATE(\tau) := F_{CATE(W)}^{-1}(\tau) = \theta_{QCATE}(F_{Y_1|W}, F_{Y_0|W}, F_W; \tau)$$

where  $\theta_{Q_{CATE}(F_{Y_1|W}, F_{Y_0|W}, F_W; \tau)} := F_{CATE(W)}^{-1}(\tau)$  with

$$F_{CATE(W)}(z) = \int \mathbb{1}(\theta_{CATE}(F_{Y_1|W}, F_{Y_0|W}; w) \leq z) dF_W(w).$$

This parameter is motivated by the sorted effects studied in Chernozhukov et al. (2018).

**Example 4.2.8** (Average Weighted Welfare). For a weight (or assignment) function  $\omega : \text{supp}(W) \rightarrow [0, 1]$ , denote the average weighted welfare (AWW) as

$$AWW(\omega) := \mathbb{E}[\omega(W)Y_1 + (1 - \omega(W))Y_0] = \theta_{AWW}(F_{Y_1|W}, F_{Y_0|W}, F_W; \omega)$$

where

$$\begin{aligned} \theta_{AWW}(F_{Y_1|W}, F_{Y_0|W}, F_W; \omega) := \\ \int \left( \omega(w) \int y_1 dF_{Y_1|W}(y_1 | w) + (1 - \omega(w)) \int y_0 dF_{Y_0|W}(y_0 | w) \right) dF_W(w). \end{aligned}$$

Kallus and Zhou (2018) called the AWW the policy value.

**Example 4.2.9** (Joint Distribution Function). For  $(y_1, y_0) \in \mathbb{R}^2$ , the joint cumulative distribution function (cdf) of  $(Y_1, Y_0)$  is

$$F_{Y_1, Y_0}(y_1, y_0) := \mathbb{P}(Y_1 \leq y_1, Y_0 \leq y_0) = \theta_{CDF}(F_{Y_1|X, W}, F_{Y_0|X, W}, C_{1,0|X, W}, F_{Y, X, W}; y_1, y_0)$$

where

$$\begin{aligned} \theta_{CDF}(F_{Y_1|X, W}, F_{Y_0|X, W}, C_{1,0|X, W}, F_{Y, X, W}; y_1, y_0) := \\ \sum_{x \in \{0, 1\}} \int C_{1,0|X, W}(F_{Y_1|X, W}(y_1 | x, w), F_{Y_0|X, W}(y_0 | x, w) | x, w) p_{x|w} dF_W(w). \end{aligned}$$

**Example 4.2.10** (Distribution of Treatment Effects). For  $z \in \mathbb{R}$ , the cumulative distribution function for the unit level treatment effect  $Y_1 - Y_0$  (called the DTE) is

$$DTE(z) := F_{Y_1 - Y_0}(z) = \mathbb{P}(Y_1 - Y_0 \leq z) = \theta_{DTE}(F_{Y_1|X, W}, F_{Y_0|X, W}, C_{1,0|X, W}, F_{Y, X, W}; z)$$

where

$$\begin{aligned} \theta_{DTE}(F_{Y_1|X, W}, F_{Y_0|X, W}, C_{1,0|X, W}, F_{Y, X, W}; z) := \\ \sum_{x \in \{0, 1\}} \int \left( \int_{\{y_1 - y_0 \leq z\}} dC_{1,0|X, W}(F_{Y_1|X, W}(y_1 | x, w), F_{Y_0|X, W}(y_0 | x, w) | x, w) \right) p_{x|w} dF_W(w). \end{aligned}$$

**Example 4.2.11** (Quantiles of Treatment Effects). For  $\tau \in (0, 1)$ , the quantiles of the distribution function of treatment effect (QDTE)  $Y_1 - Y_0$  is

$$Q_{Y_1 - Y_0}(\tau) := \inf\{z \in \mathbb{R} : F_{Y_1 - Y_0}(z) \geq \tau\} = \theta_{QDTE}(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W}; \tau)$$

where

$$\begin{aligned} & \theta_{QDTE}(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W}; \tau) \\ & := \inf\{z \in \mathbb{R} : \theta_{DTE}(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W}; z) \geq \tau\}. \end{aligned}$$

The parameters in examples 4.2.1–4.2.8 only depend on the distribution of potential outcomes through their marginal distributions given  $(X, W)$ , while the parameters in examples 4.2.9–4.2.11 also depend on their copulas. Under overlap and unconfoundedness, the parameters in 4.2.1–4.2.8 are all point-identified. The parameters 4.2.9–4.2.11 are partially identified under overlap and unconfoundedness since the conditional copulas  $C_{1,0|X,W}$  are not identified from the joint distribution of  $(Y, X, W)$ . In other words, these parameters depend on the type of dependence between  $Y_1$  and  $Y_0$ , and unconfoundedness does not reveal any information about this dependence. For example, Fan and Park (2010) show the identified set for  $F_{Y_1 - Y_0}(z)$ , the DTE, is an interval and they provide a closed-form expression for its lower and upper bounds.

However, if unconfoundedness fails, all these parameters will be partially identified. The identified set for parameters that are partially identified under unconfoundedness becomes larger under failures of unconfoundedness.

### **4.3 Relaxing Unconfoundedness**

We now consider relaxations of the unconfoundedness assumption. We will consider two related, general relaxations of unconfoundedness that encompass several disparate relaxations that were studied in the literature. We begin by considering a class of assumptions on the probabilities of treatment when conditioning on covariates  $W$  and one of the potential outcomes.

**Assumption 12** (Marginal  $c$ -dependence). *Let  $(\underline{c}(w, \eta), \bar{c}(w, \eta))$  satisfy  $0 < \underline{c}(w, \eta) \leq p_{1|w} \leq \bar{c}(w, \eta) < 1$  for all  $w \in \text{supp}(W)$ . The potential outcomes satisfy marginal  $c$ -dependence if, for  $x \in \{0, 1\}$ ,*

$$p_x(Y_x, w) := \mathbb{P}(X = 1 \mid Y_x, W = w) \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$$

*almost surely conditional on  $W = w$  for all  $w \in \text{supp}(W)$ .*

This assumption restricts the manner in which potential outcomes affect the treatment probability  $p_x(y, w)$ , which we call a *latent propensity score*. We will use  $c$ -dependence assumptions to conduct sensitivity analyses for unconfoundedness. Here the sensitivity parameters are  $\underline{c}(w, \eta)$  and  $\bar{c}(w, \eta)$ , which we refer to as bound functions. Like the notation in Rambachan et al. (2023), we let  $\eta$  be a possibly infinite-dimensional nuisance parameter that is point-identified from the distribution  $F_{Y, X, W}$ . The bound functions are also allowed to depend on the covariate value  $w$ . In principle, we can also allow the bounds to differ across  $x \in \{0, 1\}$ , but we do not include an  $x$  subscript for simplicity. The specification of  $\underline{c}(w, \eta)$  and  $\bar{c}(w, \eta)$  is left implicit, which allows them to be functions of low-dimensional or scalar sensitivity parameters. For example, the marginal sensitivity model of Tan (2006), which depends on a single sensitivity parameter, can be viewed as a special case of marginal  $c$ -dependence. We show this in section 4.3.1.

We can also see that setting  $\underline{c}(w, \eta) = \bar{c}(w, \eta) = p_{1|w}$  yields unconfoundedness as a special case of marginal  $c$ -dependence, while letting  $(\underline{c}(w, \eta), \bar{c}(w, \eta))$  approach  $(0, 1)$  implies that no restrictions on the dependence between  $X$  and  $Y_x$  (given covariates) are imposed. Note that we restrict the propensity score  $p_{1|w}$  to lie within our specified bounds for  $p_x(Y_x, w)$ . If the propensity score were outside these bounds, then the assumption would be misspecified because, by the law of iterated expectations,  $p_{1|w} = \mathbb{E}[\mathbb{P}(X = 1 \mid Y_x, W = w) \mid W = w] \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$ .

We also consider a closely related class of assumptions that restricts the probability of treatment given both potential outcomes.

**Assumption 13** (Joint  $c$ -dependence). *Let  $(\underline{c}(w, \eta), \bar{c}(w, \eta))$  satisfy  $0 < \underline{c}(w, \eta) \leq p_{1|w} \leq \bar{c}(w, \eta) < 1$  for all  $w \in \text{supp}(W)$ . The potential outcomes satisfy joint  $c$ -dependence if*

$$p(Y_1, Y_0, w) := \mathbb{P}(X = 1 \mid Y_1, Y_0, W = w) \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$$

*almost surely conditional on  $W = w$  for all  $w \in \text{supp}(W)$ .*

Joint  $c$ -dependence with bound functions  $\underline{c}(w, \eta)$  and  $\bar{c}(w, \eta)$  implies marginal  $c$  dependence with the same bound functions. This is due to the law of iterated expectations.

**Lemma 4.3.1.** *Let Assumption 13 hold for  $(\underline{c}(w, \eta), \bar{c}(w, \eta))$ . Then, Assumption 12 holds for  $(\underline{c}(w, \eta), \bar{c}(w, \eta))$ .*

We next show that several unconfoundedness relaxations from recent related literature can be viewed as special cases of either marginal or joint  $c$ -dependence.

### 4.3.1 The marginal sensitivity Model

Tan (2006) proposed the Marginal Sensitivity Model (MSM), which restricts the odds ratio between propensity scores and treatment probabilities that also condition on the potential outcome  $Y_x$ , for  $x = 0, 1$ . For  $x \in \{0, 1\}$ , let

$$R_x(Y_x, W) := \frac{\mathbb{P}(X = 1 \mid Y_x, W)}{\mathbb{P}(X = 0 \mid Y_x, W)} \bigg/ \frac{\mathbb{P}(X = 1 \mid W)}{\mathbb{P}(X = 0 \mid W)}$$

denote this odds ratio. When  $Y_x$  is continuously distributed with respect to the Lebesgue measure, this ratio can also be expressed as ratios of conditional densities of  $Y_x \mid X = 1, W$  and  $Y_x \mid X = 0, W$ .

Tan (2006)'s MSM posits known bounds for these odds ratios.

**Definition 4.3.1** (Marginal Sensitivity Model (MSM)). *Let  $\Lambda \in [1, +\infty)$  be known. The potential outcomes satisfy the Marginal Sensitivity Model if*

$$R_x(Y_x, w) \in [\Lambda^{-1}, \Lambda] \text{ for } x \in \{0, 1\}$$

*almost surely conditional on  $W = w$  for all  $w \in \text{supp}(W)$ .*

$\Lambda$  is a scalar sensitivity parameter. Setting  $\Lambda = 1$  is equivalent to assuming unconfoundedness, and increasing  $\Lambda$  allows for more dependence of latent propensity scores on potential outcomes. Variants of Tan (2006)'s MSM whose odds ratios condition on both potential outcomes have also been considered. Similarly, these odds ratios may instead condition on an abstract confounder  $U$  rather than potential outcomes. See, for example, Dorn and Guo (2023) and Dorn et al. (2024) for recent examples of these two variants. To distinguish it from the case where one conditions on the potential outcomes one at a time, we call the version that conditions on both potential outcomes the *Joint Sensitivity Model* (JSM).

**Definition 4.3.2** (Joint Sensitivity Model (JSM)). *Let  $\Lambda \in [1, +\infty)$  be known. The potential outcomes satisfy the Joint Sensitivity Model if*

$$R(Y_1, Y_0, w) \in [\Lambda^{-1}, \Lambda] \text{ for } x \in \{0, 1\}$$

*almost surely conditional on  $W = w$  for all  $w \in \text{supp}(W)$ , where*

$$R(Y_1, Y_0, W) := \frac{\mathbb{P}(X = 1 \mid Y_1, Y_0, W)}{\mathbb{P}(X = 0 \mid Y_1, Y_0, W)} \bigg/ \frac{\mathbb{P}(X = 1 \mid W)}{\mathbb{P}(X = 0 \mid W)}.$$

We now state generalizations of the MSM and JSM that allow their odds ratios to have arbitrary bounds, as opposed to bounds that have product equal to 1. We also allow their bounds to depend on covariates or nuisance parameters. We will continue distinguishing between *marginal* sensitivity models, which condition on one potential outcome at a time, and *joint* sensitivity models, which condition on both potential outcomes.

**Definition 4.3.3** (Generalized Sensitivity Models). *Let  $\underline{\Lambda}(w, \eta) \in (0, 1]$  and  $\bar{\Lambda}(w, \eta) \in [1, +\infty)$  for all  $w \in \text{supp}(W)$  where  $\underline{\Lambda}(w, \eta)$  and  $\bar{\Lambda}(w, \eta)$  are known. The potential outcomes satisfy the Generalized Marginal Sensitivity Model (GMSM) if*

$$R_x(Y_x, w) \in [\underline{\Lambda}(w, \eta), \bar{\Lambda}(w, \eta)] \text{ for } x \in \{0, 1\}$$

*almost surely conditional on  $W = w$  for all  $w \in \text{supp}(W)$ . They satisfy the Generalized Joint Sensitivity Model (GJSM) if*

$$R(Y_1, Y_0, w) \in [\underline{\Lambda}(w, \eta), \bar{\Lambda}(w, \eta)]$$

almost surely conditional on  $W = w$  for all  $w \in \text{supp}(W)$ .

We can see that the MSM is a special case of the GMSM by setting  $[\underline{\Lambda}(w, \eta), \bar{\Lambda}(w, \eta)] = [\Lambda^{-1}, \Lambda]$ . Similarly, the JSM is a special case of the GJSM.

The GMSM is equivalent to marginal  $c$ -dependence because, for each bound function pair  $[\underline{c}(w, \eta), \bar{c}(w, \eta)]$  under marginal  $c$ -dependence, there exists exactly one corresponding bound function pair  $[\underline{\Lambda}(w, \eta), \bar{\Lambda}(w, \eta)]$  under the GMSM. The same link exists between joint  $c$ -dependence and the GJSM. We show this in the following proposition.

**Proposition 4.3.1** (Equivalence of Sensitivity Models).

1. Let marginal (joint)  $c$ -dependence hold with bound functions  $[\underline{c}(w, \eta), \bar{c}(w, \eta)]$ . Then the GMSM (GJSM) holds with bound functions

$$[\underline{\Lambda}(w, \eta), \bar{\Lambda}(w, \eta)] = \left[ \frac{\underline{c}(w, \eta) p_{0|w}}{1 - \underline{c}(w, \eta) p_{1|w}}, \frac{\bar{c}(w, \eta) p_{0|w}}{1 - \bar{c}(w, \eta) p_{1|w}} \right]. \quad (4.4)$$

2. Let the GMSM (GJSM) hold with bound functions  $[\underline{\Lambda}(w, \eta), \bar{\Lambda}(w, \eta)]$ . Then marginal (joint)  $c$ -dependence holds with bound functions

$$[\underline{c}(w, \eta), \bar{c}(w, \eta)] = \left[ \frac{p_{1|w} \underline{\Lambda}(w, \eta)}{p_{0|w} + p_{1|w} \underline{\Lambda}(w, \eta)}, \frac{p_{1|w} \bar{\Lambda}(w, \eta)}{p_{0|w} + p_{1|w} \bar{\Lambda}(w, \eta)} \right]. \quad (4.5)$$

This proposition shows that the marginal  $c$ -dependence is equivalent to the generalized marginal sensitivity model. Similarly, joint  $c$ -dependence is equivalent to the generalized marginal sensitivity model.

### 4.3.2 Conditional $c$ -dependence

Masten and Poirier (2018a) studied a relaxation of unconfoundedness they called *conditional  $c$ -dependence*, which assumed symmetric bounds on the latent propensity score.

**Definition 4.3.4** (Conditional  $c$ -dependence). Let  $c \in [0, 1]$  be a known scalar sensitivity parameter. The potential outcomes satisfy conditional  $c$ -dependence if

$$p_x(Y_x, w) := \mathbb{P}(X = 1 \mid Y_x, W = w) \in [p_{1|w} - c, p_{1|w} + c]$$

almost surely conditional on  $W = w$  for all  $w \in \text{supp}(W)$ .

This is a special case of marginal  $c$ -dependence where the bounds equal

$$\underline{c}(w, \eta) = p_{1|w} - c \quad \text{and} \quad \bar{c}(w, \eta) = p_{1|w} + c.$$

Here the nuisance parameter is  $p_{1|(\cdot)}$ , the propensity score function. Unconfoundedness is obtained by setting  $c = 0$ , while the no-assumption bounds are obtained for  $c$ 's equal to or larger than  $\sup_{w \in \text{supp}(W)} \max\{p_{1|w}, p_{0|w}\}$ . Masten and Poirier (2018a) provided closed-form expressions for sharp bounds on the CQTE, CATE, ATE, QTE, and ATT when potential outcomes are continuously distributed or binary. Masten et al. (2024) describe flexible parametric estimators of these bounds and provide nonstandard inference methods.

## 4.4 General Identification Results

Next we derive sharp bounds on a large class of target parameters under the relaxations described in section 4.3. We will study a class of parameters that includes all eleven examples in section 4.2 as special cases. Specifically, we will compute these parameters' sharp bounds, or their identified set, under marginal and joint  $c$ -dependence, which are equivalent to the GSM and GJSM, respectively.

### 4.4.1 Bounds on marginal distributions

Before studying our general class of treatment effect parameters, we first consider bounds on the distribution functions of each potential outcome, given covariates. These cdfs are building blocks for these parameters and, as we will see, analytical bounds on these cdfs will directly map into analytical bounds on these parameters.

Specifically, we begin by analyzing the conditional cdf of the potential outcome  $Y_x$  given the covariate value  $w$ ,  $F_{Y_x|W}(y | w) := \mathbb{P}(Y_x \leq y | W = w)$ . Under either marginal or joint  $c$ -dependence, we can show that this cdf is bounded above and below by two cdfs which form an envelope for  $F_{Y_x|W}(y | w)$  for all values of  $(y, w) \in \mathbb{R} \times \text{supp}(W)$ .

Define the following functions:

$$\begin{aligned}
\underline{F}_{Y_1|W}(y | w) &= \max \left\{ F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\underline{c}(w, \eta)}, \frac{\underline{c}(w, \eta) - p_{1|w}}{\underline{c}(w, \eta)} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\underline{c}(w, \eta)} \right\} \\
\overline{F}_{Y_1|W}(y | w) &= \min \left\{ F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\underline{c}(w, \eta)}, \frac{\overline{c}(w, \eta) - p_{1|w}}{\overline{c}(w, \eta)} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\overline{c}(w, \eta)} \right\} \\
\underline{F}_{Y_0|W}(y | w) &= \max \left\{ F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1 - \underline{c}(w, \eta)}, \frac{p_{1|w} - \overline{c}(w, \eta)}{1 - \overline{c}(w, \eta)} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1 - \overline{c}(w, \eta)} \right\} \\
\overline{F}_{Y_0|W}(y | w) &= \min \left\{ F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1 - \overline{c}(w, \eta)}, \frac{p_{1|w} - \underline{c}(w, \eta)}{1 - \underline{c}(w, \eta)} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1 - \underline{c}(w, \eta)} \right\}.
\end{aligned}$$

Viewed as functions of  $y$ , these four functions are cdfs since they are nondecreasing, right-continuous, and their limits as  $y \rightarrow -\infty, +\infty$  equal 0 and 1, respectively. We show these four cdfs form bounds for  $F_{Y_x|W}$  under marginal or joint  $c$ -dependence.

**Lemma 4.4.1.** *Let Assumption 11 hold. Let either Assumption 12 or 13 hold. Then, for all  $(y, w) \in \mathbb{R} \times \text{supp}(W)$ ,*

$$\mathbb{P}(Y_1 \leq y | W = w) \in \left[ \underline{F}_{Y_1|W}(y | w), \overline{F}_{Y_1|W}(y | w) \right]$$

and

$$\mathbb{P}(Y_0 \leq y | W = w) \in \left[ \underline{F}_{Y_0|W}(y | w), \overline{F}_{Y_0|W}(y | w) \right].$$

We note a few properties of these bounds. The bounds for  $F_{Y_x|W}(y | w)$  collapse to a point if either  $\underline{c}(w, \eta) = p_{1|w}$  or  $\overline{c}(w, \eta) = p_{1|w}$ . Also note that  $F_{Y|X,W}(y | x, w)$  always lies within the bounds for  $F_{Y_x|W}(y | w)$ . This is because  $c$ -dependence never rules out unconfoundedness, and unconfoundedness implies that the distribution of  $Y$  given  $(X, W) = (x, w)$  equals that of  $Y_x$  given  $W = w$ .

These cdf bounds also yield cdf bounds under the GMSM or GJSM since they are equivalent to  $c$ -dependence as shown in Proposition 4.3.1. These bounds are also valid for the standard MSM or JSM, as they are special cases of marginal or joint  $c$ -dependence.

We now show these cdf bounds are sharp, or that they cannot be improved upon. This is the case under marginal or joint  $c$ -dependence. The cdf of  $Y_x | W = w$  can also lie in the interior of these bounds, as we show that any convex linear combination of the upper and lower cdf bounds can be attained.

Before establishing this, let  $\mathcal{C}$  denote the set of all bivariate copulas and let

$$\mathcal{C}_{1,0|X,W} = \left\{ \{C_{1,0|x,w}\}_{x \in \{0,1\}, w \in \text{supp}(W)} \text{ such that } C_{1,0|x,w} \in \mathcal{C} \right\}$$

denote the collection of all bivariate copulas across all treatment and covariate values  $(x, w) \in \{0, 1\} \times \text{supp}(W)$ . We also say that a distribution function for  $(Y_1, Y_0) | X, W$  is compatible with the observed distribution  $F_{Y,X,W}$  if

$$F_{Y_1|X,W}(\cdot | 1, w) = F_{Y|X,W}(\cdot | 1, w) \quad \text{and} \quad F_{Y_0|X,W}(\cdot | 0, w) = F_{Y|X,W}(\cdot | 0, w) \quad (4.6)$$

for all  $w \in \text{supp}(W)$ .

We now define the identified set for the distribution of  $(Y_1, Y_0) | X, W$  from the observable distribution  $F_{Y,X,W}$  under  $c$ -dependence. This set consists of all conditional cdfs and copulas that imply a distribution for  $(Y_1, Y_0) | X, W$  that is both compatible with the data distribution  $F_{Y,X,W}$  and with a  $c$ -dependence condition.

**Definition 4.4.1** (Identified Set). *For a given distribution of the observables  $F_{Y,X,W}$  and bound functions  $c := (\underline{c}(w, \eta), \bar{c}(w, \eta))$ , the identified set for  $(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W})$  under marginal  $c$ -dependence is given by*

$$\begin{aligned} \mathcal{I}^{\text{marg}}(F_{Y,X,W}; c) := \\ \{(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}) : F_{Y_1, Y_0|X,W} = C_{1,0|X,W}(F_{Y_1|X,W}, F_{Y_0|X,W}) \\ \text{and } p_{1|(\cdot)} \text{ satisfy equation (4.6) and Assumption 12}\}. \end{aligned}$$

The identified set under joint  $c$ -dependence is given by

$$\begin{aligned} \mathcal{I}^{\text{joint}}(F_{Y,X,W}; c) := \\ \{(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}) : F_{Y_1, Y_0|X,W} = C_{1,0|X,W}(F_{Y_1|X,W}, F_{Y_0|X,W}) \\ \text{and } p_{1|(\cdot)} \text{ satisfy equation (4.6) and Assumption 13}\}. \end{aligned}$$

In our later derivations, we may refer to the identified set for  $(F_{Y_1|W}, F_{Y_0|W}, C_{1,0|X,W})$  instead, which we denote by  $\mathcal{I}_0^i(F_{Y,X,W}; c)$  for  $i \in \{\text{marg}, \text{joint}\}$ . Via equation (4.3), this set can be viewed as an affine transformation of the identified set for  $(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W})$ .

We now show some key properties of the cdfs and copulas in these identified sets. We begin with marginal  $c$ -dependence.

**Theorem 4.4.1.** *Let Assumptions 11 and 12 hold. For all  $(\varepsilon, \gamma) \in [0, 1]^2$  and for any  $C_{1,0|X,W} \in \mathcal{C}_{1,0|X,W}$ ,*

$$\left( \varepsilon \underline{F}_{Y_1|W} + (1 - \varepsilon) \overline{F}_{Y_1|W}, \gamma \underline{F}_{Y_0|W} + (1 - \gamma) \overline{F}_{Y_0|W}, C_{1,0|X,W} \right) \in \mathcal{I}_0^{marg}(F_{Y,X,W}; c).$$

This theorem shows that the four pairs of cdfs, i.e.,  $(\underline{F}_{Y_1|W}, \underline{F}_{Y_0|W})$ ,  $(\overline{F}_{Y_1|W}, \overline{F}_{Y_0|W})$ ,  $(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W})$ , and  $(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W})$ , are part of the identified set. This is obtained by varying  $(\varepsilon, \gamma)$  over  $\{(1, 1), (0, 1), (1, 0), (0, 0)\}$ . We show this by explicitly constructing latent propensity scores  $p_1(Y_1, w)$  and  $p_0(Y_0, w)$  that lie in  $[\underline{c}(w, \eta), \overline{c}(w, \eta)]$  almost surely under the implied distribution of  $(Y_1, Y_0) | W = w$ . These propensity scores have a switching structure where they equal the lower/upper bound for low values of  $Y_x$  and the upper/lower bound for large values of  $Y_x$ . For example, the propensity score  $p_1(Y_1, w)$  associated with cdf upper bound  $\overline{F}_{Y_1|W}$  equals

$$p_1(Y_1, w) := \begin{cases} \underline{c}(w, \eta) & \text{if } Y_1 < \overline{Q}_1 \\ \overline{A}_1 & \text{if } Y_1 = \overline{Q}_1 \\ \overline{c}(w, \eta) & \text{if } Y_1 > \overline{Q}_1 \end{cases}$$

where

$$\overline{Q}_1 := Q_{Y|X,W} \left( \frac{(\overline{c}(w, \eta) - p_{1|w}) \underline{c}(w, \eta)}{(\overline{c}(w, \eta) - \underline{c}(w, \eta)) p_{1|w}} \mid X = 1, W = w \right)$$

and

$$\overline{A}_1 := \frac{\mathbb{P}(Y = \overline{Q}_1, X = 1 \mid W = w)}{\overline{F}_{Y_1|W}(\overline{Q}_1 \mid w) - \overline{F}_{Y_1|W}(\overline{Q}_1 - \mid w)}.$$

Note that  $\overline{A}_1 \in [\underline{c}(w, \eta), \overline{c}(w, \eta)]$ . We denote  $\lim_{q \nearrow \overline{Q}_1} \overline{F}_{Y_1|W}(q \mid w)$  by  $\overline{F}_{Y_1|W}(\overline{Q}_1 - \mid w)$ .

The propensity scores associated with the cdf bounds  $\underline{F}_{Y_1|W}$ ,  $\overline{F}_{Y_0|W}$ , and  $\underline{F}_{Y_0|W}$  can all be found in Appendix C.1.

This switching structure of the latent propensity score was observed for conditional  $c$ -dependence by Masten and Poirier (2018a, pages 335–339), and for the MSM in Proposition 2 of Dorn and Guo (2023). Our sharpness proof implies that latent propensity score

$p_1(Y_1, w)$  satisfies an integral constraint, namely that  $\mathbb{E}[p_1(Y_1, W) \mid W = w] = \mathbb{P}(X = 1 \mid W = w)$ , in order to ensure it is compatible with the observed propensity score.

Theorem 4.4.1 also shows that any convex linear combinations of these four cdf pairs lies in the identified set. As a result, the identified set for  $F_{Y_x|W}(y \mid w)$  is the entire closed interval  $[\underline{F}_{Y_x|W}(y \mid w), \overline{F}_{Y_x|W}(y \mid w)]$ . Moreover, the identified set for the pair  $(F_{Y_1|W}(y \mid w), F_{Y_0|W}(y \mid w))$  is the Cartesian product of their individual identified sets, meaning that fixing or knowing the conditional distribution of one potential outcome does not affect the identified set of the distribution of the other potential outcome.

Finally, Theorem 4.4.1 proves that no conditional copulas are ruled out by marginal  $c$ -dependence. For example, marginal  $c$ -dependence allows  $Y_1$  and  $Y_0$  to be independent, comonotonic<sup>1</sup>, or countermonotonic given  $X$  and  $W$ .

A similar result is obtained under joint  $c$ -dependence.

**Theorem 4.4.2.** *Let Assumptions 11 and 13 hold. For all  $(\varepsilon, \gamma) \in [0, 1]^2$  there exists  $C_{1,0|X,W} \in \mathcal{C}_{1,0|X,W}$  such that*

$$(\varepsilon \underline{F}_{Y_1|W} + (1 - \varepsilon) \overline{F}_{Y_1|W}, \gamma \underline{F}_{Y_0|W} + (1 - \gamma) \overline{F}_{Y_0|W}, C_{1,0|X,W}) \in \mathcal{I}_0^{joint}(F_{Y,X,W}; c).$$

The only difference between the two theorems concerns the dependence structures between  $Y_1$  and  $Y_0$ . Theorem 4.4.1 shows that all copulas are compatible with marginal  $c$ -dependence, while our proof of Theorem 4.4.2 only exhibits one copula for each pair of conditional distributions  $(\varepsilon \underline{F}_{Y_1|W} + (1 - \varepsilon) \overline{F}_{Y_1|W}, \gamma \underline{F}_{Y_0|W} + (1 - \gamma) \overline{F}_{Y_0|W})$ .

#### 4.4.2 Bounds on monotonic parameters

The sharp bounds provided in theorems 4.4.1 and 4.4.2 can be used to deliver analytical expressions for sharp bounds on a large class of treatment effect parameters. We first define the identified set for a parameter  $\theta$  defined in (4.2).

**Definition 4.4.2** (Parameter Identified Sets). *Let  $\theta(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W})$  be a parameter. Its identified set under marginal  $c$ -dependence with bounds  $c := (\underline{c}(w, \eta), \overline{c}(w, \eta))$*

<sup>1</sup> This is also referred to as *rank invariance*. For example, see the discussion in J. J. Heckman et al. (1997).

is given by

$$\mathcal{I}_\theta^{margin}(F_{Y,X,W}; c) := \{\theta(F_1, F_0, C, F_{Y,X,W}) : (F_1, F_0, C) \in \mathcal{I}^{margin}(F_{Y,X,W}; c)\}.$$

For a parameter  $\theta(F_{Y_1|X,W}, F_{Y_0|X,W}, F_{Y,X,W})$  that does not depend on the copula  $C_{1,0|X,W}$ , its identified set under joint  $c$ -dependence is given by

$$\mathcal{I}_\theta^j(F_{Y,X,W}; c) := \{\theta(F_1, F_0, F_{Y,X,W}) : (F_1, F_0, C) \in \mathcal{I}^{joint}(F_{Y,X,W}; c) \text{ for some } C \in \mathcal{C}_{1,0|X,W}\}.$$

These sets are the parameter values consistent with the known distribution of observables  $F_{Y,X,W}$  and with a  $c$ -dependence condition. Without restrictions on how  $\theta$  depends on the distribution of potential outcomes, these sets may take various shapes.

We focus on a class of scalar estimands that can be ordered with respect to first order stochastic dominance.

**Definition 4.4.3** (First-Order Stochastic Dominance). *Let  $\mathcal{F}$  be the set of all univariate cdfs and let  $F, G \in \mathcal{F}$ . Say that  $F$  first-order stochastically dominates  $G$ , denoted by  $F \geq G$ , if  $F(u) \leq G(u)$  for all  $u \in \mathbb{R}$ .*

Next we define our target class of parameters.

**Definition 4.4.4** (Monotonic Parameters). *Let  $\theta : \mathcal{F} \rightarrow \mathbb{R}$  be a parameter. Say that  $\theta$  is increasing if  $F \geq G$  implies  $\theta(F) \geq \theta(G)$ . Say that  $\theta$  is decreasing if  $-\theta$  is increasing, and say  $\theta$  is monotonic if it is either increasing or decreasing.*

Following Manski (1997), monotonic parameters are also called  $D$ -parameters, or  $D_1$ -parameters. Also see Manski (2003) or Stoye (2010) who consider parameters that are increasing with respect to second-order stochastic dominance.

As an example, consider a parameter  $\theta(F_{Y_1|W})$  that is increasing in  $F_{Y_1|W}$  and suppose  $c$ -dependence holds. Then

$$\theta(F_{Y_1|W}) \in \left[ \theta(\overline{F}_{Y_1|W}), \theta(\underline{F}_{Y_1|W}) \right]$$

since  $\overline{F}_{Y_1|W} \leq F_{Y_1|W} \leq \underline{F}_{Y_1|W}$ , which holds by Lemma 4.4.1. This interval cannot be made narrower since the cdf bounds  $[\underline{F}_{Y_1|W}, \overline{F}_{Y_1|W}]$  are sharp by theorems 4.4.1 and 4.4.2.

Therefore, the identified set for  $\theta(F_{Y_1|W})$  is a subset of this closed interval that always contains its two endpoints. The interior of this interval is also part of the identified set if the functional  $\theta$  is continuous in the sense that  $\varepsilon \mapsto \theta(\varepsilon \underline{F}_{Y_1|W} + (1-\varepsilon) \overline{F}_{Y_1|W})$  is continuous. This type of continuity is implied by the continuity of the mapping  $F \mapsto \theta(F)$  under the sup-distance metric.

Assuming monotonicity of the parameter will help derive properties of its identified set. Monotonicity is a substantive restriction, but all eleven parameters from Section 4.2 satisfy it. This is formally established in Lemma 4.4.2 below. We begin by considering monotonic parameters that do not depend on copulas.

**Theorem 4.4.3.** *Let  $\theta(F_{Y_1|W}, F_{Y_0|W}, F_{Y,X,W})$  be increasing in  $F_{Y_1|W}(\cdot | w)$  and decreasing in  $F_{Y_0|W}(\cdot | w)$  for each  $w \in \text{supp}(W)$ . Let Assumption 11 hold, and either Assumption 12 or 13 hold. Then, for  $i \in \{\text{marg}, \text{joint}\}$  the convex hull of the identified set for  $\theta(F_{Y_1|W}, F_{Y_0|W}, F_{Y,X,W})$  is the closed interval*

$$\begin{aligned} & \mathcal{I}_\theta^i(F_{Y,X,W}; c) \\ &= \left[ \inf_{(F_1, F_0, C) \in \mathcal{I}_0^i(F_{Y,X,W}; c)} \theta(F_1, F_0, F_{Y,X,W}), \sup_{(F_1, F_0, C) \in \mathcal{I}_0^i(F_{Y,X,W}; c)} \theta(F_1, F_0, F_{Y,X,W}) \right] \\ &= \left[ \theta(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_{Y,X,W}), \theta(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, F_{Y,X,W}) \right]. \end{aligned}$$

If  $(\varepsilon, \gamma) \mapsto \theta(\varepsilon \underline{F}_{Y_1|W} + (1-\varepsilon) \overline{F}_{Y_1|W}, \gamma \underline{F}_{Y_0|W} + (1-\gamma) \overline{F}_{Y_0|W}, F_{Y,X,W})$  is continuous over  $(\varepsilon, \gamma) \in [0, 1]^2$ , this interval equals the identified set.

This theorem shows that substituting the upper/lower cdf bounds delivers sharp bounds for any parameter that is monotonic in the first-order stochastic dominance sense. The result is derived under an assumption that the parameter is increasing in  $F_{Y_1|W}$  and decreasing in  $F_{Y_0|W}$ , but it immediately generalizes to parameters that are increasing or decreasing in either or both conditional cdfs. For example, the cdf pair  $(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W})$  will maximize a parameter that is decreasing in  $F_{Y_1|W}$  and increasing in  $F_{Y_0|W}$ , and the cdf pair  $(\overline{F}_{Y_1|W}, \overline{F}_{Y_0|W})$  will maximize (minimize) a parameter that is decreasing (increasing) in

both  $F_{Y_1|W}$  and  $F_{Y_0|W}$ . The identified set for these parameters always contains endpoints  $\theta(\bar{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_{Y,X,W})$  and  $\theta(\underline{F}_{Y_1|W}, \bar{F}_{Y_0|W}, F_{Y,X,W})$ . It also contains all the values between these endpoints whenever the mapping  $\theta$  is continuous in the appropriate sense.

We document the monotonicity of various building blocks for parameters of interest in the following technical lemma. We omit covariates  $W$  for simplicity here, except in part 4 on QCATE because that parameter requires covariates to be nontrivial.

**Lemma 4.4.2.** *Let Assumption 11 hold. Then, for  $x \in \{0, 1\}$  and  $\tau \in (0, 1)$ ,*

1.  $\theta_{\mathbb{E}}(F_{Y_x}) := \int y dF_{Y_x}(y)$  is increasing and continuous in the sense that  $\varepsilon \mapsto \theta_{\mathbb{E}}(\varepsilon F_{Y_x} + (1 - \varepsilon)F'_{Y_x})$  is continuous for any  $(F_{Y_x}, F'_{Y_x})$  over  $\varepsilon \in [0, 1]$ .
2.  $\theta_Q(F_{Y_x}; \tau) := F_{Y_x}^{-1}(\tau)$  is increasing.
3.  $\theta_{CQ}(F_{Y_x}; \tau) := F_{Y_x|X}^{-1}(\tau | 1 - x)$  is increasing.
4.  $\theta_{QCATE}(F_{Y_1|W}, F_{Y_0|W}, F_W; \tau)$  (see Example 4.2.7) is decreasing in  $F_{Y_1|W}$  and increasing in  $F_{Y_0|W}$ .
5.  $\theta_{CDF}(F_{Y_1}, F_{Y_0}, C; y_1, y_0) := C(F_{Y_1}(y_0), F_{Y_0}(y_0))$  is decreasing in  $F_{Y_1}$  and  $F_{Y_0}$  for all  $(y_1, y_0) \in \mathbb{R}^2$  and copulas  $C$ .
6.  $\theta_{DTE}(F_{Y_1}, F_{Y_0}, C; z) := \int_{\{y_1 - y_0 \leq z\}} dC(F_{Y_1}(y_1), F_{Y_0}(y_0))$  is decreasing in  $F_{Y_1}$  and increasing in  $F_{Y_0}$  for all  $z \in \mathbb{R}$  and copulas  $C$ .

Using this lemma, all eight parameters that are invariant to copulas are bounded by substituting the upper or lower cdf bounds from Theorem 4.4.1. This allows us to compute analytical bounds for these parameters.

## 4.5 Analytical Bounds on Treatment Effect Parameters

We explore these analytical bounds by focusing on five of our examples to illustrate these expressions. The first three parameters are independent from the copula, while the last two are copula-dependent.

### 4.5.1 Average treatment effects (Example 4.2.3)

From Lemma 4.4.2.1, we have that the ATE satisfies

$$\mathbb{E}[Y_1 - Y_0] = \theta_{\text{ATE}}(F_{Y_1|W}, F_{Y_0|W}, F_W) \in [\theta_{\text{ATE}}(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_W), \theta_{\text{ATE}}(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, F_W)].$$

This interval equals the identified set by the monotonicity and continuity of the expectation functional which was established in Lemma 4.4.2.1. The lower and upper bounds can be obtained by calculating  $\int y d\overline{F}_{Y_x|W}(y | w)$  and  $\int y d\underline{F}_{Y_x|W}(y | w)$  for  $x \in \{0, 1\}$ . Via the quantile transformation, these bounds can also be written as integrals of  $\underline{Q}_{Y_x|W}(u | w)$  and  $\overline{Q}_{Y_x|W}(u | w)$  over  $u \in (0, 1)$ . Thus the ATE bounds can be written as integrals of quantiles. Via Lemma C.5.2 in Appendix C.5, we show that these quantile integrals can be converted into conditional expectations of outcomes given that they exceed or fall short of a fixed conditional quantile. These are equivalent to Conditional Value at Risk (CVaR) measures that appear in Dorn et al. (2024). These bounds are stated explicitly in equations (C.45)–(C.48) in Appendix C.5 in the general case. When  $Y | X, W$  is continuously distributed, we obtain simpler expressions for these bounds that we give here:

$$\begin{aligned} \theta_{\text{ATE}}(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_W) = & \mathbb{E} \left[ \left( \mathbb{E}[Y | Y \leq \overline{Q}_1, X = 1, W] - \mathbb{E}[Y | Y \leq \underline{Q}_0, X = 0, W] \right) \frac{\bar{c} - p_{1|W}}{\bar{c} - \underline{c}} \right] \\ & + \mathbb{E} \left[ \left( \mathbb{E}[Y | Y > \overline{Q}_1, X = 1, W] - \mathbb{E}[Y | Y > \underline{Q}_0, X = 0, W] \right) \frac{p_{1|W} - \underline{c}}{\bar{c} - \underline{c}} \right] \end{aligned}$$

and

$$\begin{aligned} \theta_{\text{ATE}}(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, F_W) = & \mathbb{E} \left[ \left( \mathbb{E}[Y | Y \leq \underline{Q}_1, X = 1, W] - \mathbb{E}[Y | Y \leq \overline{Q}_0, X = 0, W] \right) \frac{p_{1|W} - \underline{c}}{\bar{c} - \underline{c}} \right] \\ & + \mathbb{E} \left[ \left( \mathbb{E}[Y | Y > \underline{Q}_1, X = 1, W] - \mathbb{E}[Y | Y > \overline{Q}_0, X = 0, W] \right) \frac{\bar{c} - p_{1|W}}{\bar{c} - \underline{c}} \right]. \end{aligned}$$

Note that the dependence of  $(\underline{c}, \bar{c})$  on  $(W, \eta)$  was suppressed for convenience.

### 4.5.2 Quantile treatment effects (Example 4.2.5)

We now consider bounds on the quantile treatment effect for a fixed quantile  $\tau \in (0, 1)$ . By Lemma 4.4.2.2, the functional  $\theta_{\text{QTE}}$  is increasing in  $F_{Y_1|W}$  and decreasing in  $F_{Y_0|W}$ . Therefore, by Theorem 4.4.3,  $\text{QTE}(\tau)$  has the following sharp bounds:

$$\text{QTE}(\tau) \in \left[ \underline{Q}_{Y_1}(\tau) - \overline{Q}_{Y_0}(\tau), \overline{Q}_{Y_1}(\tau) - \underline{Q}_{Y_0}(\tau) \right]$$

where  $\overline{Q}_{Y_x}$  is the left-inverse of cdf  $\underline{F}_{Y_x}(\cdot) := \mathbb{E}[\underline{F}_{Y_x|W}(\cdot | W)]$  for  $x \in \{0, 1\}$ . Analogously,  $\underline{Q}_{Y_x}$  is the left-inverse of cdf  $\overline{F}_{Y_x}(\cdot) := \mathbb{E}[\overline{F}_{Y_x|W}(\cdot | W)]$ . Analytical expressions for the unconditional cdf bounds for the treated potential outcome are given by

$$\begin{aligned} \overline{F}_{Y_1}(y) &= \mathbb{E} \left[ \min \left\{ F_{Y|X,W}(y | 1, W) \frac{p_{1|W}}{\underline{c}}, \frac{\bar{c} - p_{1|W}}{\bar{c}} + F_{Y|X,W}(y | 1, W) \frac{p_{1|W}}{\bar{c}} \right\} \right] \\ \underline{F}_{Y_1}(y) &= \mathbb{E} \left[ \max \left\{ F_{Y|X,W}(y | 1, W) \frac{p_{1|W}}{\bar{c}}, \frac{\underline{c} - p_{1|W}}{\underline{c}} + F_{Y|X,W}(y | 1, W) \frac{p_{1|W}}{\underline{c}} \right\} \right] \end{aligned}$$

and similar expressions can be obtained for  $Y_0$ . The left-inverses of the previous expressions yield bounds on quantiles of  $Y_1$  and  $Y_0$ , and which can be used to compute the QTE bounds.

### 4.5.3 Average weighted welfare (Example 4.2.8)

Consider a policy  $\omega : \text{supp}(W) \rightarrow [0, 1]$  that treats units with covariate value  $w$  with probability  $\omega(w)$ . The average welfare in a population under such policy is given by

$$\text{AWW}(\omega) = \theta_{\text{AWW}}(F_{Y_1|W}, F_{Y_0|W}, F_W, \omega) = \mathbb{E}[\omega(W)\mathbb{E}[Y_1 | W] + (1 - \omega(W))\mathbb{E}[Y_0 | W]].$$

By adapting Lemma 4.4.2.1, this functional is increasing in  $F_{Y_1|W}$ , increasing in  $F_{Y_0|W}$ , and continuous in the sense defined in the lemma. Therefore, by Theorem 4.4.3, its identified set is the closed interval given by

$$\left[ \theta_{\text{AWW}}(\overline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, F_W, \omega), \theta_{\text{AWW}}(\underline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_W, \omega) \right].$$

An analytical expression for these bounds can be obtained by substituting in the expressions for the cdf bounds in the previous functionals. When  $Y | X, W$  is continuously distributed,

the bounds are given by

$$\begin{aligned} & \theta_{\text{AWW}}(\bar{F}_{Y_1|W}, \bar{F}_{Y_0|W}, F_W, \omega) \\ &= \mathbb{E} \left[ \omega(W) \left( \mathbb{E}[Y | Y \leq \bar{Q}_1, X = 1, W] \frac{\bar{c} - p_{1|W}}{\bar{c} - \underline{c}} + \mathbb{E}[Y | Y > \bar{Q}_1, X = 1, W] \frac{p_{1|W} - \underline{c}}{\bar{c} - \underline{c}} \right) \right] \\ &+ \mathbb{E} \left[ (1 - \omega(W)) \left( \mathbb{E}[Y | Y \leq \bar{Q}_0, X = 0, W] \frac{p_{1|W} - \underline{c}}{\bar{c} - \underline{c}} + \mathbb{E}[Y | Y > \bar{Q}_0, X = 0, W] \frac{\bar{c} - p_{1|W}}{\bar{c} - \underline{c}} \right) \right] \end{aligned}$$

and

$$\begin{aligned} & \theta_{\text{AWW}}(\underline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_W, \omega) \\ &= \mathbb{E} \left[ \omega(W) \left( \mathbb{E}[Y | Y \leq \underline{Q}_1, X = 1, W] \frac{p_{1|W} - \underline{c}}{\bar{c} - \underline{c}} + \mathbb{E}[Y | Y > \underline{Q}_1, X = 1, W] \frac{\bar{c} - p_{1|W}}{\bar{c} - \underline{c}} \right) \right] \\ &+ \mathbb{E} \left[ (1 - \omega(W)) \left( \mathbb{E}[Y | Y \leq \underline{Q}_0, X = 0, W] \frac{\bar{c} - p_{1|W}}{\bar{c} - \underline{c}} + \mathbb{E}[Y | Y > \underline{Q}_0, X = 0, W] \frac{p_{1|W} - \underline{c}}{\bar{c} - \underline{c}} \right) \right]. \end{aligned}$$

#### 4.5.4 Copula-dependent parameters

We now consider identification of the parameters in examples 4.2.9–4.2.11 which all depend on the copulas  $C_{1,0|X,W}$ . Even under unconfoundedness these parameters are not point-identified. Relaxing unconfoundedness will yield larger identified sets for these parameters when compared to the unconfoundedness baseline. We will focus on marginal  $c$ -dependence since it does not restrict the dependence structure between the potential outcomes.

##### The joint distribution function

Consider identification of the joint cdf  $F_{Y_1, Y_0}(y_1, y_0)$  under marginal  $c$ -dependence. Consider the functional

$$\begin{aligned} & \theta_{\text{CDF}}(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W}; y_1, y_0) \\ &:= \int (C_{1,0|X,W}(F_{Y_1|X,W}(y_1 | 1, w), F_{Y_0|X,W}(y_0 | 1, w) | 1, w) p_{1|w} \\ &\quad + C_{1,0|X,W}(F_{Y_1|X,W}(y_1 | 0, w), F_{Y_0|X,W}(y_0 | 0, w) | 0, w) p_{0|w}) dF_W(w). \end{aligned}$$

Fix the conditional copula function  $C_{1,0|X,W}(\cdot, \cdot | \cdot, \cdot)$ . Then by Lemma 4.4.2.4, this functional is decreasing in  $F_{Y_0|X,W}(y_0 | 1, w)$  and  $F_{Y_1|X,W}(y_1 | 0, w)$ . Thus it is bounded below

by

$$\theta_{\text{CDF}}(\underline{F}_{Y_1|X,W}, \underline{F}_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W}; y_1, y_0)$$

and above by

$$\theta_{\text{CDF}}(\overline{F}_{Y_1|X,W}, \overline{F}_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W}; y_1, y_0).$$

Moreover, by Theorem 4.4.3, these bounds are sharp.

Since  $C_{1,0|X,W}$  is unknown, we then compute the maximum and minimum of these bounds over the set of copulas that are consistent with marginal  $c$ -dependence; this is simply the set of all copulas. The Fréchet-Hoeffding bounds show that all copulas  $C$  satisfy

$$C(u, v) \in [\max\{u + v - 1, 0\}, \min\{u, v\}] =: [\underline{C}(u, v), \overline{C}(u, v)]$$

for all  $(u, v) \in [0, 1]^2$ . The copula bounds  $\underline{C}$  and  $\overline{C}$  are themselves copulas. Combining these facts, we obtain the following analytical bounds on the joint cdf of potential outcomes.

**Proposition 4.5.1** (Identified set for joint cdf). *Let assumptions 11 and 12 hold. Then, for any  $(y_1, y_0) \in \mathbb{R}^2$ , the identified set for  $F_{Y_1, Y_0}(y_1, y_0)$  is given by the closed interval*

$$\begin{aligned} \mathcal{I}_{\theta_{\text{CDF}}}^{\text{marg}}(F_{Y,X,W}; c) = & \left[ \mathbb{E} \left( \max\{\underline{F}_{Y_1|X,W}(y_1 | X, W) + \underline{F}_{Y_0|X,W}(y_0 | X, W) - 1, 0\} \right), \right. \\ & \left. \mathbb{E} \left( \min\{\overline{F}_{Y_1|X,W}(y_1 | X, W), \overline{F}_{Y_0|X,W}(y_0 | X, W)\} \right) \right]. \end{aligned} \quad (4.7)$$

The bounds in (4.7) are themselves cdfs, so these bounds can be attained simultaneously for all  $(y_1, y_0) \in \mathbb{R}^2$ . The bounds for  $F_{Y_1, Y_0}$  under unconfoundedness are obtained as a special case when  $\underline{c} = p_{1|W} = \overline{c}$ , which implies that  $\overline{F}_{Y_x|X,W} = \underline{F}_{Y_x|X,W} = F_{Y|X,W}(\cdot | x, \cdot)$ . Making this substitution in equation (4.7) yields these bounds under unconfoundedness.

### The distribution of treatment effects

Identification of this parameter under unconfoundedness was studied in Fan and Park (2010), by applying results first shown in Makarov (1982) and later studied in Williamson and Downs (1990). Masten and Poirier (2020) also studied this parameter under conditional

$c$ -dependence and under a range of assumptions on copulas for  $(Y_1, Y_0)$ . By Lemma 2.1 in Fan and Park (2010), the cdf of  $Y_1 - Y_0$  given  $(X, W) = (x, w)$  satisfies

$$F_{Y_1 - Y_0 | X, W}(z | x, w) \in \left[ \max \left\{ \sup_{y \in \mathbb{R}} (F_{Y_1 | X, W}(y | x, w) - F_{Y_0 | X, W}(y - z | x, w)), 0 \right\}, \right. \\ \left. 1 + \min \left\{ \inf_{y \in \mathbb{R}} (F_{Y_1 | X, W}(y | x, w) - F_{Y_0 | X, W}(y - z | x, w)), 0 \right\} \right]$$

and these bounds are sharp for any pair of cdfs  $(F_{Y_1 | X, W}, F_{Y_0 | X, W})$ . These bounds are decreasing in  $F_{Y_1 | X, W}$  and increasing in  $F_{Y_0 | X, W}$ , therefore substituting the upper/lower cdf bounds for  $F_{Y_x | X, W}$  results in sharp bounds for  $F_{Y_1 - Y_0 | X, W}$  under  $c$ -dependence. This was established in Lemma 4.4.2.6. Integrating these bounds over the marginal distribution of  $(X, W)$  yields sharp bounds for the unconditional cdf of  $Y_1 - Y_0$ . This result is summarized in the following proposition.

**Proposition 4.5.2** (Identified set for DTE). *Let assumptions 11 and 12 hold. For any  $z \in \mathbb{R}$ , the convex hull of the identified set for  $F_{Y_1 - Y_0}(z)$  is given by*

$$\mathcal{I}_{\theta_{DTE}}^{marg}(F_{Y, X, W}; c) = \\ \left[ \mathbb{E} \left( \max \left\{ \sup_{y \in \mathbb{R}} (\underline{F}_{Y_1 | X, W}(y | X, W) - \overline{F}_{Y_0 | X, W}(y - z | X, W)), 0 \right\} \right), \right. \\ \left. 1 + \mathbb{E} \left( \min \left\{ \inf_{y \in \mathbb{R}} (\overline{F}_{Y_1 | X, W}(y | X, W) - \underline{F}_{Y_0 | X, W}(y - z | X, W)), 0 \right\} \right) \right].$$

This expression involves two one-dimensional optimization problems, but the objective functions are known, closed-form functionals of the distribution of the observables. Bounds on the QDTE can be obtained as a corollary by taking the left-inverse of the cdf bounds.

## 4.6 Conclusion

In this paper we proposed a general class of relaxations of unconfoundedness, and showed how it includes several previous approaches as special cases. We then derived closed form identification results for many different target parameters under this general class of relaxations. There are at least three natural next steps. First, in this paper we focused

on population level identification results. Corresponding estimation and inference results can likely be derived by using standard sample analog estimators and arguments, but we leave the details to future work. Second, it would be interesting to explore whether our bounds have either the double-sharpness or double-validity properties defined in Dorn et al. (2024), and if not, whether alternative bounds that had these properties could be derived. Third, it would be interesting to extend our results to independence assumptions beyond unconfoundedness, such as IV exogeneity (e.g., section 4 of Masten and Poirier 2018b).

## 5. Conclusions

This dissertation consists of three essays in microeconometrics. The first essay addresses the weak instrument problem in MTE models by introducing uniformly valid inference procedures for causal parameters extrapolated by MTEs, achieving the robustness of policy effect inference when instrumental variables exhibit limited variation. The second essay develops three testing procedures, based on asymptotic approximation, bootstrap, and permutations, to assess the marginal homogeneity assumption in panel data. The third essay presents a general framework for relaxing the unconfoundedness assumption, which can include several existing approaches as special cases. This framework enables a broader sensitivity analysis for various treatment effect parameters.

## Appendix A. Additional Results for Chapter 2

### A.1 Proofs

#### A.1.1 Notation

Throughout the appendix, we employ the notation in Table A.1, which was not necessarily introduced in the text.

Table A.1: Important Notation in Appendix A.1

$q_F(z_\ell)$	$\mathbb{P}_F(Z = z_\ell)$
$q_F(d, z_\ell)$	$\mathbb{P}_F(D = d, Z = z_\ell)$
$p_F(z_\ell)$	$\mathbb{P}_F(D = 1 \mid Z = z_\ell)$
$p_F$	$(p_F(z_0), p_F(z_1), \dots, p_F(z_K))'$
$A_F$	The matrix of propensity scores, defined in equations (2.5) and (2.10)
$\beta_{F,d\ell}$	$\mathbb{E}_F[Y \mid D = d, Z = z_\ell]$
$\beta_{F,d}$	$(\beta_{F,d0}, \dots, \beta_{F,dK})'$
$\beta_F$	$(\beta'_{F,1}, \beta'_{F,0})'$
$\sigma_{F,d\ell}^2$	$\text{var}_F(Y \mid D = d, Z = z_\ell)$
$\{\tau_{j,F}\}_{j=1}^{2(M+1)}$	The singular values of $A_F$ (in a descending order)

In addition, let  $\mathbb{S}_{++}^k$  denote the space of positive definite matrices with  $k$  rows (columns).

#### A.1.2 Proof of uniform validity in Section 2.3

The proof of asymptotic similarity of the AR and conditional tests in Proposition 2.3.1 and Theorem 2.3.1 uses the sub-sequencing techniques from D. W. Andrews et al., 2020. Specifically, we verify their Assumption B\* in the following Proposition A.1.1.

**Proposition A.1.1.** *For any subsequence  $\{p_n\}$  of  $\{n\}$  and any sequence  $\{(\lambda_{p_n}, F_{p_n}) \in \mathcal{P}_0\}$  for which*

1.  $\theta_{F_{p_n}} \rightarrow \theta_\infty \in \Theta$ , where  $\theta_{F_{p_n}}$  is a sequence of parameters such that  $(\theta_{F_{p_n}}, F_{p_n}) \in \mathcal{P}$  and  $\lambda_{p_n} = c'\theta_{F_{p_n}}$ ;
2.  $\beta_{F_{p_n},d\ell} \rightarrow \beta_{\infty,d\ell}$  for  $d = 0, 1$  and  $\ell = 0, 1, \dots, K$ ;
3.  $\sigma_{F_{p_n},d\ell}^2 \rightarrow \sigma_{\infty,d\ell}^2$  for  $d = 0, 1$  and  $\ell = 0, 1, \dots, K$ ;
4.  $p_{F_{p_n}}(z_\ell) \rightarrow p_\infty(z_\ell)$  for all  $\ell = 0, 1, \dots, K$ ;

5.  $q_{F_{p_n}}(z_\ell) \rightarrow q_\infty(z_\ell)$  for all  $\ell = 0, 1, \dots, K$ .

The following result holds:

(a) the convergence of the type-I error rate of AR test:

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} \left( AR_{p_n, k}(\lambda_{p_n}) > q_{\chi_1^2}(1 - \alpha) \right) = \alpha, \text{ for } k = 1, \dots, K,$$

(b) Suppose further that

$$\sqrt{p_n} \|\pi_{F_{p_n}}\| \rightarrow s_\infty \in [0, \infty], \text{ and } \pi_{F_{p_n}} / \|\pi_{F_{p_n}}\| \rightarrow \iota_\infty,$$

where  $\iota_\infty$  is a unit vector, then we have the convergence of the type-I error rate of the conditional test

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} (W_{p_n}(\lambda_{p_n}) > q_{W^*}(1 - \alpha)) = \alpha.$$

Based on the above result along subsequences, the asymptotic similarity of the AR and conditional test can be established similarly as D. W. Andrews et al. (2020, Theorem 2.1).

*Proof of Proposition 2.3.1.* Let  $\{a_n\}$  be a subsequence of  $\{n\}$  such that

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_{a_n}} \left( AR_{a_n, k}(\lambda_{a_n}) > q_{\chi_1^2}(1 - \alpha) \right) = \limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F \left( AR_{n, k}(\lambda) > q_{\chi_1^2}(1 - \alpha) \right).$$

Such a sequence always exists. Along the sequence  $\{a_n\}$ , we can choose a subsequence  $\{p_n\} \subset \{a_n\}$  such that the conditions 1-5 in Proposition A.1.1 hold for some

$$\vartheta_\infty \equiv (\theta_\infty, \beta_\infty, \{\sigma_{\infty, d\ell}^2\}_{d, \ell}, p_\infty, q_\infty).$$

That is,

$$\vartheta_{p_n} \equiv \left( \theta_{F_{p_n}}, \beta_{F_{p_n}}, \{\sigma_{F_{p_n}, d\ell}^2\}_{d, \ell}, p_{F_{p_n}}, q_{F_{p_n}} \right) \rightarrow \vartheta_\infty.$$

Such converging subsequence  $\{\vartheta_{p_n}\}$  always exists since each element of  $\vartheta_{a_n}$  is contained in a compact set as restricted by the parameter space  $\mathcal{P}$ . By Proposition A.1.1(a), we have

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} \left( AR_{p_n, k}(\lambda_{p_n}) > q_{\chi_1^2}(1 - \alpha) \right) = \alpha.$$

Since  $\{p_n\}$  is a subsequence of  $\{a_n\}$ , and the rejection probability converges along the sequence  $\{a_n\}$ , we have

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F \left( \text{AR}_{n,k}(\lambda_{p_n}) > q_{\chi_1^2}(1 - \alpha) \right) \\ &= \lim_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} \left( \text{AR}_{p_n,k}(\lambda_{p_n}) > q_{\chi_1^2}(1 - \alpha) \right) \\ &= \alpha. \end{aligned}$$

By redefining the subsequence under inf operator, similar arguments also imply that

$$\liminf_{n \rightarrow \infty} \inf_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F \left( \text{AR}_{n,k}(\lambda_{p_n}) > q_{\chi_1^2}(1 - \alpha) \right) = \alpha$$

Then the proof is complete.  $\square$

*Proof of Theorem 2.3.1.* The proof follows the same arguments in the proof of Proposition 2.3.1 by replacing AR statistic with the Wald statistic  $W_n(\lambda)$ , replacing the critical value  $q_{\chi_1^2}(1 - \alpha)$  with the conditional critical value  $q_{W^*}(1 - \alpha)$ , and redefining  $\vartheta_\infty$  as:

$$\vartheta_\infty \equiv (\theta_\infty, \beta_\infty, \{\sigma_{\infty, d\ell}^2\}_{d, \ell}, p_\infty, q_\infty, s_\infty, \iota_\infty).$$

The compactness of the space of  $s_\infty$  follows from the compactness of  $[0, \infty]$ . Instead of using Proposition A.1.1(a), we use Proposition A.1.1(b) to show the size of the conditional test along a subsequence  $\{p_n\}$ .  $\square$

*Proof of Corollary 2.3.1.* Note that Delta method gives a different asymptotic expansion for the moment along a converging sequence:

$$\begin{aligned} \sqrt{n}\hat{g}(\lambda_n) & \xrightarrow{d} (\partial_p g(\lambda_\infty) - \Delta_\mu[\partial_{p'} c_\mu] - \Delta_\rho[\partial_{p'} c_\rho]) \mathcal{Z}_p \\ & \quad - (\Delta_\mu[\partial_{q'} c_\mu] + \Delta_\rho[\partial_{q'} c_\rho]) \mathcal{Z}_q \\ & \quad + \partial_{\beta_1} g(\lambda_\infty) \mathcal{Z}_{\beta_1} + \partial_{\beta_0} g(\lambda_\infty) \mathcal{Z}_{\beta_0} \\ & \sim \mathcal{N}(0, S^\dagger(\lambda_\infty)). \end{aligned}$$

The asymptotic covariance matrix  $S^\dagger(\lambda_\infty)$  is uniformly positive definite by noting that Lemma A.1.2 has already established that  $\partial_{\beta_1} g(\lambda_\infty) \mathcal{Z}_{\beta_1} + \partial_{\beta_0} g(\lambda_\infty) \mathcal{Z}_{\beta_0}$  is non-degenerate uniformly as long as  $c$  is nonzero. Since  $\mathcal{Z}_p$ ,  $\mathcal{Z}_q$ , and  $\mathcal{Z}_\beta$  are jointly independent,  $\hat{S}^\dagger(\lambda)$  consistently estimates the asymptotic variance  $S^\dagger(\lambda_\infty)$ . The rest of the proof follows by the same arguments in Proposition A.1.1 and Theorem 2.3.1.  $\square$

*Proof of Proposition A.1.1.* For simplicity of notations, the proof is shown for the full sequence  $\{n\}$ . Then, we note that the same proof goes through with  $p_n$  in place of  $n$ .

**Part (a):** Note that  $g_k(\lambda)$  is a continuously differentiable function of  $\lambda$ ,  $p$ ,  $\beta_1$ , and  $\beta_0$ . Applying Delta method and the convergence in Lemma A.1.1(a) to the sample moment  $\hat{g}_k(\lambda)$  yields

$$\begin{aligned} \sqrt{n} \hat{g}_k(\lambda_n) &= [\partial_{p'} g_k(\lambda_n)] \hat{Z}_p + [\partial_{\beta_1'} g_k(\lambda_n)] \hat{Z}_{\beta_1} + [\partial_{\beta_0'} g_k(\lambda_n)] \hat{Z}_{\beta_0} + o_p(1) \\ &\xrightarrow{d} \mathcal{N}(0, s_k^2) \end{aligned}$$

where

$$s_k^2 = [\partial_{p'} g_k(\lambda_\infty)] \Sigma_{p,\infty} [\partial_{p'} g_k(\lambda_\infty)] + [\partial_{\beta_1'} g_k(\lambda_\infty)] \Sigma_{\beta_1,\infty} [\partial_{\beta_1'} g_k(\lambda_\infty)] + [\partial_{\beta_0'} g_k(\lambda_\infty)] \Sigma_{\beta_0,\infty} [\partial_{\beta_0'} g_k(\lambda_\infty)],$$

where  $\lambda_\infty$  is the limit of  $\lambda_n = c' \theta_n$  as  $n$  goes to infinity, and  $g_k(\lambda_\infty)$  equals  $g_k(\lambda)$  by sending  $\lambda$ ,  $p$ ,  $\beta_1$ , and  $\beta_0$  to their limiting values.

Again by Lemma A.1.1(b),  $\Sigma_{p,\infty}$ ,  $\Sigma_{\beta_1,\infty}$ , and  $\Sigma_{\beta_0,\infty}$  can be consistently estimated by their sample analogs  $\hat{\Sigma}_p$ ,  $\hat{\Sigma}_{\beta_1}$ , and  $\hat{\Sigma}_{\beta_0}$  along the drifting sequence, respectively. The continuous mapping theorem then implies that

$$\hat{s}_k^2(\lambda_n) = \partial_{p'} \hat{g}_k(\lambda_n) \hat{\Sigma}_p \partial_{p'} \hat{g}_k(\lambda_n) + \partial_{\beta_1'} \hat{g}_k(\lambda_n) \hat{\Sigma}_{\beta_1} \partial_{\beta_1'} \hat{g}_k(\lambda_n) + \partial_{\beta_0'} \hat{g}_k(\lambda_n) \hat{\Sigma}_{\beta_0} \partial_{\beta_0'} \hat{g}_k(\lambda_n) \xrightarrow{p} s_k^2.$$

Note that Lemma A.1.2 implies  $s_k^2 \in (0, \infty)$ . By Slutsky's theorem, we have

$$\frac{\sqrt{n} \hat{g}_k(\lambda_n)}{\hat{s}_k(\lambda_n)} \xrightarrow{d} \mathcal{N}(0, 1),$$

which implies the desired result.

**Part (b):** Following similar arguments in part (a), Lemma A.1.1 implies that

$$\begin{aligned}
\sqrt{n}\hat{g}(\lambda_n) &\xrightarrow{d} \mathcal{Z}_{\mathfrak{g}} \equiv \partial_p g(\lambda_\infty)\mathcal{Z}_p + \partial_{\beta_1} g(\lambda_\infty)\mathcal{Z}_{\beta_1} + \partial_{\beta_0} g(\lambda_\infty)\mathcal{Z}_{\beta_0} \sim \mathcal{N}(0_{K \times 1}, S_\infty) \\
\partial_x \hat{g}(\lambda_n) &\xrightarrow{p} \partial_x g(\lambda_\infty) \quad \text{for } x \in \{p, \beta_1, \beta_0\} \\
\hat{\Sigma}_x &\xrightarrow{p} \Sigma_{x, \infty} \quad \text{for } x \in \{p, \beta_1, \beta_0\} \\
\hat{S}(\lambda_n) &\xrightarrow{p} S_\infty
\end{aligned} \tag{A.1}$$

where

$$S_\infty \equiv \partial_p g(\lambda_\infty)\Sigma_{p, \infty}\partial_{p'} g(\lambda_\infty) + \partial_{\beta_1} g(\lambda_\infty)\Sigma_{\beta_1, \infty}\partial_{\beta_1'} g(\lambda_\infty) + \partial_{\beta_0} g(\lambda_\infty)\Sigma_{\beta_0, \infty}\partial_{\beta_0'} g(\lambda_\infty).$$

and  $S_\infty$  is bounded and positive definite by Lemma A.1.2.

**Case 1:**  $s_\infty < \infty$ .

In this case, we have

$$\begin{aligned}
\sqrt{n}\hat{\pi} &= \sqrt{n}\pi_{F_n} + [\partial_p \pi]\hat{Z}_p \\
&\xrightarrow{d} s_\infty \iota_\infty + [\partial_p \pi]\mathcal{Z}_p \\
&\equiv \mathcal{Z}_\pi
\end{aligned} \tag{A.2}$$

and

$$\begin{aligned}
\hat{h}(\lambda_n) &= \sqrt{n}\hat{\pi} - [\partial_p \pi]\hat{\Sigma}_p[\partial_{p'} \hat{g}(\lambda_n)]\hat{S}(\lambda_n)^{-1}\sqrt{n}\hat{g}(\lambda_n) \\
&\xrightarrow{d} s_\infty \iota_\infty + [\partial_p \pi]\mathcal{Z}_p - [\partial_p \pi]\Sigma_{p, \infty}[\partial_{p'} g(\lambda_\infty)]S_\infty^{-1}\mathcal{Z}_{\mathfrak{g}} \\
&\equiv \mathcal{Z}_h
\end{aligned}$$

From the independence  $(\mathcal{Z}_{\beta_1}, \mathcal{Z}_{\beta_0}) \perp\!\!\!\perp \mathcal{Z}_p$ , it follows that  $\mathcal{Z}_h \perp\!\!\!\perp \mathcal{Z}_{\mathfrak{g}}$  since they are jointly normal and uncorrelated.

By continuous mapping theorem, we obtain the asymptotic distribution of the Wald statistic under the null:

$$\begin{aligned}
W_n(\lambda_n) &= \frac{n\hat{g}(\lambda_n)'\hat{S}(\lambda_n)^{-1}\hat{\pi}\hat{\pi}'\hat{S}(\lambda_n)^{-1}\hat{g}(\lambda_n)}{\hat{\pi}'\hat{S}(\lambda_n)^{-1}\hat{\pi}} \\
&\xrightarrow{d} \frac{\mathcal{Z}_{\mathfrak{g}}'S_\infty^{-1}\mathcal{Z}_\pi\mathcal{Z}_\pi'S_\infty^{-1}\mathcal{Z}_{\mathfrak{g}}}{\mathcal{Z}_\pi'S_\infty^{-1}\mathcal{Z}_\pi}.
\end{aligned}$$

On the other hand, we note that the simulated statistic  $W_n^*(\lambda_n)$  can be written as a function of standard normal draw  $\eta^*$  and a conditioning statistic

$$\hat{\Upsilon} \equiv (\hat{h}(\lambda_n), \hat{S}, \hat{\Sigma}_p, \partial_p \hat{g}(\lambda_n)).$$

Specifically, we write

$$W_n^*(\lambda_n) = W(\eta^*, \hat{\Upsilon}) \equiv \frac{(\eta^*)' \hat{S}(\lambda_n)^{-1/2} \pi_s(\eta^*, \hat{\Upsilon}) \pi_s(\eta^*, \hat{\Upsilon})' \hat{S}(\lambda_n)^{-1/2} \eta^*}{\pi_s(\eta^*, \hat{\Upsilon})' \hat{S}(\lambda_n)^{-1} \pi_s(\eta^*, \hat{\Upsilon})} \quad (\text{A.3})$$

where

$$\pi_s(\eta^*, \hat{\Upsilon}) \equiv \hat{h}(\lambda_n) + [\partial_p \pi] \hat{\Sigma}_p [\partial_{p'} \hat{g}(\lambda_n)] \hat{S}(\lambda_n)^{-1/2} \eta^*.$$

Therefore,  $\hat{\Upsilon}$  is the source of sampling uncertainty on the simulated statistic  $W_n^*(\lambda_n)$ .

Let  $G(\cdot \mid \hat{\Upsilon})$  denote the CDF of the simulated Wald statistic  $W_n^*(\lambda_n)$  conditional on data, i.e.,

$$\begin{aligned} G(x \mid \hat{\Upsilon}) &\equiv \mathbb{P}(W_n^*(\lambda_n) \leq x \mid \{Y_i, D_i, Z_i\}_{i=1}^n) \\ &= \mathbb{P}(W(\eta^*, \hat{\Upsilon}) \leq x \mid \hat{\Upsilon}). \end{aligned}$$

Next, we show that  $q(1 - \alpha, \hat{\Upsilon})$ , the  $(1 - \alpha)$ -quantile of  $G(\cdot \mid \hat{\Upsilon})$ , is a continuous function of  $\hat{\Upsilon}$  on the set  $\mathcal{U}$ , where

$$\begin{aligned} \mathcal{U} &= \{(u_1, u_2, u_3, u_4) : u_1 + [\partial_p \pi] u_3 u_4' u_2^{-1/2} \eta^* \neq 0_{K \times 1} \text{ a.s.}, \\ &\quad u_1 \in \mathbb{R}^K, u_2 \in \mathbb{S}_{++}^K, u_3 \in \mathbb{S}_{++}^{K+1}, u_4 \in \mathbb{R}^{K \times (K+1)}\}. \end{aligned} \quad (\text{A.4})$$

Let  $\{\Upsilon_n\}$  be a sequence in  $\mathcal{U}$  such that  $\Upsilon_n \rightarrow \Upsilon \in \mathcal{U}$  as  $n \rightarrow \infty$ . Given zero probability of discontinuity in the limit by the definition of  $\mathcal{U}$ ,  $\{W(\eta^*, \Upsilon_n)\}_{n \geq 1}$  converges almost surely to  $W(\eta^*, \Upsilon)$ . This implies the convergence in distribution for any continuity point  $x$  of  $G(\cdot \mid \Upsilon)$ :

$$G(x \mid \Upsilon_n) = \mathbb{P}_{\eta^*}(W(\eta^*, \Upsilon_n) \leq x) \rightarrow \mathbb{P}_{\eta^*}(W(\eta^*, \Upsilon) \leq x) = G(x \mid \Upsilon).$$

The distribution function  $G(\cdot \mid \Upsilon)$  is increasing at its  $(1 - \alpha)$ -quantile  $q(1 - \alpha, \Upsilon)$  because the random variable  $W(\eta^*, \Upsilon)$  is continuously distributed. By D. W. Andrews and

Guggenberger (2010, Lemma 5), it follows that  $q(1 - \alpha, \Upsilon_n) \rightarrow q(1 - \alpha, \Upsilon)$ . This establishes continuity of quantile function on the set  $\mathcal{U}$ .

The convergence results in (A.1) and (A.2) give

$$\hat{\Upsilon} \xrightarrow{d} \Upsilon_\infty \equiv (\mathcal{Z}_h, S_\infty, \Sigma_{p,\infty}, \partial_p g(\lambda_\infty)),$$

and by Lemma A.1.3, we have  $\mathbb{P}(\Upsilon_\infty \in \mathcal{U}) = 1$ . From the continuity of  $q(1 - \alpha, \Upsilon)$ , it follows by continuous mapping theorem that

$$W_n(\lambda_n) - q(1 - \alpha, \hat{\Upsilon}) \xrightarrow{d} \frac{\mathcal{Z}'_{\mathbf{g}} S_\infty^{-1} \mathcal{Z}_\pi \mathcal{Z}'_\pi S_\infty^{-1} \mathcal{Z}_{\mathbf{g}}}{\mathcal{Z}_\pi S_\infty^{-1} \mathcal{Z}_\pi} - q(1 - \alpha, \Upsilon_\infty),$$

which implies

$$\mathbb{P}_{F_n} \left( W_n(\lambda_n) > q(1 - \alpha, \hat{\Upsilon}) \right) \rightarrow \mathbb{P} \left( \frac{\mathcal{Z}'_{\mathbf{g}} S_\infty^{-1} \mathcal{Z}_\pi \mathcal{Z}'_\pi S_\infty^{-1} \mathcal{Z}_{\mathbf{g}}}{\mathcal{Z}_\pi S_\infty^{-1} \mathcal{Z}_\pi} > q(1 - \alpha, \Upsilon_\infty) \right).$$

Next, we show that the limit probability on the right-hand side equals  $\alpha$ . We first examine the conditional probability as follows.

$$\begin{aligned} & \mathbb{P} \left( \frac{\mathcal{Z}'_{\mathbf{g}} S_\infty^{-1} \mathcal{Z}_\pi \mathcal{Z}'_\pi S_\infty^{-1} \mathcal{Z}_{\mathbf{g}}}{\mathcal{Z}_\pi S_\infty^{-1} \mathcal{Z}_\pi} > q(1 - \alpha, \Upsilon_\infty) \mid \Upsilon_\infty \right) \\ &= \mathbb{P} \left( \frac{(\eta^*)' S_\infty^{-1/2} \pi_s(\eta^*, \Upsilon_\infty) \pi_s(\eta^*, \Upsilon_\infty)' S_\infty^{-1/2} \eta^*}{\pi_s(\eta^*, \Upsilon_\infty) S_\infty^{-1} \pi_s(\eta^*, \Upsilon_\infty)} > q(1 - \alpha, \Upsilon_\infty) \mid \Upsilon_\infty \right) \\ &= \mathbb{P}(W(\eta^*, \Upsilon_\infty) > q(1 - \alpha, \Upsilon_\infty) \mid \Upsilon_\infty) \\ &= \alpha \quad \text{a.s.} \end{aligned}$$

The second line holds by the observation that  $\eta^*$  and  $S_\infty^{-1/2} \mathcal{Z}_{\mathbf{g}}$  are both independent of  $\Upsilon_\infty$ , and that  $\mathcal{Z}_\pi = \pi_s(S_\infty^{-1/2} \mathcal{Z}_{\mathbf{g}}, \Upsilon_\infty)$ , which has the same distribution as  $\pi_s(\eta^*, \Upsilon_\infty)$ . The third line holds by the definition of  $W(\eta^*, \Upsilon_\infty)$ . The last line holds by the definition of  $q(1 - \alpha, \Upsilon)$ , and the fact that  $W(\eta^*, \Upsilon)$  is continuously distributed for any  $\Upsilon \in \mathcal{U}$ . By taking the law of total probability, it follows that the unconditional rejection probability equals  $\alpha$  as well. This completes the proof for the case  $s_\infty < \infty$ .

**Case 2:**  $s_\infty = \infty$ .

In this case, by the assumption  $\sqrt{n}\|\pi_{F_n}\| \rightarrow \infty$ , we have

$$\begin{aligned} \frac{\hat{\pi}}{\|\pi_{F_n}\|} &= \frac{\pi_{F_n}}{\|\pi_{F_n}\|} + O_p(n^{-1/2}\|\pi_{F_n}\|^{-1}) \\ &= \iota_\infty + o_p(1). \end{aligned} \quad (\text{A.5})$$

and

$$\begin{aligned} \frac{\hat{h}(\lambda_n)}{\sqrt{n}\|\pi_{F_n}\|} &= \frac{\hat{\pi}}{\|\pi_{F_n}\|} - [\partial_p \pi] \hat{\Sigma}_p [\partial_p \hat{g}(\lambda_n)] \hat{S}^{-1} \sqrt{n} \hat{g}(\lambda_n) \|\sqrt{n} \pi_{F_n}\|^{-1} \\ &= \iota_\infty + o_p(1). \end{aligned} \quad (\text{A.6})$$

From the condition  $\iota_\infty \neq 0_{K \times 1}$ , (A.1), and (A.5), it follows that

$$W_n(\lambda_n) = \frac{n \hat{g}(\lambda_n)' \hat{S}(\lambda_n)^{-1} (\hat{\pi}' / \|\pi_{F_n}\|) (\hat{\pi}' / \|\pi_{F_n}\|) \hat{S}(\lambda_n)^{-1} \hat{g}(\lambda_n)}{(\hat{\pi}' / \|\pi_{F_n}\|) \hat{S}(\lambda_n)^{-1} (\hat{\pi}' / \|\pi_{F_n}\|)} \xrightarrow{d} \chi_1^2. \quad (\text{A.7})$$

Now we examine the stochastic behavior of critical value  $q_{W^*}(1 - \alpha)$ , defined as  $(1 - \alpha)$ -quantile of  $W_n^*(\lambda_n)$  conditional on data. First, we define the normalized conditioning statistic in the construction of  $W_n^*(\lambda_n)$ :

$$\bar{\Upsilon} \equiv \left( \frac{\hat{h}(\lambda_n)}{\sqrt{n}\|\pi_{F_n}\|}, \hat{S}, \hat{\Sigma}_p, \frac{\partial_p \hat{g}(\lambda_n)}{\sqrt{n}\|\pi_{F_n}\|} \right).$$

By convergence result (A.1), (A.6), and the condition that  $\sqrt{n}\|\pi_{F_n}\| \rightarrow \infty$ , it can be seen that

$$\bar{\Upsilon} \xrightarrow{p} \bar{\Upsilon}_\infty \equiv (\iota_\infty, S_\infty, \Sigma_{p,\infty}, 0_{K \times (K+1)}).$$

This and the continuity of the quantile function  $q(1 - \alpha, \cdot)$  give  $q(1 - \alpha, \bar{\Upsilon}) \xrightarrow{p} q(1 - \alpha, \bar{\Upsilon}_\infty)$ . This shows that the  $(1 - \alpha)$ -quantile of  $W(\eta^*, \bar{\Upsilon})$  converges in probability to  $q(1 - \alpha, \bar{\Upsilon}_\infty)$ , which equals the  $(1 - \alpha)$ -quantile of  $\chi_1^2$  distribution by replacing  $\hat{\Upsilon}$  with  $\bar{\Upsilon}_\infty$  in (A.3). Note that  $W_n^*(\lambda_n) = W(\eta^*, \bar{\Upsilon})$ . From this, we conclude that

$$q_{W^*}(1 - \alpha) = q(1 - \alpha, \bar{\Upsilon}) \xrightarrow{p} q_{\chi_1^2}(1 - \alpha). \quad (\text{A.8})$$

By (A.7) and (A.8), we have  $W_n(\lambda_n) - q_{W^*}(1 - \alpha) \xrightarrow{d} \chi_1^2 - q_{\chi_1^2}(1 - \alpha)$ . Then the desired conclusion follows by the definition of convergence in distribution.  $\square$

### A.1.3 Proof of uniform validity in Section 2.4

The proof of uniform size control in Theorem 2.4.1 uses the sub-sequencing techniques from D. W. Andrews et al., 2020. Specifically, we verify part of their Assumption B in the following Proposition A.1.2.

**Proposition A.1.2.** *For any subsequence  $\{p_n\}$  of  $\{n\}$  and any sequence  $\{(\lambda_{p_n}, F_{p_n}) \in \mathcal{P}_0\}$  for which*

1.  $\theta_{F_{p_n}} \rightarrow \theta_\infty \in \Theta$ , where  $\theta_{F_{p_n}}$  is a sequence of parameters such that  $(\theta_{F_{p_n}}, F_{p_n}) \in \mathcal{P}$  and  $\lambda_{p_n} = c'\theta_{F_{p_n}}$ ;
2.  $\beta_{F_{p_n}, d\ell} \rightarrow \beta_{\infty, d\ell}$  for  $d = 0, 1$  and  $\ell = 0, 1, \dots, K$ ;
3.  $\sigma_{F_{p_n}, d\ell}^2 \rightarrow \sigma_{\infty, d\ell}^2$  for  $d = 0, 1$  and  $\ell = 0, 1, \dots, K$ ;
4.  $p_{F_{p_n}}(z_\ell) \rightarrow p_\infty(z_\ell)$  for all  $\ell = 0, 1, \dots, K$ ;
5.  $q_{F_{p_n}}(z_\ell) \rightarrow q_\infty(z_\ell)$  for all  $\ell = 0, 1, \dots, K$ ;
6.  $B_{F_{p_n}} \rightarrow B_\infty$  and  $C_{F_{p_n}} \rightarrow C_\infty$ ;
7.  $\sqrt{p_n}\tau_{j, F_{p_n}} \rightarrow t_j \in [0, \infty]$  for all  $j = 1, \dots, 2(M + 1)$ ;
8.  $\iota_n \equiv \frac{S_{F_{p_n}} B'_{F_{p_n}} c}{\|S_{F_{p_n}} B'_{F_{p_n}} c\|} \rightarrow \iota_\infty \in \mathbb{R}^{2(M+1)}$ , where  $S_{F_{p_n}}$  is defined in equation (A.22)<sup>1</sup>.

we have

$$\limsup_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} \left( \inf_{c'\theta = \lambda_{p_n}} MLC_{p_n}(\theta) > q_{(1+\alpha)\chi_1^2 + \alpha\chi_{2K+1}^2}(1 - \alpha) \right) \leq \alpha.$$

Based on the above proposition, we show the uniform size control in the following proof using the similar arguments as in D. W. Andrews et al. (2020, Theorem 2.1).

*Proof of Theorem 2.4.1.* The proof is similar to Proposition 2.3.1 but the asymptotic ex-

<sup>1</sup> Note that the denominator  $\|S_{F_n} B'_{F_n} c\| \neq 0$  since both  $S_{F_n}$  and  $B_{F_n}$  are full-rank matrices and the weight  $c$  is nonzero. Hence  $\iota_n$  is properly defined for each  $n \geq 1$  and satisfies  $\|\iota_n\| = 1$ .

actness and similarity are not preserved. Let  $\{a_n\}$  be a subsequence of  $\{n\}$  such that

$$\begin{aligned} & \lim_{n \rightarrow \infty} \mathbb{P}_{F_{a_n}} \left( \inf_{c'\theta = \lambda_{a_n}} \text{MLC}_{a_n}(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right) \\ &= \limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F \left( \inf_{c'\theta = \lambda} \text{MLC}_n(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right). \end{aligned}$$

Such sequence always exists. Along the sequence  $\{a_n\}$ , we can choose a subsequence  $\{p_n\} \subset \{a_n\}$  such that the conditions 1-8 in Proposition A.1.2 hold for some

$$\vartheta_\infty \equiv \left( \theta_\infty, \beta_\infty, \{\sigma_{\infty, d\ell}^2\}_{d, \ell}, p_\infty, q_\infty, B_\infty, C_\infty, \{t_j\}_{j=1}^{2(M+1)}, \iota_\infty \right).$$

That is,

$$\vartheta_{p_n} \equiv \left( \theta_{F_{p_n}}, \beta_{F_{p_n}}, \{\sigma_{F_{p_n}, d\ell}^2\}_{d, \ell}, p_{F_{p_n}}, q_{F_{p_n}}, B_{F_{p_n}}, C_{F_{p_n}}, \{\sqrt{p_n} \tau_{j, F_{p_n}}\}_{j=1}^{2(M+1)}, \iota_{p_n} \right) \rightarrow \vartheta_\infty.$$

Such converging subsequence  $\{\vartheta_{p_n}\}$  always exists since each element of  $\vartheta_{a_n}$  is contained in a compact set as restricted by the parameter space  $\mathcal{P}$ .<sup>2</sup> By Proposition A.1.2, it follows that

$$\limsup_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} \left( \inf_{c'\theta = \lambda_{p_n}} \text{MLC}_{p_n}(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right) \leq \alpha.$$

Since  $\{p_n\}$  is a subsequence of  $\{a_n\}$ , and the rejection probability converges along the sequence  $\{a_n\}$ , we have

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \sup_{(\lambda, F) \in \mathcal{P}_0} \mathbb{P}_F \left( \inf_{c'\theta = \lambda} \text{MLC}_n(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right) \\ &= \limsup_{n \rightarrow \infty} \mathbb{P}_{F_{p_n}} \left( \inf_{c'\theta = \lambda_{p_n}} \text{MLC}_{p_n}(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right) \\ &\leq \alpha. \end{aligned}$$

Then the proof is complete. □

---

<sup>2</sup> Note that  $\{t_j\}$  belongs to a compact set  $[0, \infty]^{2(M+1)}$ , and the space of orthogonal matrices is also compact.

*Proof of Proposition A.1.2.* For simplicity of notations, the proof is shown for the full sequence  $\{n\}$ . Then, we note that the same proof goes through with  $p_n$  in place of  $n$ .

By Lemma A.1.1(a), we have

$$\sqrt{n} \begin{pmatrix} \hat{p} - p_{F_n} \\ \hat{\beta} - \beta_{F_n} \end{pmatrix} = \begin{bmatrix} \hat{Z}_p \\ \hat{Z}_\beta \end{bmatrix} \xrightarrow{d} \mathcal{N} \left( \mathbf{0}_{3(K+1) \times 1}, \begin{bmatrix} \Sigma_{p, \infty} & \mathbf{0}_{(K+1) \times 2(K+1)} \\ \mathbf{0}_{2(K+1) \times (K+1)} & \Sigma_{\beta, \infty} \end{bmatrix} \right)$$

This implies

$$\sqrt{n}(\hat{A} - A_{F_n}) = \begin{bmatrix} (L_0(p_{F_n})\hat{Z}_p, \dots, L_M(p_{F_n})\hat{Z}_p) & \mathbf{0}_{(K+1) \times (M+1)} \\ \mathbf{0}_{(K+1) \times (M+1)} & (R_0(p_{F_n})\hat{Z}_p, \dots, R_M(p_{F_n})\hat{Z}_p) \end{bmatrix} + o_p(1) \quad (\text{A.9})$$

where

$$L_m(p) = \text{diag}\{\lambda'_{1m}(p(z_0)), \dots, \lambda'_{1m}(p(z_K))\}$$

$$R_m(p) = \text{diag}\{\lambda'_{0m}(p(z_0)), \dots, \lambda'_{0m}(p(z_K))\}.$$

By continuous mapping theorem, we have

$$\begin{aligned} \sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) &= \sqrt{n}(\hat{A} - A_{F_n})\theta_{F_n} - \sqrt{n}(\hat{\beta} - \beta_{F_n}) \\ &= \begin{bmatrix} (L_0(p_{F_n})\hat{Z}_p, \dots, L_M(p_{F_n})\hat{Z}_p) & \mathbf{0}_{(K+1) \times (M+1)} \\ \mathbf{0}_{(K+1) \times (M+1)} & (R_0(p_{F_n})\hat{Z}_p, \dots, R_M(p_{F_n})\hat{Z}_p) \end{bmatrix} \begin{bmatrix} \theta_{1, F_n} \\ \theta_{0, F_n} \end{bmatrix} \\ &\quad - \hat{Z}_\beta + o_p(1) \\ &\xrightarrow{d} \begin{bmatrix} \text{diag} \left\{ \sum_{m=0}^M \theta_{1m, \infty} \lambda'_{1m}(p_\infty(z_\ell)) : \ell = 0, 1, \dots, K \right\} \\ \text{diag} \left\{ \sum_{m=0}^M \theta_{0m, \infty} \lambda'_{0m}(p_\infty(z_\ell)) : \ell = 0, 1, \dots, K \right\} \end{bmatrix} \mathcal{Z}_p - \mathcal{Z}_\beta \\ &= H(p_\infty, \theta_\infty) \mathcal{Z}_p - \mathcal{Z}_\beta. \end{aligned}$$

The fourth line holds by definition:

$$H(p, \theta) \equiv \begin{bmatrix} \text{diag} \left\{ \sum_{m=0}^M \theta_{1m} \lambda'_{1m}(p(z_\ell)) : \ell = 0, 1, \dots, K \right\} \\ \text{diag} \left\{ \sum_{m=0}^M \theta_{0m} \lambda'_{0m}(p(z_\ell)) : \ell = 0, 1, \dots, K \right\} \end{bmatrix}.$$

Let

$$\mathcal{Z}_m \equiv H(p_\infty, \theta_\infty) \mathcal{Z}_p - \mathcal{Z}_\beta.$$

Following Lemma A.1.1, the condition  $\theta_{F_n} \rightarrow \theta_\infty$ , and note that  $\{\lambda'_{dm}(\cdot)\}$  are continuously differentiable by the continuity of  $h_m(\cdot)$ , continuous mapping theorem implies

$$\begin{aligned}\hat{\Omega}(\theta_{F_n}) &= H(\hat{p}, \theta_{F_n}) \hat{\Sigma}_p H(\hat{p}, \theta_{F_n})' + \hat{\Sigma}_\beta \\ &\xrightarrow{p} H(p_\infty, \theta_\infty) \Sigma_{p,\infty} H(p_\infty, \theta_\infty)' + \Sigma_{\beta,\infty} \\ &= \text{var}(\mathcal{Z}_m)\end{aligned}\tag{A.10}$$

and

$$\begin{aligned}\hat{\Gamma}_j(\theta_{F_n}) &= M_j(\hat{p}) \hat{\Sigma}_p H(\hat{p}, \theta_{F_n})' \\ &\xrightarrow{p} M_j(p_\infty) \Sigma_{p,\infty} H(p_\infty, \theta_\infty)' \\ &= \text{cov}(M_j(p_\infty) \mathcal{Z}_p, \mathcal{Z}_m).\end{aligned}\tag{A.11}$$

In particular,  $\text{var}(\mathcal{Z}_m)$  is positive definite by the positive definiteness of  $\Sigma_{\beta,\infty}$ , which is in turn implied by the parameter space restriction on  $\mathcal{P}$ . For each  $j = 1, \dots, 2(M+1)$ , then we have

$$\begin{aligned}\sqrt{n}(\hat{d}_j(\theta_{F_n}) - a_{j,F_n}) &= \sqrt{n}(\hat{a}_j - a_{j,F_n}) - \hat{\Gamma}_j(\theta_{F_n}) \hat{\Omega}(\theta_{F_n})^{-1} \sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) \\ &\xrightarrow{d} \underbrace{M_j(p_\infty) \mathcal{Z}_p - \text{cov}(M_j(p_\infty) \mathcal{Z}_p, \mathcal{Z}_m) \text{var}(\mathcal{Z}_m)^{-1} \mathcal{Z}_m}_{\equiv \mathcal{Z}_{d_j}},\end{aligned}$$

where the second line holds by combining results from (A.9), (A.10), and (A.11). Then we observe that

$$\text{cov}(\mathcal{Z}_{d_j}, \mathcal{Z}_m) = 0_{2(K+1) \times 2(K+1)}.$$

This shows

$$\sqrt{n} \begin{pmatrix} \hat{A}\theta_{F_n} - \hat{\beta} \\ \text{vec}(\hat{D}(\theta_{F_n})) - \text{vec}(A_{F_n}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{Z}_m \\ \text{vec}(\mathcal{Z}_D) \end{pmatrix},\tag{A.12}$$

where  $\mathcal{Z}_D = (\mathcal{Z}_{d_1}, \dots, \mathcal{Z}_{d_{2(M+1)}})$  is independent of  $\mathcal{Z}_m$ .

Based on (A.12), applying Lemma A.1.5 yields

$$n^{1/2} \tilde{D}(\theta_{F_n}) B_{F_n} S_{F_n} \xrightarrow{d} \mathcal{D}_\xi$$

where  $\mathcal{D}_\xi$  is of full column rank with probability one and is also independent of  $\mathcal{Z}_m$ , and  $S_{F_n}$  is a diagonal matrix with positive diagonal elements

$$S_{F_n} = \begin{bmatrix} \left( (\sqrt{n}\tau_{1,F_n})^{-1} & & \\ & \ddots & \\ & & (\sqrt{n}\tau_{q,F_n})^{-1} \right) & 0_{q \times (2(M+1)-q)} \\ 0_{(2(M+1)-q) \times q} & & & I_{2(M+1)-q} \end{bmatrix}.$$

Define a normalizing constant  $\gamma_n \equiv \|\sqrt{n}S_{F_n}B'_{F_n}c\|^{-1}$ . By continuous mapping theorem and given that  $\mathcal{D}_\xi$  is of full rank with probability one, we have

$$\begin{aligned} \hat{Q}(\theta_{F_n})\gamma_n &= \hat{\Omega}(\theta_{F_n})^{-1/2} \left( \sqrt{n}\tilde{D}(\theta_{F_n})B_{F_n}S_{F_n} \right) \\ &\quad \left[ \left( \sqrt{n}\tilde{D}(\theta_{F_n})B_{F_n}S_{F_n} \right)' \hat{\Omega}(\theta_{F_n})^{-1} \left( \sqrt{n}\tilde{D}(\theta_{F_n})B_{F_n}S_{F_n} \right) \right]^{-1} \underbrace{\frac{\sqrt{n}S_{F_n}B'_{F_n}c}{\|\sqrt{n}S_{F_n}B'_{F_n}c\|}}_{\iota_n} \\ &\xrightarrow{d} \mathcal{Q} \equiv \text{var}(\mathcal{Z}_m)^{-1/2} \mathcal{D}'_\xi [\mathcal{D}'_\xi \text{var}(\mathcal{Z}_m)^{-1} \mathcal{D}_\xi]^{-1} \iota_\infty. \end{aligned} \quad (\text{A.13})$$

Since  $\iota_\infty \neq 0_{2(M+1) \times 1}$ , we note  $\mathcal{Q}$  is nonzero almost surely and independent of  $\mathcal{Z}_m$  by the independence between  $\mathcal{D}_\xi$  and  $\mathcal{Z}_m$ .

Then it follows that

$$\begin{aligned} \text{MRLM}_n(\theta_{F_n}) &= \sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})' \hat{\Omega}(\theta_{F_n})^{-1/2} P_{\hat{Q}(\theta_{F_n})} \hat{\Omega}(\theta_{F_n})^{-1/2} \sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) \\ &= \sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})' \hat{\Omega}(\theta_{F_n})^{-1/2} P_{\hat{Q}(\theta_{F_n})\gamma_n} \hat{\Omega}(\theta_{F_n})^{-1/2} \sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) \\ &\xrightarrow{d} \mathcal{Z}'_m \text{var}(\mathcal{Z}_m)^{-1/2} P_{\mathcal{Q}} \text{var}(\mathcal{Z}_m)^{-1/2} \mathcal{Z}_m \\ &\sim \chi_1^2 \end{aligned}$$

The third line holds by continuous mapping theorem based on (A.10), (A.12), and (A.13).

The last line follows by that  $\mathcal{Q}$  is nonzero almost surely and  $\mathcal{Q} \perp \mathcal{Z}_m$ , thus

$$\mathcal{Z}'_m \text{var}(\mathcal{Z}_m)^{-1/2} P_{\mathcal{Q}} \text{var}(\mathcal{Z}_m)^{-1/2} \mathcal{Z}_m \sim \chi_1^2 \quad \text{conditional on } \mathcal{Q}$$

implies the unconditional distribution

$$\mathcal{Z}'_m \text{var}(\mathcal{Z}_m)^{-1/2} P_{\mathcal{Q}} \text{var}(\mathcal{Z}_m)^{-1/2} \mathcal{Z}_m \sim \chi_1^2.$$

We apply similar arguments to the difference between AR and MRLM statistics, yielding

$$\begin{bmatrix} \text{AR}_n(\theta_{F_n}) - \text{MRLM}_n(\theta_{F_n}) \\ \text{MRLM}_n(\theta_{F_n}) \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \mathbf{Z}'_m \text{var}(\mathbf{Z}_m)^{-1/2} M_{\mathcal{Q}} \text{var}(\mathbf{Z}_m)^{-1/2} \mathbf{Z}_m \\ \mathbf{Z}'_m \text{var}(\mathbf{Z}_m)^{-1/2} P_{\mathcal{Q}} \text{var}(\mathbf{Z}_m)^{-1/2} \mathbf{Z}_m \end{bmatrix} \sim \begin{bmatrix} \chi_{2K+1}^2 \\ \chi_1^2 \end{bmatrix},$$

where  $M_{\mathcal{Q}} \equiv I - P_{\mathcal{Q}}$  denotes the annihilator operator. Note that  $\chi_{2K+1}^2$  is independent of  $\chi_1^2$  because  $(P_{\mathcal{Q}} \text{var}(\mathbf{Z}_m)^{-1/2} \mathbf{Z}_m, M_{\mathcal{Q}} \text{var}(\mathbf{Z}_m)^{-1/2} \mathbf{Z}_m)$  are uncorrelated:

$$\begin{aligned} & \text{cov}(P_{\mathcal{Q}} \text{var}(\mathbf{Z}_m)^{-1/2} \mathbf{Z}_m, M_{\mathcal{Q}} \text{var}(\mathbf{Z}_m)^{-1/2} \mathbf{Z}_m) \\ &= \mathbb{E} \left( \text{cov}(P_{\mathcal{Q}} \text{var}(\mathbf{Z}_m)^{-1/2} \mathbf{Z}_m, M_{\mathcal{Q}} \text{var}(\mathbf{Z}_m)^{-1/2} \mathbf{Z}_m \mid \mathcal{Q}) \right) \\ &= \mathbb{E}(P_{\mathcal{Q}} M_{\mathcal{Q}}) \\ &= \mathbf{0}_{2(K+1) \times 2(K+1)}. \end{aligned}$$

The first line holds by the law of total covariance and the fact that  $\mathbb{E}[\mathbf{Z}_m \mid \mathcal{Q}] = \mathbb{E}[\mathbf{Z}_m] = \mathbf{0}$ . The second line holds by the independence between  $\mathcal{Q}$  and  $\mathbf{Z}_m$ . The third line holds by the definition  $M_{\mathcal{Q}} = I - P_{\mathcal{Q}}$ . So we have

$$\text{MLC}_n(\theta_{F_n}) = \text{MRLM}_n(\theta_{F_n}) + a \cdot \text{AR}_n(\theta_{F_n}) \xrightarrow{d} (1+a)\chi_1^2 + a\chi_{2K+1}^2$$

by continuous mapping theorem.

As a result, we show that

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \mathbb{P}_{F_n} \left( \inf_{c' \theta = \lambda_n} \text{MLC}_n(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right) \\ & \leq \lim_{n \rightarrow \infty} \mathbb{P}_{F_n} \left( \text{MLC}_n(\theta_{F_n}) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1 - \alpha) \right) \\ & = \alpha. \end{aligned}$$

Then the proof is complete. □

### A.1.4 Proof of bias formula in section 2.5.1

*Proof of Lemma 2.5.1.* Applying FWL theorem to partial out the linear effects of  $W$  in the misspecified regression (2.24), we have

$$\begin{aligned} Y^{\perp W|D=1} &= \tilde{\rho}_1 \lambda_1(P)^{\perp W|D=1} + Y^{\perp W, \lambda_1(P)|D=1} && \text{conditional on } D = 1 \\ Y^{\perp W|D=0} &= \tilde{\rho}_0 \lambda_0(P)^{\perp W|D=0} + Y^{\perp W, \lambda_0(P)|D=0} && \text{conditional on } D = 0 \end{aligned}$$

Conditional on  $D = d$ , simple OLS regression gives

$$\begin{aligned} \tilde{\rho}_d &= \frac{\text{cov}(Y^{\perp W|D=d}, \lambda_d(P)^{\perp W|D=d} \mid D = d)}{\text{var}(\lambda_d(P)^{\perp W|D=d} \mid D = d)} \\ &= \frac{\text{cov}((\rho_d \lambda_d(P) + W' \eta_d \lambda_d(P))^{\perp W|D=d}, \lambda_d(P)^{\perp W|D=d} \mid D = d)}{\text{var}(\lambda_d(P)^{\perp W|D=d} \mid D = d)} \\ &= \rho_d + \frac{\text{cov}((W' \lambda_d(P))^{\perp W|D=d}, \lambda_d(P)^{\perp W|D=d} \mid D = d)}{\text{var}(\lambda_d(P)^{\perp W|D=d} \mid D = d)} \eta_d. \end{aligned}$$

The second line holds by noting that correct regression equation (2.23) gives

$$Y = \mu_d + W' \tau_d + \rho_d \lambda_d(P) + W' \eta_d \lambda_d(P) + \epsilon_d \quad \text{conditional on } D = d.$$

where  $\epsilon_d \equiv Y - \mathbb{E}[Y \mid D = d, W, P]$ . This implies

$$Y^{\perp W|D=d} = (\rho_d \lambda_d(P) + W' \eta_d \lambda_d(P))^{\perp W|D=d}$$

since  $\epsilon_d$  is uncorrelated with  $W$  conditional on  $D = d$ . So we establish the bias formula for  $\tilde{\rho}_d - \rho_d$ .

Alternatively, we can also apply the FWL theorem to partial out the linear effects of  $\lambda_d(W)$  in the misspecified regression, giving

$$\begin{aligned} Y^{\perp \lambda_1(P)|D=1} &= \tilde{\tau}'_1 W^{\perp \lambda_1(P)|D=1} + Y^{\perp W, \lambda_1(P)|D=1} && \text{conditional on } D = 1 \\ Y^{\perp \lambda_0(P)|D=0} &= \tilde{\tau}'_0 W^{\perp \lambda_0(P)|D=0} + Y^{\perp W, \lambda_0(P)|D=0} && \text{conditional on } D = 0 \end{aligned}$$

In this way,  $\tilde{\tau}_d$  is the OLS coefficient on  $W$  in a regression of  $Y^{\perp \lambda_d(P)|D=d}$  on  $W^{\perp \lambda_d(P)|D=d}$ :

$$\tilde{\tau}_d = \mathbb{E}[(W^{\perp \lambda_d(P)|D=d})(W^{\perp \lambda_d(P)|D=d})' \mid D = d]^{-1} \mathbb{E}[(W^{\perp \lambda_d(P)|D=d})(Y^{\perp \lambda_d(P)|D=d}) \mid D = d].$$

The correct regression (2.23) implies that  $Y^{\perp\lambda_d(P)|D=d} = (W'\tau_d + W'\eta_d\lambda_d(P))^{\perp\lambda_d(P)|D=d}$ .

Plugging it into the OLS estimand  $\tilde{\tau}_d$  then gives

$$\begin{aligned}\tilde{\tau}_d - \tau_d &= \mathbb{E}[(W^{\perp\lambda_d(P)|D=d})(W^{\perp\lambda_d(P)|D=d})' | D = d]^{-1} \\ &\quad \times \mathbb{E}[(W^{\perp\lambda_1(P)|D=d})(W'\lambda_d(P))^{\perp\lambda_d(P)|D=d} | D = d] \eta_d \\ &= \mathbb{E}[(W^{\perp\lambda_d(P)|D=d})(W^{\perp\lambda_d(P)|D=d})' | D = d]^{-1} \\ &\quad \times \mathbb{E}[(W^{\perp\lambda_1(P)|D=d})(W'\lambda_d(P)) | D = d] \eta_d\end{aligned}$$

which establishes the bias formula for  $\tilde{\tau}_d - \tau_d$ .  $\square$

*Proof of Theorem 2.5.1.* We begin by noting that  $\text{cov}(W, \lambda_d(P) | D = d) = 0_{L \times 1}$  implies

$$W^{\perp\lambda_d(P)|D=d} = W - \mathbb{E}[W | D = d] \quad \text{and} \quad \lambda_d(P)^{\perp W|D=d} = \lambda_d(P) - \mathbb{E}[\lambda_d(P) | D = d]. \quad (\text{A.14})$$

From  $\text{cov}(WW', \lambda_d(P) | D = d) = 0_{L \times L}$ , it follows that

$$\begin{aligned}& [W\lambda_d(P)]^{\perp W|D=d} \\ &= W\lambda_d(P) - \mathbb{E}[W\lambda_d(P) | D = d] \\ &\quad - \mathbb{E}[\lambda_d(P)W(W - \mathbb{E}[W | D = d])' | D = d] \text{var}(W | D = d)^{-1}(W - \mathbb{E}[W | D = d]) \\ &= W\lambda_d(P) - \mathbb{E}[W | D = d]\mathbb{E}[\lambda_d(P) | D = d] \\ &\quad - (\mathbb{E}[\lambda_d(P)WW' | D = d] - \mathbb{E}[\lambda_d(P)W | D = d]\mathbb{E}[W | D = d]') \\ &\quad \quad \times \text{var}(W | D = d)^{-1}(W - \mathbb{E}[W | D = d]) \\ &= W\lambda_d(P) - \mathbb{E}[W | D = d]\mathbb{E}[\lambda_d(P) | D = d] \\ &\quad - (\mathbb{E}[\lambda_d(P) | D = d](\mathbb{E}[WW' | D = d] - \mathbb{E}[W | D = d]\mathbb{E}[W | D = d]')) \\ &\quad \quad \times \text{var}(W | D = d)^{-1}(W - \mathbb{E}[W | D = d]) \\ &= W\lambda_d(P) - \mathbb{E}[W | D = d]\mathbb{E}[\lambda_d(P) | D = d] - \mathbb{E}[\lambda_d(P) | D = d](W - \mathbb{E}[W | D = d]) \\ &= W(\lambda_d(P) - \mathbb{E}[\lambda_d(P) | D = d]). \quad (\text{A.15})\end{aligned}$$

The first equality follows by definition of regression residuals. The second equality follows by the assumption  $\text{cov}(W, \lambda_d(P) | D = d) = 0_{L \times 1}$ . The third equality follows by the assumption  $\text{cov}(W, \lambda_d(P) | D = d) = 0_{L \times 1}$  and  $\text{cov}(WW', \lambda_d(P) | D = d) = 0_{L \times L}$ .

First, we simplify the bias formula in Lemma 2.5.1, which would lead us to compute the bias on causal parameter of interests. The general bias formula on  $\tau_d$  can be simplified by the following derivations:

$$\begin{aligned}
\tilde{\tau}_d &= \tau_d + \mathbb{E}[(W^{\perp\lambda_d(P)|D=d})(W^{\perp\lambda_d(P)|D=d})' | D = d]^{-1} \mathbb{E}[(W^{\perp\lambda_1(P)|D=d})(W' \lambda_d(P)) | D = d] \eta_d \\
&= \tau_d + \text{var}(W | D = d)^{-1} \mathbb{E}[(W - \mathbb{E}[W | D = d])W' \lambda_d(P) | D = d] \eta_d \\
&= \tau_d + \text{var}(W | D = d)^{-1} (\mathbb{E}[WW' \lambda_d(P) | D = d] - \mathbb{E}[W | D = d] \mathbb{E}[W' \lambda_1(P) | D = d]) \eta_d \\
&= \tau_d + \text{var}(W | D = d)^{-1} \text{var}(W | D = d) \mathbb{E}[\lambda_d(P) | D = d] \eta_d \\
&= \tau_d + \mathbb{E}[\lambda_d(P) | D = d] \eta_d.
\end{aligned} \tag{A.16}$$

The first equality is given by the Lemma 2.5.1. The second equality holds by (A.14) and (A.15). The fourth equality holds by assumption  $\text{cov}(WW', \lambda_d(P) | D = d) = 0_{L \times L}$  and  $\text{cov}(W, \lambda_d(P) | D = d) = 0_{L \times 1}$ .

On the other hand, the general bias formula on  $\rho_d$  can be simplified by the following derivations:

$$\begin{aligned}
\tilde{\rho}_d &= \rho_d + \frac{\text{cov}((W' \lambda_d(P))^{\perp W|D=d}, \lambda_d(P)^{\perp W|D=d} | D = d)}{\text{var}(\lambda_d(P)^{\perp W|D=d} | D = d)} \eta_d \\
&= \rho_d + \frac{\text{cov}(W'(\lambda_d(P) - \mathbb{E}[\lambda_d(P) | D = d]), \lambda_d(P) - \mathbb{E}[\lambda_d(P) | D = d] | D = d)}{\text{var}(\lambda_d(P) | D = d)} \eta_d \\
&= \rho_d + \frac{\mathbb{E}[W' \lambda_1(P)^2 | D = d] - \mathbb{E}[W' | D = d] \mathbb{E}[\lambda_d(P) | D = d]^2}{\text{var}(\lambda_d(P) | D = d)} \eta_d \\
&= \rho_d + \mathbb{E}[W' \eta_d | D = d].
\end{aligned} \tag{A.17}$$

The first equality is given by the Lemma 2.5.1. The second equality holds by (A.14) and (A.15). The third equality holds by assumption that  $\text{cov}(W, \lambda_d(P) | D = d) = 0_{L \times 1}$ . The final equality holds by the assumption that  $\text{cov}(W, \lambda_d(P)^2) = 0_{L \times 1}$ .

Having simplified the bias formulae in Lemma 2.5.1, now we can derive the bias formula

for  $\mu_d$  as follows:

$$\begin{aligned}
\tilde{\mu}_d &= \mathbb{E}[Y - W'\tilde{\tau}_d - \tilde{\rho}_d\lambda_d(P) \mid D = d] \\
&= \mathbb{E}[\mu_d + W'\tau_d + \rho_d\lambda_d(P) + [W\lambda_d(P)]'\eta_d - W'\tilde{\tau}_d - \tilde{\rho}_d\lambda_d(P) \mid D = d] \\
&= \mathbb{E}[\mu_d + [W\lambda_d(P)]'\eta_d - W'\eta_d\mathbb{E}[\lambda_d(P) \mid D = d] - \mathbb{E}[W'\eta_d \mid D = d]\lambda_d(P) \mid D = d] \\
&= \mu_d - \mathbb{E}[W'\eta_d \mid D = d]\mathbb{E}[\lambda_d(P) \mid D = d]
\end{aligned} \tag{A.18}$$

The first equality follows by misspecified regression (2.24). The second equality follows by the correctly specified regression (2.23). The third equality follows from the bias formula in (A.16) and (A.17). The last equality holds by the assumption  $\text{cov}(W, \lambda_d(P) \mid D = d) = \mathbf{0}_{L \times 1}$ .

Given the bias formula in (A.16), (A.17), and (A.18), now we compute the bias on estimating the slope of MTE curve:

$$\begin{aligned}
\widetilde{\text{Slope}} - \text{Slope} &= [\tilde{\rho}_1 - \tilde{\rho}_0] - \mathbb{E}[\rho_1(W) - \rho_0(W)] \\
&= \mathbb{E}[W \mid D = 1]'\eta_1 - \mathbb{E}[W \mid D = 0]'\eta_0 - \mathbb{E}[W]'\eta_1 + \mathbb{E}[W]'\eta_0 \\
&= (\mathbb{E}[W \mid D = 1] - \mathbb{E}[W \mid D = 0])'(\mathbb{P}(D = 0)\eta_1 + \mathbb{P}(D = 1)\eta_0).
\end{aligned}$$

The second line holds by (A.17). This establishes the bias formula for the slope of MTE curve.

Next note that (A.16) and (A.18) imply

$$\begin{aligned}
\widetilde{\text{CATE}} - \text{CATE} &= [\tilde{\mu}_1 - \tilde{\mu}_0] + w'(\tilde{\tau}_1 - \tilde{\tau}_0) - [\mu_1 - \mu_0] - w'(\tau_1 - \tau_0) \\
&= (w - \mathbb{E}[W \mid D = 1])'\eta_1 \times \mathbb{E}[\lambda_1(P) \mid D = 1] \\
&\quad - (w - \mathbb{E}[W \mid D = 0])'\eta_0 \times \mathbb{E}[\lambda_0(P) \mid D = 0].
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{\text{ATE}} - \text{ATE} &= [\tilde{\mu}_1 - \tilde{\mu}_0] + \mathbb{E}[W]'\eta_1(\tilde{\tau}_1 - \tilde{\tau}_0) - [\mu_1 - \mu_0] - \mathbb{E}[W]'\eta_1(\tau_1 - \tau_0) \\
&= (\mathbb{E}[W] - \mathbb{E}[W \mid D = 1])'\eta_1 \times \mathbb{E}[\lambda_1(P) \mid D = 1] \\
&\quad - (\mathbb{E}[W] - \mathbb{E}[W \mid D = 0])'\eta_0 \times \mathbb{E}[\lambda_0(P) \mid D = 0].
\end{aligned}$$

So we have established the bias formula for ATE, CATE, and slope of MTE curve under Assumption 7.1. From this, note that the bias on ATE and slope would vanish if we additionally impose Assumption 7.2, i.e.,  $\mathbb{E}[W] = \mathbb{E}[W \mid D = 1] = \mathbb{E}[W \mid D = 0]$ .  $\square$

### A.1.5 Lemmas for main results

**Lemma A.1.1.** *For a sequence of distributions  $\{(\theta_{F_n}, F_n)\}_{n=1}^\infty \subset \mathcal{P}$ , suppose*

1.  $q_{F_n}(z_\ell) \rightarrow q_\infty(z_\ell)$  for each  $\ell = 0, 1, \dots, K$ ,
2.  $p_{F_n} \rightarrow p_\infty$ ,
3.  $\beta_{F_n} = (\beta'_{F_n,1}, \beta'_{F_n,0})' \rightarrow \beta_\infty$ ,
4.  $\sigma_{F_n,d\ell}^2 \rightarrow \sigma_{\infty,d\ell}^2$  for  $d = 0, 1$  and  $\ell = 0, 1, \dots, K$ .

Then we have

(a) *The following convergence holds*

$$\begin{pmatrix} \hat{Z}_p \\ \hat{Z}_{\beta_1} \\ \hat{Z}_{\beta_0} \\ \hat{Z}_q \end{pmatrix} \equiv \sqrt{n} \begin{pmatrix} \hat{p} - p_{F_n} \\ \hat{\beta}_1 - \beta_{F_n,1} \\ \hat{\beta}_0 - \beta_{F_n,0} \\ \hat{q} - q_{F_n} \end{pmatrix} \xrightarrow{d} \begin{bmatrix} \mathcal{Z}_p \\ \mathcal{Z}_{\beta_1} \\ \mathcal{Z}_{\beta_0} \\ \mathcal{Z}_q \end{bmatrix}, \quad (\text{A.19})$$

where

$$(\mathcal{Z}'_p, \mathcal{Z}'_{\beta_1}, \mathcal{Z}'_{\beta_0}, \mathcal{Z}'_q)' \sim \mathcal{N}(0_{3(K+1) \times 1}, \text{diag}\{\Sigma_{p,\infty}, \Sigma_{\beta_1,\infty}, \Sigma_{\beta_0,\infty}, \Sigma_{q,\infty}\}),$$

$$\Sigma_{p,\infty} \equiv \text{diag} \left\{ \frac{p_\infty(z_\ell)(1 - p_\infty(z_\ell))}{q_\infty(z_\ell)} : \ell = 0, 1, \dots, K \right\}$$

$$\Sigma_{\beta_1,\infty} \equiv \text{diag} \left\{ \frac{\sigma_{\infty,1\ell}^2}{q_\infty(z_\ell)p_\infty(z_\ell)} : \ell = 0, 1, \dots, K \right\}$$

$$\Sigma_{\beta_0,\infty} \equiv \text{diag} \left\{ \frac{\sigma_{\infty,0\ell}^2}{q_\infty(z_\ell)(1 - p_\infty(z_\ell))} : \ell = 0, 1, \dots, K \right\}$$

$$\Sigma_{q,\infty} \equiv \{\Sigma_{q,\infty}[i, j]\}_{i,j=0,1,\dots,K},$$

with

$$\Sigma_{q,\infty}[i, j] = \begin{cases} p_\infty(z_i)(1 - p_\infty(z_i)) \\ -p_\infty(z_i)p_\infty(z_j). \end{cases}$$

(b) We have consistent estimators for the asymptotic variance:

$$\begin{aligned}\hat{\Sigma}_p &\xrightarrow{p} \Sigma_{p,\infty} \\ \hat{\Sigma}_\beta &\xrightarrow{p} \Sigma_{\beta,\infty} \equiv \text{diag}\{\Sigma_{\beta_1,\infty}, \Sigma_{\beta_0,\infty}\} \\ \hat{\Sigma}_q &\xrightarrow{p} \Sigma_{q,\infty}.\end{aligned}$$

(c) The convergence results in part (a) and (b) also hold for any subsequence  $\{p_n\}$  provided that the conditions given in the lemma hold along  $\{p_n\}$ .

**Lemma A.1.2.** Suppose  $c_\mu \neq 0$  or  $c_\rho \neq 0$  are fixed weights and let  $\lambda = c'\theta$ . For any  $(\theta, F) \in \mathcal{P}$ , the asymptotic variance matrix

$$S \equiv [\partial_p g(\lambda)]\Sigma_p[\partial_{p'} g(\lambda)] + [\partial_{\beta_1} g(\lambda)]\Sigma_{\beta_1}[\partial_{\beta_1'} g(\lambda)] + [\partial_{\beta_0} g(\lambda)]\Sigma_{\beta_0}[\partial_{\beta_0'} g(\lambda)]$$

is uniformly bounded and positive definite. That is,

$$\inf_{(\theta, F) \in \mathcal{P}} \lambda_{\min}(S) > 0, \quad (\text{A.20})$$

and

$$\sup_{(\theta, F) \in \mathcal{P}} \lambda_{\max}(S) < \infty, \quad (\text{A.21})$$

where  $\lambda_{\min}(S)$  and  $\lambda_{\max}(S)$  denote the smallest and largest eigenvalues of  $S$ , respectively.

**Remark A.1.1.** This lemma shows that any choice of nonzero linear weights  $c_\mu$  and  $c_\rho$  would not entail an asymptotic singular covariance matrix for AR test, which is essential for establishing uniform validity.

**Remark A.1.2.** Indeed, we can relax the restriction that  $\lambda = c'\theta$  and show this result for all  $\lambda$  that belongs to a compact subset on the real line. The above statement is sufficient to prove the uniform validity of the conditional Wald test so I do not pursue this extension.

**Lemma A.1.3.** Suppose the conditions in Proposition A.1.1 hold for a sequence  $(\lambda_n, F_n)$  for which  $\sqrt{n}\|\pi_{F_n}\| \rightarrow s_\infty < \infty$ . Then  $\Upsilon_\infty = (\mathcal{Z}_h, S_\infty, \Sigma_{p,\infty}, \partial_p g(\lambda_\infty))$  belongs to  $\mathcal{U}$  defined in (A.4) almost surely. That is,

$$\mathbb{P}_{\eta^*} \left( \mathcal{Z}_h + [\partial_p \pi]_{\Sigma_{p,\infty}} \partial_{p'} g(\lambda_\infty) S_\infty^{-1/2} \eta^* \neq 0_{K \times 1} \mid \mathcal{Z}_h \right) = 1 \quad a.s.$$

**Lemma A.1.4.** [D. W. Andrews and Guggenberger (2017, Lemma 10.3)] Assume the following conditions hold under a sequence of DGPs  $\{(\theta_{F_n}, F_n)\}_{n \geq 1}$ :

1. The scaled singular values converge

$$\sqrt{n}\tau_{j,F_n} \rightarrow \begin{cases} \infty & \text{if } j \leq q \\ t_j \in [0, \infty) & \text{if } j > q \end{cases}$$

2.  $C_{F_n}$  and  $B_{F_n}$  converge to their limits  $C_\infty$  and  $B_\infty$ , respectively.

3. The moment condition and the normalized  $\hat{D}(\theta_{F_n})$  converge jointly:

$$\sqrt{n} \begin{pmatrix} \hat{A}\theta_{F_n} - \hat{\beta} \\ \text{vec}(\hat{D}(\theta_{F_n})) - \text{vec}(A_{F_n}) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathcal{Z}_m \\ \text{vec}(\mathcal{Z}_D) \end{pmatrix}$$

where  $\mathcal{Z}_D = (\mathcal{Z}_{d_1}, \dots, \mathcal{Z}_{d_{2(M+1)}})$  is independent of  $\mathcal{Z}_m$ .

Then we have

$$n^{1/2} \hat{D}(\theta_{F_n}) B_{F_n} S_{F_n} \xrightarrow{d} \mathcal{D} = O_p(1)$$

where

$$S_{F_n} \equiv \text{diag}\{(\sqrt{n}\tau_{1,F_n})^{-1}, \dots, (\sqrt{n}\tau_{q,F_n})^{-1}, 1, \dots, 1\} \in \mathbb{R}^{2(M+1) \times 2(M+1)}. \quad (\text{A.22})$$

Moreover,  $\mathcal{D}$  is independent of  $\mathcal{Z}_m$ .

**Lemma A.1.5.** Suppose the assumptions in Lemma A.1.4 hold. Let  $\xi \in \mathbb{R}^{k \times d_\theta}$  be a matrix of i.i.d. standard normal random variables and  $\kappa > 0$ . Then

$$n^{1/2}(\hat{D}(\theta_{F_n}) + \kappa n^{-1/2}\xi) B_{F_n} S_{F_n} \xrightarrow{d} \mathcal{D}_\xi$$

where  $\mathcal{D}_\xi$  has full rank with probability one and is independent of  $\mathcal{Z}_m$ .

*Proof of Lemma A.1.1. Part (a):* We first show the influence function representation of  $\sqrt{n}(\hat{\beta} - \beta_{F_n})$ . Recall that

$$\begin{aligned} q_F(d, z_\ell) &= \mathbb{P}_F(D = d, Z = z_\ell) \\ &= \begin{cases} q_F(z_\ell) p_F(z_\ell) & \text{if } d = 1 \\ q_F(z_\ell)(1 - p_F(z_\ell)) & \text{if } d = 0. \end{cases} \end{aligned}$$

Denote  $q_\infty(d, z_\ell) \equiv \lim_{n \rightarrow \infty} q_{F_n}(d, z_\ell) \in [\epsilon^2, (1 - \epsilon)^2]$  under the parameter space restriction.

Note that for each  $d = 0, 1$  and  $\ell = 0, 1, \dots, K$ .

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{d\ell} - \beta_{F_n, d\ell}) &= \sqrt{n} \left( \frac{\sum_{i=1}^n Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]}{\sum_{i=1}^n \mathbb{1}[D_i = d, Z_i = z_\ell]} - \beta_{F_n, d\ell} \right) \\
&= \sqrt{n} \left( \frac{q_{F_n}(d, z_\ell)}{\hat{q}(d, z_\ell)} \cdot \frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(z_\ell)} - \beta_{F_n, d\ell} \right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} - \beta_{F_n, d\ell} \right) \\
&\quad + \frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} \cdot \sqrt{n} \left( \frac{q_{F_n}(d, z_\ell)}{\hat{q}(d, z_\ell)} - 1 \right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \frac{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} - \beta_{F_n, d\ell} \right) \\
&\quad - \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbb{1}[D_i = d, Z_i = z_\ell] - q_{F_n}(d, z_\ell)) \cdot \left( \frac{\beta_{F_n, d\ell}}{q_{F_n}(d, z_\ell)} + o_p(1) \right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell] - \mathbb{1}[D_i = d, Z_i = z_\ell] \beta_{F_n, d\ell}}{q_{F_n}(d, z_\ell)} + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n (Y_i - \mathbb{E}_{F_n}(Y \mid D = d, Z = z_\ell)) \frac{\mathbb{1}[D_i = d, Z_i = z_\ell]}{\mathbb{P}_{F_n}(D_i = d, Z_i = z_\ell)} + o_p(1).
\end{aligned}$$

The fourth equality holds by applying the WLLN (Lemma A.1.7) to  $\{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]/q_{F_n}(d, z_\ell)\}_{i=1}^n$  and to  $\{\mathbb{1}[D_i = d, Z_i = z_\ell]\}_{i=1}^n$ , which yields

$$\frac{1}{n} \sum_{i=1}^n \frac{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} = \beta_{F_n, d\ell} + o_p(1)$$

$$\hat{q}(d, z_\ell) = q_{F_n}(d, z_\ell) + o_p(1).$$

Note that  $L^{1+\delta}$ -integrability (A.29) required by WLLN holds here since  $\{\mathbb{1}[D_i = d, Z_i =$

$z_\ell\}_{i=1}^n$  are uniformly bounded and for  $\{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell] / q_{F_n}(d, z_\ell)\}_{i=1}^n$ , we have

$$\begin{aligned} \mathbb{E}_{F_n} \left| \frac{Y_i \mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} \right|^{1+\delta} &= \frac{\mathbb{E}_{F_n}(|Y_i|^{1+\delta} \mid D_i = d, Z_i = z_\ell)}{q_{F_n}(d, z_\ell)^\delta} \\ &\leq \frac{\mathbb{E}_{F_n}(|Y_i|^{2+\delta} \mid D_i = d, Z_i = z_\ell)^{\frac{1+\delta}{2+\delta}}}{q_{F_n}(d, z_\ell)^\delta} \quad \text{by Hölder's inequality} \\ &< \frac{\zeta^{\frac{1+\delta}{2+\delta}}}{\epsilon^{2\delta}} < \infty. \end{aligned}$$

The fifth equality holds by applying Lyapunov CLT to  $\{n^{-1/2}(\mathbb{1}[D_i = d, Z_i = z_\ell] - \mathbb{E}_{F_n}(D_i = d, Z_i = z_\ell))\}_{i=1}^n$ :

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbb{1}[D_i = d, Z_i = z_\ell] - q_{F_n}(d, z_\ell)) = O_p(1)$$

and by noting that  $O_p(1)o_p(1) = o_p(1)$ . Here Lyapunov condition (A.31) holds by noting that the sequence is uniformly bounded, and condition (A.30) holds since

$$\text{var}_{F_n}(\mathbb{1}[D_i = d, Z_i = z_\ell]) = q_{F_n}(d, z_\ell)(1 - q_{F_n}(d, z_\ell)) \rightarrow q_\infty(d, z_\ell)(1 - q_\infty(d, z_\ell)) > 0,$$

where  $q_\infty(d, z_\ell) > 0$  by the parameter space restriction.

Likewise, we can derive the influence function representation for  $\sqrt{n}(\hat{p} - p_{F_n})$ . For each  $\ell = 0, 1, \dots, K$ , we have

$$\sqrt{n}(\hat{p}(z_\ell) - p_{F_n}(z_\ell)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (D_i - \mathbb{P}_{F_n}(D = 1 \mid Z = z_\ell)) \frac{\mathbb{1}[Z_i = z_\ell]}{\mathbb{P}_{F_n}(Z = z_\ell)} + o_p(1).$$

Therefore,

$$\begin{aligned} \sqrt{n} \begin{pmatrix} \{\hat{p}(z_\ell) - p_{F_n}(z_\ell)\}_\ell \\ \{\hat{\beta}_{1\ell} - \beta_{F_n,1\ell}\}_\ell \\ \{\hat{\beta}_{0\ell} - \beta_{F_n,0\ell}\}_\ell \\ \{\hat{q}(z_\ell) - q_{F_n}(z_\ell)\}_\ell \end{pmatrix} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \left\{ (D_i - \mathbb{P}_{F_n}(D = 1 \mid Z = z_\ell)) \frac{\mathbb{1}[Z_i = z_\ell]}{\mathbb{P}_{F_n}(Z = z_\ell)} \right\}_\ell \\ \left\{ (Y_i - \mathbb{E}_{F_n}(Y \mid D = 1, Z = z_\ell)) \frac{\mathbb{1}[D_i = 1, Z_i = z_\ell]}{\mathbb{P}_{F_n}(D_i = 1, Z_i = z_\ell)} \right\}_\ell \\ \left\{ (Y_i - \mathbb{E}_{F_n}(Y \mid D = 0, Z = z_\ell)) \frac{\mathbb{1}[D_i = 0, Z_i = z_\ell]}{\mathbb{P}_{F_n}(D_i = 0, Z_i = z_\ell)} \right\}_\ell \\ \left\{ (\mathbb{1}[Z_i = z_\ell] - \mathbb{P}_{F_n}(Z = z_\ell)) \right\}_\ell \end{pmatrix} + o_p(1) \\ &\equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_n(Y_i, D_i, Z_i) + o_p(1). \end{aligned}$$

As  $n$  goes to infinity,

$$\text{var}_{F_n}(\phi_n(Y_i, D_i, Z_i)) = \text{diag}\{\Sigma_{p,F_n}, \Sigma_{\beta_1,F_n}, \Sigma_{\beta_0,F_n}, \Sigma_{q,F_n}\} \rightarrow \text{diag}\{\Sigma_{p,\infty}, \Sigma_{\beta_1,\infty}, \Sigma_{\beta_0,\infty}, \Sigma_{q,\infty}\} \geq 0$$

where

$$\Sigma_{p,F_n} = \text{diag} \left\{ \frac{p_{F_n}(z_\ell)(1 - p_{F_n}(z_\ell))}{q_{F_n}(z_\ell)} : \ell = 0, 1, \dots, K \right\}$$

$$\Sigma_{\beta_1,F_n} = \text{diag} \left\{ \frac{\sigma_{F_n,1\ell}^2}{q_{F_n}(z_\ell)p_{F_n}(z_\ell)} : \ell = 0, 1, \dots, K \right\}$$

$$\Sigma_{\beta_0,F_n} = \text{diag} \left\{ \frac{\sigma_{F_n,0\ell}^2}{q_{F_n}(z_\ell)(1 - p_{F_n}(z_\ell))} : \ell = 0, 1, \dots, K \right\}$$

$$\Sigma_{q,F_n} = \{\Sigma_{q,F_n}[i, j]\}_{i,j=0,1,\dots,K}$$

with

$$\Sigma_{q,F_n}[i, j] = \begin{cases} p_{F_n}(z_i)(1 - p_{F_n}(z_i)) & \text{if } i = j \\ -p_{F_n}(z_i)p_{F_n}(z_j) & \text{if } i \neq j. \end{cases}$$

Here the positive semi-definiteness of  $\text{diag}\{\Sigma_{p,\infty}, \Sigma_{\beta_1,\infty}, \Sigma_{\beta_0,\infty}, \Sigma_{q,\infty}\}$  is implied by the positive semi-definiteness of each covariance matrix in the diagonal position.

To apply the Lyapunov CLT to  $\phi_n(Y_i, D_i, Z_i)$ , it suffices to verify the Lyapunov condition (A.31). We note that  $\left\{ (D_i - p_{F_n}(z_\ell)) \frac{\mathbb{1}[Z_i=z_\ell]}{q_{F_n}(z_\ell)} \right\}_\ell$  and  $\{\mathbb{1}[Z_i = z_\ell] - q_{F_n}(z_\ell)\}_\ell$  are uniformly bounded, and

$$\begin{aligned} & \mathbb{E}_{F_n} \left| (Y_i - \mathbb{E}_{F_n}(Y_i | D = d, Z = z_\ell)) \frac{\mathbb{1}[D_i = d, Z_i = z_\ell]}{\mathbb{P}_{F_n}(D = d, Z = z_\ell)} \right|^{2+\delta} \\ &= \frac{\mathbb{E}_{F_n}[|Y_i - \mathbb{E}_{F_n}(Y_i | D = d, Z = z_\ell)|^{2+\delta} | D = d, Z = z_\ell]}{\mathbb{P}_{F_n}(D = d, Z = z_\ell)^{1+\delta}} \\ &< \frac{2^{2+\delta}\zeta}{\epsilon^{2(1+\delta)}} < \infty, \end{aligned}$$

where the last line holds by the  $L^{2+\delta}$ -integrability of the centralized moment:

$$\begin{aligned}
& \mathbb{E}[|Y - \mathbb{E}[Y \mid D = d, Z = z]|^{2+\delta} \mid D = d, Z = z] \\
& \leq 2^{1+\delta} \left( \mathbb{E}[|Y|^{2+\delta} \mid D = d, Z = z] + |\mathbb{E}[Y \mid D = d, Z = z]|^{2+\delta} \right) \\
& \leq 2^{2+\delta} \mathbb{E}[|Y|^{2+\delta} \mid D = d, Z = z] \\
& \leq 2^{2+\delta} \zeta.
\end{aligned} \tag{A.23}$$

Note that the second line of (A.23) follows by  $c_r$  inequality given in Lemma A.1.6, and the last line holds by the parameter space restriction on the existence of  $(2 + \delta)$ 'th moment of outcomes.

Hence the desired convergence (A.19) follows by Lyapunov CLT.

**Part (b):** Next, we establish the consistency of the variance estimators. First note that

$$\hat{\Sigma}_p - \Sigma_{p, F_n} = \text{diag} \left\{ \frac{\hat{p}(z_\ell)(1 - \hat{p}(z_\ell))}{\hat{q}(d, z_\ell)} - \frac{p_{F_n}(z_\ell)(1 - p_{F_n}(z_\ell))}{q_{F_n}(d, z_\ell)} \right\} = o_p(1)$$

since  $\hat{p} - p_{F_n} = o_p(1)$  by the convergence (A.19) and for  $d = 0, 1$  and  $\ell = 0, 1, \dots, K$ ,

$$\hat{q}(d, z_\ell) - q_{F_n}(d, z_\ell) = o_p(1), \tag{A.24}$$

which is implied by WLLN for bounded arrays. Therefore,  $\hat{\Sigma}_p$  is consistent to  $\Sigma_{p, \infty}$  since  $\Sigma_{p, F_n} \rightarrow \Sigma_{p, \infty}$  as  $n \rightarrow \infty$ . Applying the similar arguments to  $\hat{\Sigma}_q$  also yields the consistency of  $\hat{\Sigma}_q$  to  $\Sigma_{q, \infty}$ .

Now we examine the variance estimators  $\hat{\sigma}_{d\ell}^2$ :

$$\begin{aligned}
\hat{\sigma}_{d\ell}^2 &= \frac{1}{n} \sum_{i=1}^n \frac{(Y_i - \hat{\beta}_{d\ell})^2 \mathbb{1}[D_i = d, Z_i = z_\ell]}{\hat{q}(d, z_\ell)} \\
&= \frac{1}{\hat{q}(d, z_\ell)} \left[ \frac{1}{n} \sum_{i=1}^n \left( (Y_i - \beta_{F_n, d\ell})^2 + 2(Y_i - \beta_{F_n, d\ell})(\hat{\beta}_{d\ell} - \beta_{F_n, d\ell}) + (\hat{\beta}_{d\ell} - \beta_{F_n, d\ell})^2 \right) \mathbb{1}[D_i = d, Z_i = z_\ell] \right] \\
&= \frac{q_{F_n}(d, z_\ell)}{\hat{q}(d, z_\ell)} \left[ \frac{1}{n} \sum_{i=1}^n (Y_i - \beta_{F_n, d\ell})^2 \frac{\mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} \right] \\
&\quad + 2(\hat{\beta}_{d\ell} - \beta_{F_n, d\ell}) \frac{q_{F_n}(d, z_\ell)}{\hat{q}(d, z_\ell)} \left[ \frac{1}{n} \sum_{i=1}^n (Y_i - \beta_{F_n, d\ell}) \frac{\mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} \right] \\
&\quad + (\hat{\beta}_{d\ell} - \beta_{F_n, d\ell})^2
\end{aligned}$$

Note that equation (A.23) implies that  $L^{1+\delta}$ -integrability (A.29) holds for

$$X_{ni} \in \left\{ (Y_i - \beta_{F_n, d\ell})^2 \frac{\mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)}, (Y_i - \beta_{F_n, d\ell}) \frac{\mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} \right\}$$

for appropriately chosen  $\delta$ .

By WLLN, we have

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n (Y_i - \beta_{F_n, d\ell})^2 \frac{\mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} - \sigma_{F_n, d\ell}^2 &= o_p(1) \\
\frac{1}{n} \sum_{i=1}^n (Y_i - \beta_{F_n, d\ell}) \frac{\mathbb{1}[D_i = d, Z_i = z_\ell]}{q_{F_n}(d, z_\ell)} &= o_p(1)
\end{aligned} \tag{A.25}$$

As implied by the convergence (A.19), we have  $\hat{\beta}_{d\ell} - \beta_{F_n, d\ell} = o_p(1)$ . This combined with equations (A.24) and (A.25) then yield

$$\hat{\sigma}_{d\ell}^2 - \sigma_{F_n, d\ell}^2 \xrightarrow{p} 0.$$

Hence we have

$$\hat{\Sigma}_\beta - \Sigma_{\beta, F_n} = \text{diag} \left\{ \frac{\hat{\sigma}_{d\ell}^2}{\hat{q}(d, z_\ell)} - \frac{\sigma_{F_n, d\ell}^2}{q_{F_n}(d, z_\ell)} : d = 1, 0; \ell = 0, 1, \dots, K \right\} = o_p(1).$$

This establish the consistency of  $\hat{\Sigma}_\beta$  since  $\Sigma_{\beta, F_n} \rightarrow \Sigma_{\beta, \infty}$  as  $n \rightarrow \infty$ . Finally, note that part (c) can be proved by replacing  $n$  with the subsequence  $p_n$  throughout the arguments given above. Then the proof is complete.  $\square$

*Proof of Lemma A.1.2.* First, we show the inequality (A.20). Note that  $S$  is a sum of three positive semi-definite matrices. It suffices to show that the minimal eigenvalue of the sum

$$\bar{S} \equiv [\partial_{\beta_1} g(\lambda)] \Sigma_{\beta_1} [\partial_{\beta_1'} g(\lambda)] + [\partial_{\beta_0} g(\lambda)] \Sigma_{\beta_0} [\partial_{\beta_0'} g(\lambda)]$$

is bounded away from zero because

$$\lambda_{\min}(A + B) \geq \lambda_{\min}(A) + \lambda_{\min}(B) \geq \max\{\lambda_{\min}(A), \lambda_{\min}(B)\}$$

given that  $A$  and  $B$  are both positive semi-definite.<sup>3</sup>

Next, note that  $\Sigma_{\beta_1}$  and  $\Sigma_{\beta_0}$  are diagonal matrices with positive elements  $\Sigma_{\beta_d}[\ell, \ell] = \frac{\sigma_{d\ell}^2}{q(d, z_\ell)}$  for  $\ell = 0, 1, \dots, K$  on the diagonal. Let  $\nu_d$  denote the first column of  $[\partial_{\beta_d} g(\lambda)]$ , whose  $k$ -th row ( $1 \leq k \leq K$ ) is defined in equation (2.9), then it follows that

$$[\partial_{\beta_1} g(\lambda)] \Sigma_{\beta_1} [\partial_{\beta_1'} g(\lambda)] = \Sigma_{\beta_1}[0, 0] \cdot \nu_1' \nu_1 + \text{diag} \begin{pmatrix} \Sigma_{\beta_1}[1, 1]([1 - p(z_0)]c_\mu + 2c_\rho)^2 \\ \vdots \\ \Sigma_{\beta_1}[K, K]([1 - p(z_0)]c_\mu + 2c_\rho)^2 \end{pmatrix}$$

and

$$[\partial_{\beta_0} g(\lambda)] \Sigma_{\beta_0} [\partial_{\beta_0'} g(\lambda)] = \Sigma_{\beta_0}[0, 0] \cdot \nu_0' \nu_0 + \text{diag} \begin{pmatrix} \Sigma_{\beta_0}[1, 1](-p(z_0)c_\mu + 2c_\rho)^2 \\ \vdots \\ \Sigma_{\beta_0}[K, K](-p(z_0)c_\mu + 2c_\rho)^2 \end{pmatrix}$$

Hence we have

$$\begin{aligned} \bar{S} &= (\Sigma_{\beta_1}[0, 0] \cdot \nu_1' \nu_1 + \Sigma_{\beta_0}[0, 0] \cdot \nu_0' \nu_0) \\ &\quad + \text{diag} \begin{pmatrix} \Sigma_{\beta_1}[1, 1]([1 - p(z_0)]c_\mu + 2c_\rho)^2 + \Sigma_{\beta_0}[1, 1](-p(z_0)c_\mu + 2c_\rho)^2 \\ \vdots \\ \Sigma_{\beta_1}[K, K]([1 - p(z_0)]c_\mu + 2c_\rho)^2 + \Sigma_{\beta_0}[K, K](-p(z_0)c_\mu + 2c_\rho)^2 \end{pmatrix} \end{aligned}$$

We now show that each diagonal element in the second term has a positive lower bound that does not depend on the distribution of data. For each  $\ell = 0, 1, \dots, K$ , we have

<sup>3</sup> This inequality holds by noting that  $\lambda_{\min}(A + B) = \min_{\|x\|=1} x'(A + B)x \geq \min_{\|x\|=1} x'Ax + \min_{\|x\|=1} x'Bx = \lambda_{\min}(A) + \lambda_{\min}(B)$ .

$\Sigma_{\beta_d}[\ell, \ell] \geq \epsilon/(1 - \epsilon)^2 \equiv \underline{\delta}(\epsilon) > 0$ . This implies the  $k$ -th diagonal element is bounded below by

$$\begin{aligned} & \Sigma_{\beta_1}[k, k]([1 - p(z_0)]c_\mu + 2c_\rho)^2 + \Sigma_{\beta_0}[k, k](-p(z_0)c_\mu + 2c_\rho)^2 \\ & > \underline{\delta}(\epsilon) \underbrace{([1 - p(z_0)]c_\mu + 2c_\rho)^2 + (-p(z_0)c_\mu + 2c_\rho)^2}_{\text{LB}} \end{aligned}$$

On the one hand,

$$\text{LB} \geq \frac{1}{2}c_\mu^2 \tag{A.26}$$

by Cauchy-Schwarz inequality:  $a^2 + b^2 \geq \frac{1}{2}(a - b)^2$ . On the other hand,

$$\begin{aligned} \text{LB} &= 8c_\rho^2 + [p(z_0)^2 + (1 - p(z_0))^2] c_\mu^2 + 4c_\rho c_\mu [1 - 2p(z_0)] \\ &\geq 8c_\rho^2 + \frac{1}{2}c_\mu^2 + 4c_\rho c_\mu [1 - 2p(z_0)] \\ &= \left( 2\sqrt{2}(1 - 2p(z_0))c_\rho + \frac{1}{\sqrt{2}}c_\mu \right)^2 + 8[1 - (1 - 2p(z_0))^2] c_\rho^2 \\ &\geq 32p(z_0)(1 - p(z_0)) c_\rho^2 \\ &\geq 32\epsilon(1 - \epsilon)c_\rho^2 \end{aligned} \tag{A.27}$$

Combining inequalities (A.26) and (A.27) then yields

$$\text{LB} \geq \max \left\{ \frac{1}{2}c_\mu^2, \quad 32\epsilon(1 - \epsilon)c_\rho^2 \right\} > 0$$

since  $c_\mu \neq 0$  or  $c_\rho \neq 0$ . This implies that

$$\begin{aligned} \lambda_{\min}(\bar{S}) &\geq \min \left( \begin{array}{c} \Sigma_{\beta_1}[1, 1]([1 - p(z_0)]c_\mu + 2c_\rho)^2 + \Sigma_{\beta_0}[1, 1](-p(z_0)c_\mu + 2c_\rho)^2 \\ \vdots \\ \Sigma_{\beta_1}[K, K]([1 - p(z_0)]c_\mu + 2c_\rho)^2 + \Sigma_{\beta_0}[K, K](-p(z_0)c_\mu + 2c_\rho)^2 \end{array} \right) \\ &> \underline{\delta}(\epsilon) \max \left\{ \frac{1}{2}c_\mu^2, \quad 32\epsilon(1 - \epsilon)c_\rho^2 \right\} \end{aligned}$$

This establishes the inequality (A.20), as desired.

Next, we show the inequality (A.21). Following the inequality

$$\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$$

given that  $A$  and  $B$  are both positive semi-definite, it suffices to show that

$$\lambda_{\max}([\partial_p g(\lambda)]_{\Sigma_p}[\partial_{p'} g(\lambda)]), \quad \lambda_{\max}([\partial_{\beta_1} g(\lambda)]_{\Sigma_{\beta_1}}[\partial_{\beta_1} g(\lambda)]), \quad \text{and} \quad \lambda_{\max}([\partial_{\beta_0} g(\lambda)]_{\Sigma_{\beta_0}}[\partial_{\beta_0} g(\lambda)])$$

are bounded above by some positive constant that does not depend on the distribution of data. For simplicity, we establish this statement for  $\lambda_{\max}([\partial_p g(\lambda)]_{\Sigma_p}[\partial_{p'} g(\lambda)])$ . Let  $\nu_p$  denote the first column of  $[\partial_p g(\lambda)]$ . For each  $\ell = 0, 1, \dots, K$ , we have  $\Sigma_p[\ell, \ell] = \frac{p(z_\ell)(1-p(z_\ell))}{q(z_\ell)} \leq \frac{1}{4\epsilon} \equiv \bar{\delta}(\epsilon) < \infty$ . Then it follows that

$$\begin{aligned} \lambda_{\max}([\partial_p g(\lambda)]_{\Sigma_p}[\partial_{p'} g(\lambda)]) &\leq \Sigma_p[0, 0] \nu_p' \nu_p + \left( \max_{1 \leq k \leq K} \Sigma_p[k, k] \right) |\lambda - (\beta_{10} - \beta_{00}) c_\mu|^2 \\ &< \bar{\delta}(\epsilon) \sum_{\ell=0}^K |-\lambda + (\beta_{1\ell} - \beta_{0\ell}) c_\mu|^2. \end{aligned}$$

The parameter space  $\mathcal{P}$  requires  $\lambda$  and  $\beta$  to be constrained by a compact space that depends only through  $\Theta$  and  $\zeta > 0$  in Definition 2.2.1, respectively. Thus the right-hand side can be bounded above by a positive constant that is independent of data distribution, which completes the proof.  $\square$

*Proof of Lemma A.1.3.* Let  $\mathcal{H} = [\partial_p \pi]_{\Sigma_{p, \infty}} \partial_{p'} g(\lambda_\infty) \in \mathbb{R}^{K \times K}$ . We divide the proof into two cases:

**Case 1:**  $\text{rank}(\mathcal{H}) > 0$ .

We show a stronger conclusion:  $h + \mathcal{H} S_\infty^{-1/2} \eta^* \neq 0_{K \times 1}$  a.s. for all  $h \in \mathbb{R}^K$  and prove it by contradiction. Suppose there exists  $h_0 \in \mathbb{R}^K$  such that  $\mathbb{P}(h_0 + \mathcal{H} S_\infty^{-1/2} \eta^* = 0_{K \times 1}) > 0$ . By the condition that  $\text{rank}(\mathcal{H}) > 0$ , there exists a vector  $a \in \mathbb{R}^K$  such that  $\mathcal{H}' a \neq 0_{K \times 1}$ . Therefore, the variance of the normal random variable  $a'(h_0 + \mathcal{H} S_\infty^{-1/2} \eta^*)$  is nonzero:

$$a' \mathcal{H} S_\infty^{-1} \mathcal{H}' a \neq 0.$$

This shows  $a'(h_0 + \mathcal{H} S_\infty^{-1/2} \eta^*)$  is a nondegenerate normal random variable, which implies

$$\mathbb{P}(a'(h_0 + \mathcal{H} S_\infty^{-1/2} \eta^*) = 0) = 0.$$

However, this contradicts with the claim that  $\mathbb{P}(h_0 + \mathcal{H}S_\infty^{-1/2}\eta^* = 0_{K \times 1}) > 0$  since

$$\mathbb{P}(h_0 + \mathcal{H}S_\infty^{-1/2}\eta^* = 0_{K \times 1}) \leq \mathbb{P}(a'(h_0 + \mathcal{H}S_\infty^{-1/2}\eta^*) = 0) = 0.$$

So the desired conclusion is established.

**Case 2:**  $\text{rank}(\mathcal{H}) = 0$ .

Note that  $\text{rank}(\mathcal{H}) = 0$  gives  $\mathcal{H} = 0_{K \times K}$ , which implies

$$\begin{aligned} \mathcal{Z}_h + \mathcal{H}S_\infty^{-1/2}\eta^* &= \mathcal{Z}_h \\ &= s_{\infty\ell_\infty} + [\partial_p\pi]\mathcal{Z}_p - \mathcal{H}S_\infty^{-1}\mathcal{Z}_g \\ &= s_{\infty\ell_\infty} + [\partial_p\pi]\mathcal{Z}_p. \end{aligned}$$

Note that the variance of the random vector  $s_{\infty\ell_\infty} + [\partial_p\pi]\mathcal{Z}_p$  equals a full-rank matrix  $[\partial_p\pi]\Sigma_{p,\infty}[\partial_p\pi]$  since  $\partial_p\pi$  and  $\Sigma_{p,\infty}$  have full rank. This implies  $a'(s_{\infty\ell_\infty} + [\partial_p\pi]\mathcal{Z}_p)$  is a non-degenerate normal random variable for all  $a \neq 0_{K \times 1}$ . Following similar arguments in Case 1, we conclude that  $s_{\infty\ell_\infty} + [\partial_p\pi]\mathcal{Z}_p \neq 0_{K \times 1}$  almost surely. Thus the desired conclusion has been established.  $\square$

*Proof of Lemma A.1.4.* This proof closely follows the arguments in D. W. Andrews and Guggenberger (2017, lemma 10.3(d)). For notational simplicity, denote  $d_\theta \equiv 2(M + 1)$  and  $d_m = 2(K + 1)$ , which represent the numbers of columns and rows of matrix  $\hat{D}(\theta)$ , respectively. Let  $S_q$  denote the upper  $q$ -block of  $S_{F_n}$  defined in equation (A.22). Then

$$S_q = \text{diag}\{(\sqrt{n}\tau_{1,F_n})^{-1}, \dots, (\sqrt{n}\tau_{q,F_n})^{-1}\} \quad \text{and} \quad S_{F_n} = \begin{bmatrix} S_q & 0_{q \times (d_\theta - q)} \\ 0_{(d_\theta - q) \times q} & I_{d_\theta - q} \end{bmatrix}. \quad (\text{A.28})$$

Let  $B_{F_n} \equiv (B_q \dot{\vdash} B_{d_\theta - q})$  and  $C_{F_n} \equiv (C_q \dot{\vdash} C_{d_m - q})$ . Then

$$\sqrt{n}\hat{D}(\theta_{F_n})B_{F_n}S_{F_n} = (\sqrt{n}\hat{D}(\theta_{F_n})B_qS_q, \sqrt{n}\hat{D}(\theta_{F_n})B_{d_\theta - q}).$$

Now we analyze each part. Note that

$$\begin{aligned}
\sqrt{n}\hat{D}(\theta_{F_n})B_qS_q &= \sqrt{n}(\hat{D}(\theta_{F_n}) - A_{F_n})B_qS_q + A_{F_n}B_q(\sqrt{n}S_q) \\
&= o_p(1) + C_{F_n}\Pi_{F_n}(B_q, B_{d_\theta-q})'B_q(\sqrt{n}S_q) \\
&= o_p(1) + C_{F_n}\Pi_{F_n}(I_q, 0_{q \times (d_\theta-q)})'(\sqrt{n}S_q) \\
&= o_p(1) + C_{F_n}(I_q, 0_{q \times (d_\theta-q)})' \\
&\xrightarrow{p} C_{q,\infty}
\end{aligned}$$

where  $C_{q,\infty}$  denotes the first  $q$ -columns of matrix  $C_\infty$ . The second line holds by  $\sqrt{n}(\hat{D}(\theta_{F_n}) - A_{F_n}) = O_p(1)$ ,  $S_q = o(1)$ , and recalling that  $A_{F_n} = C_{F_n}\Pi_{F_n}B_{F_n}'$  via singular value decomposition. The third line holds by the construction that  $B_{F_n}$  is an orthogonal matrix.

For the second part, note that

$$\begin{aligned}
\sqrt{n}A_{F_n}B_{d_\theta-q} &= \sqrt{n}C_{F_n}\Pi_{F_n}(B_q, B_{d_\theta-q})'B_{d_\theta-q} \\
&= C_{F_n}(\sqrt{n}\Pi_{F_n})(0_{(d_\theta-q) \times q}, I_{d_\theta-q})' \\
&\rightarrow C_\infty \begin{bmatrix} 0_{q \times (d_\theta-q)} \\ \text{diag}\{t_{q+1}, \dots, t_{d_\theta}\} \\ 0_{(d_m-d_\theta) \times (d_\theta-q)} \end{bmatrix}.
\end{aligned}$$

On the other hand, by continuous mapping theorem,

$$\sqrt{n}(\hat{D}(\theta_{F_n}) - A_{F_n})B_{d_\theta-q} \xrightarrow{d} \mathcal{Z}_D B_{d_\theta-q,\infty}$$

where  $B_{d_\theta-q,\infty}$  denotes the last  $(d_\theta - q)$ -columns of  $B_\infty$ . Therefore, we have

$$\sqrt{n}\hat{D}(\theta_{F_n})B_{d_\theta-q} \xrightarrow{d} C_\infty \begin{bmatrix} 0_{q \times (d_\theta-q)} \\ \text{diag}\{t_{q+1}, \dots, t_{d_\theta}\} \\ 0_{(d_m-d_\theta) \times (d_\theta-q)} \end{bmatrix} + \mathcal{Z}_D B_{d_\theta-q,\infty}$$

Combine the above results, we have

$$\sqrt{n}\hat{D}(\theta_{F_n})B_{F_n}S_{F_n} \xrightarrow{d} \mathcal{D} \equiv \left( C_{q,\infty}, C_\infty \begin{bmatrix} 0_{q \times (d_\theta-q)} \\ \text{diag}\{t_{q+1}, \dots, t_{d_\theta}\} \\ 0_{(d_m-d_\theta) \times (d_\theta-q)} \end{bmatrix} + \mathcal{Z}_D B_{d_\theta-q,\infty} \right)$$

whose stochastic behavior only depends on  $\mathcal{Z}_D$ , thus is independent of  $\mathcal{Z}_m$ .  $\square$

*Proof of Lemma A.1.5.* For notational simplicity, we still denote  $d_\theta = 2(M + 1)$  and  $d_m = 2(K + 1)$ . The proof of Lemma A.1.4 shows that

$$n^{1/2} \hat{D}(\theta_{F_n}) B_{F_n} S_{F_n} \xrightarrow{d} \mathcal{D}$$

where  $\mathcal{D}$  is independent of  $\mathcal{Z}_m$ .

When  $q = d_\theta$ , we have

$$\begin{aligned} n^{1/2} \kappa n^{-1/2} \xi B_{F_n} S_{F_n} &= \kappa \xi B_{F_n} S_{F_n} \\ &\xrightarrow{p} \kappa \xi B_\infty 0_{d_\theta \times d_\theta} \\ &= 0_{d_m \times d_\theta}. \end{aligned}$$

Then

$$n^{1/2} (\hat{D}(\theta_{F_n}) + \kappa n^{-1/2} \xi) B_{F_n} S_{F_n} \xrightarrow{p} C_\infty$$

which is the same limit as the  $n^{1/2} \hat{D}(\theta_{F_n}) B_{F_n} S_{F_n}$ . Thus the introduction of this exogeneous shock would not harm the asymptotic behavior of the test under strong identification.

When  $q < d_\theta$ , we have

$$\begin{aligned} n^{1/2} \kappa n^{1/2} \xi B_{F_n} S_{F_n} &= \kappa \xi B_{F_n} S_{F_n} \\ &\xrightarrow{p} \kappa \xi B_\infty \text{diag}(0_{q \times q}, I_{d_\theta - q}) \\ &= (0_{q \times q}, \kappa \xi B_{d_\theta - q, \infty}). \end{aligned}$$

Combining with the results from Lemma A.1.4, we have

$$\begin{aligned} &n^{1/2} (\hat{D}(\theta_{F_n}) + \kappa n^{-1/2} \xi) B_{F_n} S_{F_n} \\ &\xrightarrow{d} \left( C_{q, \infty}, C_\infty \begin{bmatrix} 0_{q \times (d_\theta - q)} \\ \text{diag}\{t_{q+1}, \dots, t_{d_\theta}\} \\ 0_{(d_m - d_\theta) \times (d_\theta - q)} \end{bmatrix} + (\mathcal{Z}_D + \kappa \xi) B_{d_\theta - q, \infty} \right) \\ &\equiv \mathcal{D}_\xi. \end{aligned}$$

Now we show that  $\mathcal{D}_\xi$  has full rank of probability one. Define

$$C_{d_\theta - q, \xi} \equiv C_\infty \begin{bmatrix} 0_{q \times (d_\theta - q)} \\ \text{diag}\{t_{q+1}, \dots, t_{d_\theta}\} \\ 0_{(d_m - d_\theta) \times (d_\theta - q)} \end{bmatrix} + (\mathcal{Z}_D + \kappa \xi) B_{d_\theta - q, \infty}.$$

Since  $\mathcal{D}_\xi$  has full rank is equivalent to  $C'_\infty \mathcal{D}_\xi$  has full rank, which equals

$$\begin{aligned} C'_\infty \mathcal{D}_\xi &= \begin{bmatrix} C'_{q,\infty} \\ C'_{d_m-q,\infty} \end{bmatrix} (C_{q,\infty}, C_{d_\theta-q,\xi}) \\ &= \begin{bmatrix} C'_{q,\infty} C_{q,\infty} & C'_{q,\infty} C_{d_\theta-q,\xi} \\ C'_{d_m-q,\infty} C_{q,\infty} & C'_{d_m-q,\infty} C_{d_\theta-q,\xi} \end{bmatrix} \\ &= \begin{bmatrix} I_q & C'_{q,\infty} C_{d_\theta-q,\xi} \\ \mathbf{0}_{(d_m-q) \times q} & C'_{d_m-q,\infty} C_{d_\theta-q,\xi} \end{bmatrix}. \end{aligned}$$

Hence,  $\mathcal{D}_\xi$  has full rank is equivalent to  $C'_{d_m-q,\infty} C_{d_\theta-q,\xi}$  having full rank. By Lemma 16.1 from D. W. Andrews and Guggenberger, 2017, it suffices to show that the variance of  $\text{vec}(C'_{d_m-q,\infty} C_{d_\theta-q,\xi})$  is positive definite.

Note that

$$\begin{aligned} \text{var}(\text{vec}(C'_{d_m-q,\infty} C_{d_\theta-q,\xi})) &= \text{var}(\text{vec}(C'_{d_m-q,\infty} (\mathcal{Z}_D + \kappa\xi) B_{d_\theta-q,\infty})) \\ &= \text{var}((B'_{d_\theta-q,\infty} \otimes C'_{d_m-q,\infty}) \text{vec}(\mathcal{Z}_D + \kappa\xi)) \\ &= (B'_{d_\theta-q,\infty} \otimes C'_{d_m-q,\infty}) \text{var}(\text{vec}(\mathcal{Z}_D + \kappa\xi)) (B'_{d_\theta-q,\infty} \otimes C'_{d_m-q,\infty})'. \end{aligned}$$

By the property that  $\text{rank}(A \otimes B) = \text{rank}(A) \text{rank}(B)$ , we have

$$\begin{aligned} \text{rank}(B_{d_\theta-q,\infty} \otimes C_{d_m-q,\infty}) &= \text{rank}(B_{d_\theta-q,\infty}) \text{rank}(C_{d_m-q,\infty}) \\ &= (d_\theta - q)(d_m - q), \end{aligned}$$

implying that  $B_{d_\theta-q,\infty} \otimes C_{d_m-q,\infty}$  is a full-rank matrix. On the other hand,

$$\text{var}(\text{vec}(\mathcal{Z}_D + \kappa\xi)) = \text{var}(\text{vec}(\mathcal{Z}_D)) + \kappa^2 \text{var}(\text{vec}(\xi))$$

with the equality holding by the independence between  $\xi$  and data. Note that  $\text{var}(\text{vec}(\xi)) = I_{dq d_\theta}$  by the assumption that  $\xi$  is a matrix of i.i.d. standard normal variables. Therefore,  $\text{var}(\text{vec}(\mathcal{Z}_D + \kappa\xi))$  is the sum of two PSD matrices, with one of them being positive definite, so we conclude that it is positive definite.

For every  $x \neq \mathbf{0}_{(d_m-q)(d_\theta-q) \times 1}$ , we have  $(B_{d_\theta-q,\infty} \otimes C_{d_m-q,\infty})x \neq \mathbf{0}_{kd_\theta \times 1}$  by the full rank of  $B_{d_\theta-q,\infty} \otimes C_{d_m-q,\infty}$ , and thus

$$x'(B'_{d_\theta-q,\infty} \otimes C'_{d_m-q,\infty}) \text{var}(\text{vec}(\mathcal{Z}_D + \kappa\xi)) (B'_{d_\theta-q,\infty} \otimes C'_{d_m-q,\infty})' x \neq 0$$

by the positive definiteness of  $\text{var}(\text{vec}(\mathcal{Z}_D + \kappa\xi))$ . So we conclude that the variance of  $\text{vec}(C'_{d_m-q,\infty}C_{d_\theta-q,\xi})$  is positive definite. The desired conclusion has been established.  $\square$

### A.1.6 WLLN and CLT for triangular array

The following  $c_r$  inequality can be found in various reference, for example White, 1999, Proposition 3.8

**Lemma A.1.6** (The  $c_r$  inequality). *Let  $X$  and  $Y$  be random variables with  $\mathbb{E}|X|^r < \infty$  and  $\mathbb{E}|Y|^r < \infty$  for some  $r > 0$ . Then*

$$\mathbb{E}|X + Y|^r \leq c_r (\mathbb{E}|X|^r + \mathbb{E}|Y|^r)$$

where  $c_r = 1$  if  $r \leq 1$  and  $c_r = 2^{r-1}$  if  $r > 1$ .

**Lemma A.1.7** (WLLN for  $L^{1+\delta}$  triangular array). *Let  $\{X_{ni}\}$  be a row-wise i.i.d. triangular array of random variables. Suppose*

$$\sup_{n,i} \mathbb{E}|X_{ni}|^{1+\delta} < \infty \tag{A.29}$$

for some  $\delta > 0$ . Then

$$\frac{1}{n} \sum_{i=1}^n X_{ni} - \mathbb{E}[X_{ni}] \xrightarrow{p} 0.$$

*Proof of Lemma A.1.7.* For each  $\epsilon > 0$ , note that

$$\begin{aligned} \mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^n X_{ni} - \mathbb{E}[X_{ni}] \right| > \epsilon \right) &\leq \frac{\mathbb{E}|n^{-1} \sum_{i=1}^n (X_{ni} - \mathbb{E}X_{ni})|^{1+\delta}}{\epsilon^{1+\delta}} \\ &\leq \frac{\sum_{i=1}^n \mathbb{E}|X_{ni} - \mathbb{E}X_{ni}|^{1+\delta}}{n^{1+\delta} \epsilon^{1+\delta}} \\ &\leq \frac{2^\delta (\mathbb{E}|X_{ni}|^{1+\delta} + |\mathbb{E}X_{ni}|^{1+\delta})}{n^\delta \epsilon^{1+\delta}} \\ &\leq \frac{2^{1+\delta} \sup_{n,i} \mathbb{E}|X_{n,i}|^{1+\delta}}{n^\delta \epsilon^{1+\delta}} \\ &\rightarrow 0 \end{aligned}$$

as  $n \rightarrow \infty$ . The first line holds by Markov's inequality. The second line holds by triangular inequality. The third line holds by  $c_r$  inequality. The fourth line holds by absolute inequality. Hence the desired conclusion holds.  $\square$

The following central limit theorem for a triangular array of random vectors is borrowed from Hansen (2022, Theorem 9.4).

**Lemma A.1.8** (Multivariate Lyapunov CLT). *Let  $\{X_{ni}\}$  be a row-wise i.i.d. triangular array of random vectors in  $\mathbb{R}^k$  with expectations  $\mathbb{E}[X_{ni}] = 0$  and covariance matrices  $\Sigma_{ni} = \mathbb{E}[X_{ni}X_{ni}']$ . Set  $\bar{\Sigma}_n = n^{-1} \sum_{i=1}^n \Sigma_{ni}$ . Suppose*

$$\bar{\Sigma}_n \rightarrow \Sigma \geq 0 \tag{A.30}$$

and for some  $\delta > 0$

$$\sup_{n,i} \mathbb{E} \|X_{ni}\|^{2+\delta} < \infty. \tag{A.31}$$

Then as  $n \rightarrow \infty$ ,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

## A.2 Power Analysis of MLC Tests

### A.2.1 Main results

In this section, I analyze the power of MLC test under strong identification. To define the sequences of DGPs that are strongly identified, I impose the following additional restrictions on the space  $\mathcal{P}$ .

**Definition A.2.1** (Parameter space under strong identification). *For some  $\delta, M > 0$  and  $\epsilon \in (0, 1/2)$ , define the set  $\mathcal{P}_s$  of pairs  $(\theta, F)$  that satisfy the following conditions:*

1.  $(\theta, F) \in \mathcal{P}$ . That is,  $(\theta, F)$  satisfies the conditions imposed in Definition 2.2.1 with the given  $\delta, \zeta > 0$ , and  $\epsilon \in (0, 1/2)$ .
2. Equation (2.5) holds with functions  $\{h_m(\cdot)\}_{m=1}^M$  satisfying Assumption 1.6.

3. There exists a set of index  $\mathcal{S} \subseteq \{0, 1, \dots, K\}$  with  $|\mathcal{S}| = M + 1$  such that

$$\min_{j,k \in \mathcal{S}, j \neq k} |p(z_j) - p(z_k)| \geq \epsilon.$$

Specifically, the second condition brings back the unisolvent property on the specified functions, and the third condition assumes that there is a set of isolated propensity scores sufficient to point identify the primitive parameters in the MTE model.

**Lemma A.2.1.** *With  $\mathcal{P}_s$  defined in Definition A.2.1, We have*

$$\inf_{(\theta, F) \in \mathcal{P}_s} \lambda_{\min}(A_F) > 0.$$

where  $A_F$  is defined below equation (2.10) for some distribution  $F$ .

The parameter space  $\mathcal{P}_s$  guarantees that the smallest singular value of  $A_F$ , denoted as  $\tau_{2(M+1), F}$ , is uniformly bounded away from zero. Since  $A_F$  has full rank for each  $(\theta, F) \in \mathcal{P}_s$ ,  $\theta = (A'_F A_F)^{-1} A'_F \beta_F$  is uniquely determined by  $F$ , so does  $\lambda = c'\theta$ . Therefore,  $\mathcal{P}_s$  can be regarded as a collection of distribution  $F$ , and we denote  $\theta_F \equiv (A'_F A_F)^{-1} A'_F \beta_F$  and  $\lambda_F \equiv c'\theta_F$  for each  $F \in \mathcal{P}_s$ . The next result establishes the consistency and local power property of the MLC test under strong identification:

**Proposition A.2.1.** *Suppose  $\{F_n : n \geq 1\} \subseteq \mathcal{P}_s$ , where for some  $F_0 \in \mathcal{P}_s$ , assume*

1.  $p_{F_n}(z_\ell) \rightarrow p_{F_0}(z_\ell)$  for all  $\ell = 0, 1, \dots, K$ ,
2.  $\beta_{F_n} \rightarrow \beta_{F_0}$ ,
3.  $\sigma_{F_n, d\ell}^2 \rightarrow \sigma_{F_0, d\ell}^2$  for all  $d = 0, 1$  and  $\ell = 0, 1, \dots, K$ ,
4.  $q_{F_n}(z_\ell) \rightarrow q_{F_0}(z_\ell)$  for all  $\ell = 0, 1, \dots, K$ .

Consider a sequence of null values  $\lambda_n^* = \lambda_{F_n} + bn^{-r}$  with  $b \in \mathbb{R}$  and  $b \neq 0$ , then the following conclusion holds

(a) If  $r \in [0, 1/2)$ , then we have

$$\lim_{n \rightarrow \infty} \mathbb{P}_{F_n}(\hat{\phi}_{MLC}(\lambda_n^*) = 1) = 1. \quad (\text{A.32})$$

(b) If  $r = 1/2$ , then we have

$$\lim_{a \searrow 0} \limsup_{n \rightarrow \infty} \mathbb{E}_{F_n} \left[ \hat{\phi}_{Wald}(\lambda_n^*) - \hat{\phi}_{MLC}(\lambda_n^*) \right] = 0.$$

It is worth noting that the proposed MLC test is approximately as powerful as the asymptotic efficient Wald test under strong identification if we set the weight assigned to AR statistic sufficiently small. I. Andrews (2018, Theorem 3) establishes a related result by showing that a  $(1 - \alpha - \gamma)$  LC confidence set is contained by a  $(1 - \alpha)$  Wald confidence set with probability approaching one for all  $\gamma > 0$  based on a specific choice of weight function  $a(\gamma)$  such that the critical values of the two tests are equal:  $q_{(1+a(\gamma))\chi_1^2 + a(\gamma)\chi_{2K+1}^2}^{1-\alpha-\gamma} = q_{\chi_1^2}^{1-\alpha}$ . Nevertheless, it is less clear from his result that the MLC test would have similar local power as the Wald test at the *same significance level, regardless of how we choose the AR weight*. This result is now formally established in Proposition A.2.1(b).

It is not recommended to set  $a = 0$  for the MLC test. On the one hand, the AR statistic helps direct the optimizer of the profiled MLC statistic to converge within a small neighborhood of the true parameter  $\theta$  under strong identification. Within this neighborhood, the MRLM statistic is first-order equivalent to the Wald statistic. However, the MRLM statistic alone cannot detect deviations from the true parameter except in the direction of the target parameter  $c'\theta$ . On the other hand, assigning a higher weight to the AR statistic increases power under weak identification, as the MRLM statistic might be small for distant alternatives due to the near-singularity of  $\hat{A}$  (Kleibergen, 2005). While I. Andrews, 2016 discusses the optimal choice of  $a$  for full vector inference problems, the optimal choice of  $a$  for subvector inference remains an open question for future research.

## A.2.2 Additional lemmas

**Lemma A.2.2.** *Under the same assumptions in Proposition A.2.1, the local power of Wald statistic equals:*

$$\lim_{n \rightarrow \infty} \mathbb{E}_{F_n} [\hat{\phi}_{Wald}(\lambda_n^*)] = \mathbb{P}((\mathcal{Z} + \nu)' P_\nu (\mathcal{Z} + \nu) > q_{\chi_1^2}^{1-\alpha}).$$

where  $\mathcal{Z} \sim \mathcal{N}(0_{2(K+1) \times 1}, I_{2(K+1)})$  and  $\nu$  is defined in equation (A.42).

For each population distribution  $F$ , let

$$\Omega_F(\theta) \equiv H(p_F, \theta) \Sigma_{p, F} H(p_F, \theta)$$

$$\Gamma_{j, F}(\theta) \equiv M_j(p_F) \Sigma_{p, F} H(p_F, \theta)',$$

where  $H(p, \theta)$  and  $M_j(p)$  are defined in section 2.4.3.

**Lemma A.2.3.** *Suppose  $\{F_n\}_{n \geq 1} \subseteq \mathcal{P}_s$  is a sequence of distributions satisfying the conditions in Proposition A.2.1. The following results hold:*

$$0 < \inf_{\theta \in \Theta} \lambda_{\min}(\Omega_{F_0}(\theta)) \leq \sup_{\theta \in \Theta} \lambda_{\max}(\Omega_{F_0}(\theta)) < \infty \quad (\text{A.33})$$

$$\sup_{\theta \in \Theta} \|\hat{\Omega}(\theta) - \Omega_{F_0}(\theta)\| = o_p(1) \quad (\text{A.34})$$

$$\sup_{\theta \in \Theta} \|\hat{\Gamma}_j(\theta) - \Gamma_{j, F_0}(\theta)\| = o_p(1). \quad (\text{A.35})$$

## A.2.3 Proofs

*Proof of Lemma A.2.1.* Since the matrix  $A_F$  is the block diagonal of  $A_{1F}$  and  $A_{0F}$  defined in (2.5), it suffices to show that the minimum singular value of  $A_{dF}$  is bounded away from zero uniformly over  $\mathcal{P}_s$  for both  $d = 0, 1$ . Fix an arbitrary  $d = 0, 1$ , the proof is divided into three steps.

**Part (a):** Establish the positive lower bound of the determinant of an alternant matrix over a compact subset.

Suppose  $\{h_m(\cdot)\}_{m=1}^M$  satisfies the Assumption 1.6. Define the function

$$f(p_0, \dots, p_M) \equiv \lambda_{\min}(A(p_0, \dots, p_M)) \quad \text{for } (p_0, \dots, p_M) \in (0, 1)^{M+1},$$

where  $\lambda_{\min}$  denotes the smallest singular value of  $A(p_0, \dots, p_M) \in \mathbb{R}^{(M+1) \times (M+1)}$  whose  $(\ell, m)$ 'th element is given by

$$A_{\ell m}(p_0, \dots, p_M) \equiv \lambda_{dm}(p_\ell).$$

By Assumption 1.6,  $\{\lambda_{dm}(\cdot)\}_{m=0}^M$  are continuous functions on  $(0, 1)$ . This implies that  $f(p_0, \dots, p_M)$  is also continuous on  $(0, 1)^{M+1}$ . Consider a compact subset

$$\mathcal{E} \equiv \left\{ (p_0, \dots, p_M) : p_\ell \in [\epsilon, 1 - \epsilon] \text{ for all } \ell = 0, \dots, M, \min_{j \neq k} |p_j - p_k| \geq \epsilon \right\}. \quad (\text{A.36})$$

Then  $f$  is strictly positive and continuous on  $\mathcal{E}$  by the unisolvent property<sup>4</sup> imposed in Assumption 1.6. Therefore, there is a natural positive lower bound, denoted by  $\epsilon_d$ , such that

$$\inf_{(p_0, \dots, p_M) \in \mathcal{E}} f(p_0, \dots, p_M) > \epsilon_d > 0.$$

**Part (b):** Show that the smallest singular value of a submatrix of  $A_{dF}$  is uniformly positive.

By the third condition imposed in Definition A.2.1 and the third condition imposed in Definition 2.2.1, for each  $(\theta, F) \in \mathcal{P}$ , there exists a subset of propensity scores  $p_{\mathcal{S}} \equiv \{p_F(z_\ell) : \ell \in \mathcal{S}\}$  such that  $p_{\mathcal{S}} \in \mathcal{E}$  defined in (A.36). Then it follows from part (a) that

$$f(p_{\mathcal{S}}) = \lambda_{\min}(A(p_{\mathcal{S}})) > \epsilon_d > 0.$$

**Part (c):** Conclude the proof.

Note that  $A(p_{\mathcal{S}})$  is a submatrix of  $A_{dF}$  for row index belonging to the set  $\mathcal{S}$ . Therefore,

$$\begin{aligned} \lambda_{\min}(A_{dF}) &= \min_{\|x\|=1} \|(A_{dF})x\| \\ &\geq \min_{\|x\|=1} \|A(p_{\mathcal{S}})x\| \\ &= \lambda_{\min}(A(p_{\mathcal{S}})) \\ &> \epsilon_d > 0 \end{aligned}$$

for all  $(\theta, F) \in \mathcal{P}_s$ . The first and third line holds by the min-max principle for singular values, and the second line follows by the fact that  $A(p_{\mathcal{S}})$  is a submatrix of  $A_{dF}$ . Hence the desired conclusion has been established.  $\square$

*Proof of Proposition A.2.1. Part (a).* First, we show that

$$\inf_{c'\theta = \lambda_n^*} \|\sqrt{n}(\hat{A}\theta - \hat{\beta})\| \rightarrow \infty \quad \text{almost surely.} \quad (\text{A.37})$$

---

<sup>4</sup> The unisolvent property implies that  $A(p_0, \dots, p_M)$  has nonzero determinant, whose absolute value equals the product of all singular values of  $A(p_0, \dots, p_M)$ . From this, it implies that the smallest singular value should be positive.

To prove this, fix an arbitrary  $\theta_n^*$  that satisfies  $c'\theta_n^* = \lambda_n^*$ <sup>5</sup> Consider the following derivation

$$\begin{aligned}
\|\sqrt{n}(\hat{A}\theta_n^* - \hat{\beta})\| &= \|\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) + \sqrt{n}\hat{A}(\theta_n^* - \theta_{F_n})\| \\
&\geq \sqrt{n}\|\hat{A}(\theta_n^* - \theta_{F_n})\| - \|\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})\| \\
&\geq \lambda_{\min}(\hat{A})\sqrt{n}\|\theta_n^* - \theta_{F_n}\| - \|\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})\| \tag{A.38} \\
&\geq \lambda_{\min}(\hat{A})\frac{|c'\theta_{F_n} - \lambda_n^*|}{\sqrt{c'c}} - \|\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})\| \\
&= \lambda_{\min}(A_{F_0})\frac{|b|n^{-r+1/2}}{\sqrt{c'c}} + O_p(1).
\end{aligned}$$

The second line follows by triangle inequality. The third line holds by taking  $x = \theta_n^* - \theta_{F_n}$  in the following equality:

$$\|Ax\| \geq \|x\| \inf_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|} = \|x\| \lambda_{\min}(A). \tag{A.39}$$

The fourth line holds by the fact that the distance between  $\theta_{F_n}$  and  $\theta_n^*$  is bounded below by the distance between  $\theta_{F_n}$  and the hyperplane  $\{\theta : c'\theta = \lambda_n^*\}$ , where the latter equals  $|c'\theta_{F_n} - \lambda_n^*|/\sqrt{c'c}$ . The last line holds by the specification of null values and that  $\hat{A} \xrightarrow{p} A_{F_0}$  and  $\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})$  as implied by Lemma A.1.1. Finally, note that  $r \in [0, 1/2)$  and  $\lambda_{\min}(A_{F_0}) > 0$  (by Lemma A.2.1) implies the desired result.

Second, let  $\theta_n^*$  denote the minimizer in the profiled MLC test statistic:

$$\begin{aligned}
\theta_n^* &\in \underset{c'\theta = \lambda_n^*}{\operatorname{argmin}} \operatorname{MLC}_n(\theta) \\
&= \underset{c'\theta = \lambda_n^*}{\operatorname{argmin}} n(\hat{A}\theta - \hat{\beta})'\hat{\Omega}(\theta)^{-1/2}P_{\hat{Q}(\theta)}\hat{\Omega}(\theta)^{-1/2}(\hat{A}\theta - \hat{\beta}) + a \cdot n(\hat{A}\theta - \hat{\beta})'\hat{\Omega}(\theta)^{-1}(\hat{A}\theta - \hat{\beta}). \tag{A.40}
\end{aligned}$$

The first term of the objective function is always non-negative, we show that the second

---

<sup>5</sup> If such  $\theta_n^*$  does not exist, then the rejection probability becomes 1 by the definition of testing procedures in section 2.4.3. So the desired conclusion trivially holds.

term diverges to infinity. Note that

$$\begin{aligned}
n(\hat{A}\theta_n^* - \hat{\beta})'\hat{\Omega}(\theta_n^*)^{-1}(\hat{A}\theta_n^* - \hat{\beta}) &= \|\sqrt{n}\hat{\Omega}(\theta_n^*)^{-1/2}(\hat{A}\theta_n^* - \hat{\beta})\|^2 \\
&\geq \lambda_{\max}(\hat{\Omega}(\theta_n^*))^{-1}\|\sqrt{n}(\hat{A}\theta_n^* - \hat{\beta})\|^2 \\
&\geq [\lambda_{\max}(\Omega_{F_0}(\theta_n^*))^{-1} + o_p(1)]\|\sqrt{n}(\hat{A}\theta_n^* - \hat{\beta})\|^2 \quad (\text{A.41})
\end{aligned}$$

The second inequality holds by (A.39), and the third line holds by equation (A.34) from Lemma A.2.3. Again following this Lemma, inequality (A.33) implies  $\lambda_{\max}(\Omega_{F_0}(\theta_n^*))^{-1} > 0$ . Combining (A.37) and (A.41) implies that

$$\inf_{c'\theta = \lambda_n^*} \text{MLC}_n(\theta) = \text{MLC}_n(\theta_n^*) \rightarrow \infty \quad \text{almost surely.}$$

As a result, the desired conclusion (A.32) holds.

**Part (b).** We divide the proof into several steps:

**Step 1:** There exists a  $\bar{\theta}_n$  that satisfies  $c'\bar{\theta}_n = \lambda_n^*$  such that  $\text{MLC}_n(\bar{\theta}_n)$  converges to a mixture of noncentral  $\chi^2$  distributions.

Consider

$$\bar{\theta}_n = \theta_{F_n} + \frac{(A'_{F_0}\Omega_{F_0}(\theta_{F_0})^{-1}A_{F_0})^{-1}c}{c'(A'_{F_0}\Omega_{F_0}(\theta_{F_0})^{-1}A_{F_0})^{-1}c} \cdot \frac{b}{\sqrt{n}},$$

Note that  $\bar{\theta}_n \in \Theta$  for sufficiently large  $n$  since  $\theta \in \text{int}(\Theta)$ , and

$$\|A_{F_0}\Omega_{F_0}(\theta_{F_0})^{-1/2}(\bar{\theta}_n - \theta_{F_n})\| = \frac{|b|n^{-1/2}}{\sqrt{c'(A'_{F_0}\Omega_{F_0}(\theta_{F_0})^{-1}A_{F_0})^{-1}c}} = O(n^{-1/2}).$$

So we have  $\|\bar{\theta}_n - \theta_{F_0}\| = o(1)$  as  $\|\theta_{F_n} - \theta_{F_0}\| = o(1)$  holds by the given conditions. Then the  $j$ 'th column of  $\tilde{D}(\bar{\theta}_n)$ , denoted by  $\tilde{D}_j(\bar{\theta}_n)$ , equals

$$\begin{aligned}
\tilde{D}_j(\bar{\theta}_n) &= \hat{a}_j - \hat{\Gamma}_j(\bar{\theta}_n)\hat{\Omega}(\bar{\theta}_n)^{-1}(\hat{A}\bar{\theta}_n - \hat{\beta}) + n^{-1/2}\kappa\xi_j \\
&= \hat{a}_j - \Gamma_{j,F_0}(\theta_{F_0})\Omega_{F_0}(\theta_{F_0})^{-1}(\hat{A}\theta_{F_n} - \hat{\beta}) + o_p(1) \\
&= \hat{a}_j + o_p(1)
\end{aligned}$$

for each  $j = 1, \dots, 2(M+1)$ , where the second line follows by (A.35) from Lemma A.2.3 and the fact that  $\|\theta_{F_n} - \theta_{F_0}\| = o(1)$ . This implies  $\|\tilde{D}(\bar{\theta}_n) - \hat{A}\| = o_p(1)$ . Since  $\hat{\Omega}(\bar{\theta}_n)^{-1} =$

$\Omega_{F_0}(\theta_{F_0})^{-1} + o_p(1) = O_p(1)$ , we have

$$\begin{aligned}\hat{Q}(\bar{\theta}_n) &\equiv \hat{\Omega}(\bar{\theta}_n)^{-1/2} \tilde{D}(\bar{\theta}_n) (\tilde{D}(\bar{\theta}_n)' \hat{\Omega}(\bar{\theta}_n)^{-1} \tilde{D}(\bar{\theta}_n))^{-1} c \\ &= \hat{\Omega}(\bar{\theta}_n)^{-1/2} \hat{A} (\hat{A}' \hat{\Omega}(\bar{\theta}_n)^{-1} \hat{A})^{-1} c + o_p(1).\end{aligned}$$

Plugging this into  $\text{MRLM}_n(\bar{\theta}_n)$  yields

$$\begin{aligned}\text{MRLM}_n(\bar{\theta}_n) &= \frac{\left[ \sqrt{n}(\hat{A}\bar{\theta}_n - \hat{\beta})' \hat{\Omega}(\bar{\theta}_n)^{-1} \tilde{D}(\bar{\theta}_n) \left( \tilde{D}(\bar{\theta}_n)' \hat{\Omega}(\bar{\theta}_n)^{-1} \tilde{D}(\bar{\theta}_n) \right)^{-1} c \right]^2}{c' \left( \tilde{D}(\bar{\theta}_n)' \hat{\Omega}(\bar{\theta}_n)^{-1} \tilde{D}(\bar{\theta}_n) \right)^{-1} c} \\ &= \frac{\left[ \sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})' \hat{\Omega}(\bar{\theta}_n)^{-1} \hat{A} \left( \hat{A}' \hat{\Omega}(\bar{\theta}_n)^{-1} \hat{A} \right)^{-1} c + \sqrt{n}(\bar{\theta}_n - \theta_{F_n})' c + o_p(1) \right]^2}{c' \left( \hat{A}' \hat{\Omega}(\bar{\theta}_n)^{-1} \hat{A} \right)^{-1} c + o_p(1)} \\ &\xrightarrow{d} \left( \frac{\mathcal{Z}' \nu}{\|\nu\|} + \|\nu\| \right)^2 = (\mathcal{Z} + \nu)' P_\nu (\mathcal{Z} + \nu).\end{aligned}$$

where

$$\nu \equiv \Omega_{F_0}(\theta_{F_0})^{-1/2} A_{F_0} \sqrt{n}(\bar{\theta}_n - \theta_{F_n}) = \frac{\Omega_{F_0}(\theta_{F_0})^{-1/2} A_{F_0} (A'_{F_0} \Omega_{F_0}(\theta_{F_0})^{-1} A_{F_0})^{-1} c}{c' (A'_{F_0} \Omega_{F_0}(\theta_{F_0})^{-1} A_{F_0})^{-1} c} \cdot b \quad (\text{A.42})$$

and  $\mathcal{Z}$  is defined as the limit law of  $\sqrt{n}\Omega_{F_n}(\theta_{F_n})^{-1/2}(\hat{A}\theta_{F_n} - \hat{\beta})$ , following a normal distribution  $\mathcal{N}(0_{2(K+1) \times 1}, I_{2(K+1)})$ .

For the  $\text{AR}_n(\bar{\theta}_n)$ , observe that

$$\begin{aligned}\text{AR}_n(\bar{\theta}_n) &= \|\sqrt{n}(\hat{A}\bar{\theta}_n - \hat{\beta})' \hat{\Omega}(\bar{\theta}_n)^{-1/2}\|^2 \\ &= \|\sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta})' \hat{\Omega}(\theta_{F_n})^{-1/2} + \sqrt{n}(\bar{\theta}_n - \theta_{F_n})' \hat{A}' \hat{\Omega}(\theta_{F_n})^{-1/2} + o_p(1)\|^2 \\ &\xrightarrow{d} \|\mathcal{Z} + \nu\|^2.\end{aligned}$$

Combining the results for MRLM and AR statistics, it follows that

$$\text{MLC}_n(\bar{\theta}_n) \xrightarrow{d} (1+a) \cdot (\mathcal{Z} + \nu)' P_\nu (\mathcal{Z} + \nu) + a \cdot (\mathcal{Z} + \nu)' M_\nu (\mathcal{Z} + \nu).$$

**Step 2:** Show that the minimizer of the profiled MLC statistic is consistent to  $\theta_{F_0}$ .

Let  $\theta_n^*$  denote the minimizer of the problem (A.40). Then we have

$$a \cdot \text{AR}_n(\theta_n^*) \leq \text{MLC}_n(\theta_n^*) \leq \text{MLC}_n(\bar{\theta}_n) = O_p(1),$$

where the first inequality follows by the fact that MRLM statistic is always non-negative, the second inequality follows by the definition of  $\theta_n^*$  and that  $\bar{\theta}_n$  from step 1 is feasible under the constraint, and the last equality follows by the conclusion of step 1.

Note that this implies  $\|\sqrt{n}(\hat{A}\theta_n^* - \hat{\beta})\| = O_p(1)$ . Following inequality (A.38), we conclude that  $\|\theta_n^* - \theta_{F_n}\| = O_p(n^{-1/2})$ . Applying the same arguments in step 1 to  $\theta_n^*$  instead of  $\bar{\theta}_n$ , we find that

$$\text{MRLM}_n(\theta_n^*) \xrightarrow{d} (\mathcal{Z} + \nu)' P_\nu(\mathcal{Z} + \nu).$$

**Step 3:** Conclude the proof

To save notations, let  $q(a)$  denote the  $(1 - \alpha)$ -quantile of the mixture chi-square distributions  $(1 + a)\chi_1^2 + a\chi_{2K+1}^2$ . Note that the following inequality

$$\text{MRLM}_n(\theta_n^*) \leq \text{MLC}_n(\theta_n^*) \leq \text{MLC}_n(\bar{\theta}_n)$$

implies the inequality on rejection probabilities:

$$\mathbb{P}_{F_n}(\text{MRLM}_n(\theta_n^*) > q(a)) \leq \mathbb{E}_{F_n}[\hat{\phi}_{\text{MLC}}(\lambda_n^*)] \leq \mathbb{P}_{F_n}(\text{MLC}_n(\bar{\theta}_n) > q(a))$$

Taking the limit on both sides gives

$$L(a) \leq \liminf_{n \rightarrow \infty} \mathbb{E}_{F_n}[\hat{\phi}_{\text{MLC}}(\lambda_n^*)] \leq \limsup_{n \rightarrow \infty} \mathbb{E}_{F_n}[\hat{\phi}_{\text{MLC}}(\lambda_n^*)] \leq U(a)$$

where

$$L(a) = \mathbb{P}((\mathcal{Z} + \nu)' P_\nu(\mathcal{Z} + \nu) > q(a))$$

and

$$U(a) = \mathbb{P}((1 + a) \cdot (\mathcal{Z} + \nu)' P_\nu(\mathcal{Z} + \nu) + a \cdot (\mathcal{Z} + \nu)' M_\nu(\mathcal{Z} + \nu) > q(a)).$$

Note that both  $L(a)$  and  $U(a)$  are continuous function of  $a$ , taking  $a$  to zero from above then yields:

$$L(0) \leq \liminf_{a \searrow 0} \liminf_{n \rightarrow \infty} \mathbb{E}_{F_n}[\hat{\phi}_{\text{MLC}}(\lambda_n^*)] \leq \limsup_{a \searrow 0} \liminf_{n \rightarrow \infty} \mathbb{E}_{F_n}[\hat{\phi}_{\text{MLC}}(\lambda_n^*)] \leq U(0).$$

Since  $L(0) = U(0) = \lim_{n \rightarrow \infty} \mathbb{E}_{F_n} [\hat{\phi}_{\text{Wald}}(\lambda_n^*)]$  by Lemma A.2.2, the desired conclusion has been established.  $\square$

*Proof of Lemma A.2.2.* It suffices to show that

$$\text{Wald}_n(\lambda_n^*) \xrightarrow{d} (\mathcal{Z} + \nu)' P_\nu (\mathcal{Z} + \nu)$$

under  $\{F_n\}_{n \geq 1} \subseteq \mathcal{P}_s$ . Under strong identification such as the parameter space restriction imposed in  $\mathcal{P}_s$ , the efficient estimator  $\hat{\theta}^{\text{eff}}$  (the continuously updated GMM or two-step GMM) is consistent and asymptotic normal with the asymptotic variance (Newey & McFadden, 1994, Theorem 5.2)

$$[A'_{F_0} \Omega_{F_0}(\theta_{F_0})^{-1} A_{F_0}]^{-1}.$$

Then it follows that

$$\begin{aligned} \text{Wald}_n(\lambda_n^*) &= n(c' \hat{\theta}^{\text{eff}} - \lambda_n^*)' \left[ c' (\hat{A}' \hat{\Omega}(\hat{\theta}^{\text{eff}})^{-1} \hat{A})^{-1} c \right]^{-1} (c' \hat{\theta}^{\text{eff}} - \lambda_n^*) \\ &= \left[ \sqrt{n}(c' \hat{\theta}^{\text{eff}} - \lambda_{F_n}) - b \right]' \left[ c' (\hat{A}' \hat{\Omega}(\hat{\theta}^{\text{eff}})^{-1} \hat{A})^{-1} c \right]^{-1} \left[ \sqrt{n}(c' \hat{\theta}^{\text{eff}} - \lambda_{F_n}) - b \right] \\ &= \left( \frac{\sqrt{n}(c' \hat{\theta}^{\text{eff}} - \lambda_{F_n})}{\sqrt{c' (A'_{F_0} \Omega_{F_0}(\theta_{F_0})^{-1} A_{F_0})^{-1} c} + o_p(1)} - \|\nu\| \text{sign}(b) + o_p(1) \right)^2 \end{aligned}$$

For an efficient estimator, we have the following asymptotic representation:

$$\sqrt{n}(\hat{\theta}^{\text{eff}} - \theta_{F_n}) = -(A'_{F_n} \Omega_{F_n}(\theta_{F_n})^{-1} A_{F_n})^{-1} A'_{F_n} \Omega_{F_n}(\theta_{F_n})^{-1} \sqrt{n}(\hat{A} \theta_{F_n} - \hat{\beta}) + o_p(1).$$

which implies

$$\frac{\sqrt{n}(c' \hat{\theta}^{\text{eff}} - \lambda_{F_n})}{\sqrt{c' (A'_{F_0} \Omega_{F_0}(\theta_{F_0})^{-1} A_{F_0})^{-1} c}} \xrightarrow{d} -\frac{\mathcal{Z}' \nu}{\|\nu\|} \cdot \text{sign}(b).$$

Plugging this into the Wald test statistic yields the desired result.

$$\text{Wald}_n(\lambda_n^*) \xrightarrow{d} \left( -\frac{\mathcal{Z}' \nu}{\|\nu\|} - \|\nu\| \right)^2 = (\mathcal{Z} + \nu)' P_\nu (\mathcal{Z} + \nu).$$

Therefore the proof is complete.  $\square$

*Proof of Lemma A.2.3.* For simplicity of notation, I leave out the subscript of  $F_0$  in this proof. First, we show that the smallest and largest singular values of  $\Omega(\theta)$  are finite and bounded away from zero uniformly across  $\theta \in \Theta$  under distribution  $F_0$ . The first inequality can be established by noting that

$$\begin{aligned}\lambda_{\min}(\Omega(\theta)) &= \lambda_{\min}(H(p, \theta)' \Sigma_p H(p, \theta) + \Sigma_\beta) \\ &\geq \lambda_{\min}(\Sigma_\beta) \\ &= \min_{d, \ell} \frac{\sigma_{d\ell}^2}{q(d, z_\ell)} \\ &> 0,\end{aligned}$$

where the first inequality holds since  $H(p, \theta)' \Sigma_p H(p, \theta)$  is positive semi-definite, and the second inequality follows by the parameter space restriction on  $\mathcal{P}$ . Therefore, the lower bound inequality is established. Regarding the upper bound. Note that

$$\begin{aligned}\lambda_{\max}(\Omega(\theta)) &= \lambda_{\max}(H(p, \theta)' \Sigma_p H(p, \theta) + \Sigma_\beta) \\ &\leq \lambda_{\max}(H(p, \theta)' \Sigma_p H(p, \theta)) + \lambda_{\max}(\Sigma_\beta) \\ &= \lambda_{\max}(\Sigma_p^{1/2} H(p, \theta))^2 + \lambda_{\max}(\Sigma_\beta) \\ &\leq \lambda_{\max}(\Sigma_p) \lambda_{\max}(H(p, \theta))^2 + \lambda_{\max}(\Sigma_\beta),\end{aligned}$$

where the first inequality holds by triangle inequality for spectral norm  $\lambda_{\max}(\cdot)$ , and the second inequality holds by the Cauchy-Schwarz inequality. Since each element of  $H(p, \theta)$  is continuous in  $\theta \in \Theta$ , and  $\Theta$  is a compact set,  $\lambda_{\max}(H(p, \theta))$  is bounded uniformly over  $\theta \in \Theta$ . Note that

$$\lambda_{\max}(\Sigma_p) = \max_{\ell} \frac{p(z_\ell)(1-p(z_\ell))}{q(z_\ell)} < \infty \quad \text{and} \quad \lambda_{\max}(\Sigma_\beta) = \max_{d, \ell} \frac{\sigma_{d\ell}^2}{q(d, z_\ell)} < \infty$$

as implied by the conditions imposed on  $\mathcal{P}$ . So the inequality (A.33) is established.

Next we establish the uniform consistent estimation on  $\hat{\Omega}(\theta)$  in (A.34). First note that

the difference in spectral norm can be expressed as

$$\begin{aligned}
\|\hat{\Omega}(\theta) - \Omega(\theta)\| &= \|H(\hat{p}, \theta)' \hat{\Sigma}_p H(\hat{p}, \theta) - H(p, \theta)' \Sigma_p H(p, \theta) + \hat{\Sigma}_\beta - \Sigma_\beta\| \\
&\leq \|H(\hat{p}, \theta)' \hat{\Sigma}_p H(\hat{p}, \theta) - H(\hat{p}, \theta)' \Sigma_p H(\hat{p}, \theta)\| \\
&\quad + \|H(\hat{p}, \theta)' \Sigma_p H(\hat{p}, \theta) - H(p, \theta)' \Sigma_p H(p, \theta)\| \\
&\quad + \|\hat{\Sigma}_\beta - \Sigma_\beta\| \\
&\leq \left( \|\hat{\Sigma}_p^{1/2} H(\hat{p}, \theta)\| + \|\Sigma_p^{1/2} H(\hat{p}, \theta)\| \right) \|H(\hat{p}, \theta)\| \cdot \|\hat{\Sigma}_p^{1/2} - \Sigma_p^{1/2}\| \\
&\quad + \left( \|\Sigma_p^{1/2} H(\hat{p}, \theta)\| + \|\Sigma_p^{1/2} H(p, \theta)\| \right) \|\Sigma_p^{1/2}\| \cdot \|\hat{p} - p\| \\
&\quad + \|\hat{\Sigma}_\beta - \Sigma_\beta\|.
\end{aligned}$$

The first inequality follows by triangle inequality. The second inequality holds by noting that

$$\begin{aligned}
\|X'X - Y'Y\| &= \|X'X - X'Y + X'Y - Y'Y\| \\
&\leq \|X'(X - Y)\| + \|(X - Y)'Y\| \\
&\leq (\|X\| + \|Y\|)(\|X\| - \|Y\|),
\end{aligned}$$

for any matrices  $X$  and  $Y$  that have the same number of columns.

Note that  $\sup_{\theta \in \Theta} \|H(p, \theta)\| < \infty$  since  $H(p, \theta)$  is uniformly bounded due to its continuity in  $\theta \in \Theta$ , where  $\Theta$  is a compact set. Following the conclusions in Lemma A.1.1, we have  $\|\hat{\Sigma}_p\| = O_p(1)$ . Also note that  $\sup_{\theta \in \Theta} H(\hat{p}, \theta) = O_p(1)$ . This follows by the boundedness of  $H(p, \theta)$  on any compact set of  $(p, \theta)$  and the inequality

$$\begin{aligned}
&\mathbb{P}_{F_n} (H(\hat{p}, \theta) \leq \sup\{H(p, \theta) : \epsilon/2 \leq p(z_\ell) \leq 1 - \epsilon/2, \forall \ell = 0, 1, \dots, K; \theta \in \Theta\}) \\
&\geq \mathbb{P}_{F_n} (\|\hat{p} - p_{F_n}\| < \epsilon/2) \\
&\rightarrow 1,
\end{aligned}$$

where the second line holds since  $\epsilon \leq p_{F_n}(z_\ell) \leq 1 - \epsilon$ , and the last line holds by Lemma A.1.1.

Combining the arguments above then implies

$$\sup_{\theta \in \Theta} \|\hat{\Omega}(\theta) - \Omega(\theta)\| = O_p(1) \cdot \|\hat{\Sigma}_p^{1/2} - \Sigma_p^{1/2}\| + O_p(1) \cdot \|\hat{p} - p\| + O_p(1) \cdot \|\hat{\Sigma}_\beta - \Sigma_\beta\|.$$

Then the desired result follows by the conclusions of Lemma A.1.1.

The proof of (A.35) follows similar arguments as the proof of (A.34) and thus omitted.  $\square$

### **A.3 Inference with Estimated Weights for MLC Tests**

When the linear weights on the parameters are unknown but estimated, I maintain Assumption 4 and consider modified linear combination test for the treatment effects parameters  $\lambda = c(p, q)' \theta$ . Naively plugging in the estimator  $\hat{c} = c(\hat{p}, \hat{q})$  into the test statistic can introduce estimation errors in the constraint  $\hat{c}' \theta = \lambda$ , potentially leading to asymptotic bias in the limiting distribution of the test statistic. To address this, I modify the test statistic to account for the effects of these estimation errors.

One possible solution involves reparameterizing the model so that the target parameter  $\hat{c}' \theta$  becomes an element of the reparameterized model parameter (D. W. Andrews, 2017). However, given the broad class of causal parameters of interest, finding a universal reparameterization rule that works for all these parameters is challenging, and such reparameterization would greatly complicates the asymptotic analysis.

As an alternative, I impose a normalization on the weights of the treatment effect parameters. This assumption is broadly applicable to various causal effects, including ATE, ATT, LATE, and the normalized counterfactual policy effects  $\bar{\alpha}(\epsilon)$  considered in this paper.

**Assumption 14.** *The first and  $(K + 1)$ -th element of the weight function  $c(p, q)$  are 1 and  $-1$ , respectively, and weight is symmetric:  $c = (c'_1, c'_0)'$  with  $c_1 = -c_0$ .*

Under Assumption 14, I consider a sequence of vectors  $\tilde{\theta}_n = \theta_{F_n} + \delta_n$  such that  $\hat{c}' \tilde{\theta}_n = \lambda_n = c(p_{F_n}, q_{F_n})' \theta_{F_n}$ . Hence  $\tilde{\theta}_n$  falls into the estimated constraint set. There are two choices of  $\delta_n$  that satisfies this requirement:

$$\begin{aligned} \delta_{n,1} &= (-[\hat{c} - c(p_{F_n}, q_{F_n})]' \theta_{F_n}, 0_{1 \times (2K+1)})' \\ \delta_{n,0} &= (0_{1 \times (K+1)}, [\hat{c} - c(p_{F_n}, q_{F_n})]' \theta_{F_n}, 0_{1 \times K})'. \end{aligned}$$

For any  $r \in [0, 1]$ , define a sequence

$$\theta_{n,r} \equiv r(\theta_{F_n} + \delta_{n,1}) + (1-r)(\theta_{F_n} + \delta_{n,0}). \quad (\text{A.43})$$

This construction generates a continuum of sequences  $\{\theta_{n,r}\}_{n \geq 1}$  that satisfy the estimated constraint. By examining the asymptotic behavior of the test statistic under the sequence  $\{\theta_{n,r}\}_{n \geq 1}$ , we can establish an asymptotically valid test.

**Proposition A.3.1.** *Let Assumption 2, 4, and 14 hold, and suppose that the weight is non-degenerate:  $\inf_{(\theta,F) \in \mathcal{P}} \|c(p_F, q_F)\| > 0$ . Define  $\xi_r \equiv (-r1_{1 \times (K+1)}, (1-r)1_{1 \times (K+1)})'$  with  $r \in [0, 1]$ , where  $1_{1 \times (K+1)}$  denotes a  $1 \times (K+1)$  vector of ones. In the MLC testing procedures outlined in section 2.4.3, consider replacing  $\hat{\Omega}(\theta)$  by*

$$\hat{\Omega}(\theta; r) \equiv (H(\hat{p}, \theta) + \xi_r \theta' [\partial_p \hat{c}]) \hat{\Sigma}_p (H(\hat{p}, \theta) + \xi_r \theta' [\partial_p \hat{c}])' + (\xi_r \theta' [\partial_q \hat{c}]) \hat{\Sigma}_q (\xi_r \theta' [\partial_q \hat{c}])' + \hat{\Sigma}_\beta,$$

replacing  $\hat{\Gamma}_j(\theta)$  by

$$\hat{\Gamma}_j(\theta; r) \equiv \hat{M}_j(\hat{p}) \hat{\Sigma}_p (H(\hat{p}, \theta) + \xi_r \theta' [\partial_p \hat{c}])',$$

and replacing  $c$  with  $\hat{c}$  in the constraint set. With these modifications, the uniform validity of the MLC test is still maintained.

*Proof of Proposition A.3.1.* Under the drifting sequence satisfying the conditions outlined in Proposition A.1.2, we show that

$$\limsup_{n \rightarrow \infty} \mathbb{P} \left( \inf_{c' \theta = \lambda_n} \text{MLC}_n(\theta) > q_{(1+a)\chi_1^2 + a\chi_{2K+1}^2} (1-\alpha) \right) \leq \alpha,$$

where the two chi-square distributions are independent. Then the rest of the proof follows the same arguments in the proof of Theorem 2.4.1.

Consider  $\theta_{n,r}$  for  $r \in [0, 1]$  defined in (A.43). Then we note that  $\lim_{n \rightarrow \infty} \mathbb{P}(\tilde{\theta}_n \in \Theta) = 1$  since  $\theta_{F_n} \in \text{int}(\Theta)$  and the difference between  $\theta_{n,r}$  and  $\theta_{F_n}$  converges in probability to a zero vector. Next, it can be seen that

$$\begin{aligned} \sqrt{n}(\hat{A}\theta_{n,r} - \hat{\beta}) &= \sqrt{n}(\hat{A}\theta_{F_n} - \hat{\beta}) + r \cdot \sqrt{n}\hat{A}\delta_{n,1} + (1-r) \cdot \sqrt{n}\hat{A}\delta_{n,0} \\ &\xrightarrow{d} (H(p_\infty, \theta_\infty) + \xi_r \theta'_{F_n} [\partial_p c]) \mathcal{Z}_p + (\xi_r \theta'_{F_n} [\partial_q c]) \mathcal{Z}_q - \mathcal{Z}_\beta, \end{aligned}$$

where  $\partial_p c$  and  $\partial_q c$  denote the gradients of  $c$  with respect to  $p$  and  $q$ , respectively, by setting  $p = p_\infty$  and  $q = q_\infty$ . Note that  $\hat{\Omega}(\theta_{n,r}; r)$  consistently estimate the asymptotic variance of the above moment condition. Moreover,

$$\hat{\Gamma}_j(\theta_{n,r}; r) \xrightarrow{p} M_j(p_\infty) \Sigma_{p,\infty} (H(p_\infty, \theta_\infty) + \xi'_r \theta_\infty [\partial_p c]'),$$

which equals the asymptotic covariance between  $\sqrt{n}(\hat{A}\theta_{n,r} - \hat{\beta})$  and  $\sqrt{n}(\hat{a}_j - a_{j,F_n})$ , where  $\hat{a}_{j,F_n}$  and  $a_{j,F_n}$  denotes the  $j$ -th row of  $\hat{A}$  and  $A_{F_n}$ , respectively.

Following the same arguments in Proposition A.1.2 but replacing  $\theta_{F_n}$  with  $\theta_{n,r}$  in the argument of test statistics, the consistency of the (co)variance estimators implies

$$\text{MLC}_n(\theta_{n,r}) \xrightarrow{d} (1+a)\chi_1^2 + a\chi_{2K+1}^2. \quad (\text{A.44})$$

Let  $\mathcal{B}$  denote the event that  $\inf_{\mathcal{C}'\theta=\lambda_n} \text{MLC}_n(\theta) > q_{(1+a)\chi_1^2+a\chi_{2K+1}^2} (1-\alpha)$ . Then we have

$$\begin{aligned} \mathbb{P}_{F_n}(\mathcal{B}) &= \mathbb{P}_{F_n}(\mathcal{B}, \theta_{n,r} \in \Theta) + \mathbb{P}_{F_n}(\mathcal{B}, \theta_{n,r} \notin \Theta) \\ &\leq \mathbb{P}_{F_n}(\text{MLC}(\theta_{n,r}) > q_{(1+a)\chi_1^2+a\chi_{2K+1}^2}, \theta_{n,r} \in \Theta) + \mathbb{P}_{F_n}(\theta_{n,r} \notin \Theta) \\ &\leq \mathbb{P}_{F_n}(\text{MLC}(\theta_{n,r}) > q_{(1+a)\chi_1^2+a\chi_{2K+1}^2}) + \mathbb{P}_{F_n}(\theta_{n,r} \notin \Theta), \end{aligned}$$

where the second line holds by the fact that  $\theta_{n,r} \in \Theta$  and satisfies the constraint set under the first term of probability. Taking the limit as  $n \rightarrow \infty$  and noting  $\mathbb{P}_{F_n}(\theta_{n,r} \notin \Theta) \rightarrow 0$  along with (A.44), the desired result holds.  $\square$

## A.4 Power Comparison in Linear MTE Models

In this appendix, I compare the power of the proposed two approaches, namely, the conditional Wald test and modified linear combination test in a linear MTE model. The linear MTE model is generated by the DGP below: For each  $d = 0, 1$ ,

$$\begin{aligned} Y_d &= \mu_d + V_d \\ D &= \mathbb{1}[U \leq p(Z)] \\ V_d &= \rho_d \left( U - \frac{1}{2} \right) + e_d, \end{aligned}$$

where  $U$  is uniformly distributed over a unit interval  $[0, 1]$ ,  $Z$  has the same distribution as the one used in section 2.6, i.e., uniformly distributed over  $\{z_0, z_1, z_2\}$  and independent of  $(U, e_1, e_0)$ , and  $(e_1, e_0)$  follows the joint normal distribution with zero mean and covariance matrix  $\Sigma_e = 0.5 \cdot I_{2 \times 2}$ . Since instrument has ternary support and there are two unknown parameters to be identified for treated and control samples. The linear MTE model is over-identified and is weakly identified if all propensity scores converge to a single point. I consider the following specification of the parameters in the Monte-Carlo simulation:

- Mean potential outcomes:  $\mu_1 = \mu_0 = 0$ ,
- Slope of MTR functions:  $\rho_0 = \rho_1 = 5$
- Propensity scores:
  - (i) Strong identification:  $p^s(z) = [0.2, 0.5, 0.8]$
  - (ii) Weak identification:  $p^w(z) = [0.4, 0.5, 0.6]$

The data is generated with  $n = 500$  units to evaluate the power of the conditional Wald test, modified linear combination test, and the classical Wald test. I repeat the Monte-Carlo experiments 2,000 times to compute the average rejection rates for testing ATE  $H_0 : \mu_1 - \mu_0 = \delta_\mu$  for  $\delta_\mu \in [-5, 5]$ .

The power curves are plotted in Figure A.1 below under two different levels of identification strength. If the variation on the propensity score is strong enough, all considered approaches are valid under the true value of ATE and exhibit similar power when testing against fixed alternatives in finite samples. This result reveals that the proposed MLC test does not sacrifice too much power under strong identification when the weight assigned to AR statistic is small ( $a = 0.05$ ), which coincides with our local power analysis. Under weak identification, both MLC test and conditional Wald test control the size under the true ATE at zero, whereas Wald test over-rejects the true null under the same context. Regarding the power of tests, the MLC test may have deficient power at distant alternatives as it exhibits a non-monotonic power curve similar to the RLM test Kleibergen (2005) for full vector inference. On the other hand, the conditional Wald test has power against both positive and negative alternative values of ATE, which is therefore recommended for

practical implementation in linear MTE models.

## A.5 Testing IV Strength

Several studies have developed tests for IV strength (Lewis & Mertens, 2022; Montiel Olea & Pflueger, 2013; Stock & Yogo, 2005). These tests typically focus on the bias of IV estimators relative to OLS estimators and the size distortion of conventional Wald and t-tests under their null hypotheses. This problem is well-studied in linear IV regressions with homoskedastic error structure (I. Andrews et al., 2019, Section 4). However, little is known about nonlinear models with a focus on subvector inference. In this section, I evaluate the strength of the identification in the empirical application along multiple dimensions.

### A.5.1 Pre-testing weak identification by size distortion

Following I. Andrews, 2018, we can regard the target parameter as being weakly identified if the Wald confidence set fails to include a locally equivalent robust confidence set with a smaller coverage level.

Let  $\gamma \in [0, 1 - \alpha)$ . We choose the weight  $a = a(\gamma)$  in the MLC test such that  $a(\gamma)$  solves

$$q_{(1+a(\gamma))\chi_1^2+a(\gamma)\chi_{2K+1}^2}(1 - \alpha - \gamma) = q_{\chi_1^2}(1 - \alpha).$$

If one uses the weight  $a(\gamma)$  together with the critical value  $q_{\chi_1^2}(1 - \alpha)$ , the following set  $\mathcal{C}_P$  achieves asymptotic coverage  $1 - \alpha - \gamma$  under weak identification:

$$\mathcal{C}_P(\gamma) = \left\{ \lambda \in \mathbb{R} : \inf_{\theta=\lambda} \text{MRLM}_n(\theta) + a(\gamma) \cdot \text{AR}_n(\theta) < q_{\chi_1^2}(1 - \alpha) \right\}.$$

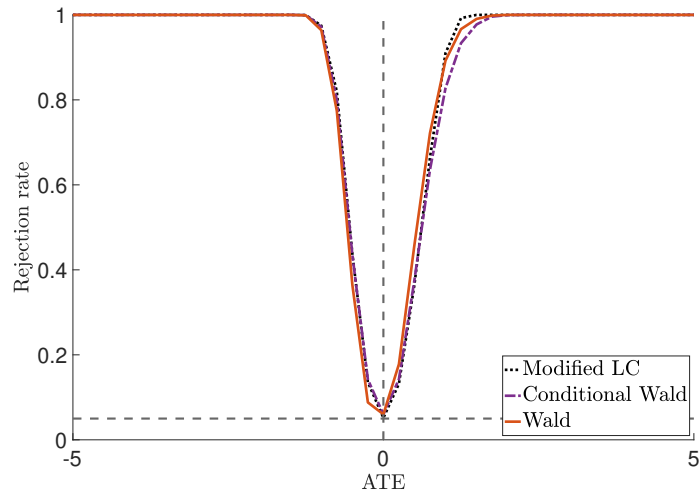
As shown by I. Andrews (2018, Theorem 3), this set is contained by a  $(1 - \alpha)$  Wald confidence set  $\mathcal{C}_W$  with probability approaching one under *strong identification*. In other words, the failure of this containment relationship suggests evidence of weak identification.

This leads to the following test of weak identification:

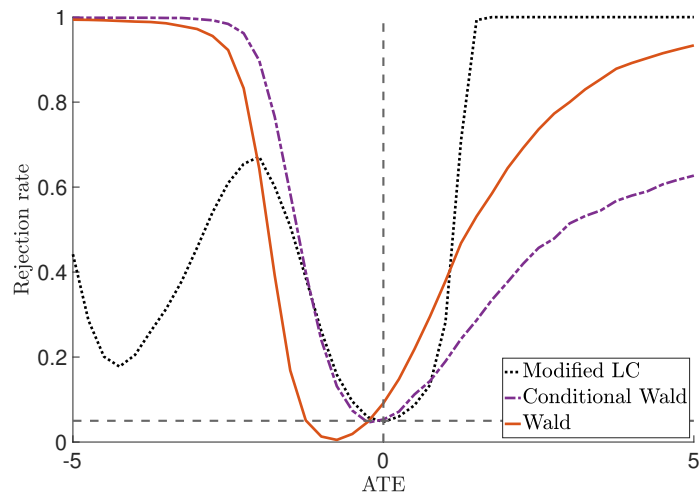
$$\phi_{ICS}(\gamma) = \mathbb{1}[\mathcal{C}_P(\gamma) \not\subseteq \mathcal{C}_W].$$

The tuning parameter  $\gamma$  measures the maximum amount of size distortion on Wald confidence set that researchers can tolerate. If  $\phi_{ICS}(\gamma) = 1$ , then researchers are willing to use

(a) Strong identification:  $p(z) = [0.2, 0.5, 0.8]$



(b) Weak identification:  $p(z) = [0.4, 0.5, 0.6]$



Note: Testing ATE at values on  $[-5, 5]$  with the true effects fixed at zero. The significance level is 5%. The sample size equals 2,000. The average rejection rates are computed with 2,000 independent Monte-Carlo simulations.

FIGURE A.1: Power Curves of the Conditional Wald, MLC, and Wald Tests

the robust confidence set with a smaller coverage level  $(1 - \alpha - \gamma)$  since it is always valid while the classical  $(1 - \alpha)$  Wald confidence set is unreliable if the test rejects. However, this type of test may not have sufficient power to detect weak identification, which implies the model might still be weakly identified even if we did not reject the test.

For the example of additive MP RTE, I report the Andrews' pretesting result in Table A.2 by specifying  $\gamma = 10\%$  and  $\alpha = 5\%$  (i.e., the researchers can tolerate at most 10% size distortion). The result indicates that cubic and quartic MTE models lead to potential concerns of weak identification conditional on most courts, while linear MTE models are strongly identified. Quadratic MTE models are strongly identified for several courts, but not all of them. This result reveals the fact that the identification strength depends on both the variation of instruments and the flexibility of the model specified by researchers, while this information is not reflected by the conventional rule-of-thumb applied to  $F$ -statistics, since the latter rule only works under homoskedastic linear IV models with a scalar endogenous coefficient (I. Andrews et al., 2019).

### A.5.2 Testing exact under-identification

Alternatively, researchers may want to test the full-rank property of the matrix  $A$  in the moment condition, i.e.,  $H_0^{\text{rank}} : \text{rank}(A) \leq 2M + 1$ . Rejections of such tests suggest that the propensity scores are sufficiently separate from each other so that the relevance condition holds under a *pointwise asymptotic* framework. Under a sequence of DGPs that may induce weak identification, we do not necessarily reject  $H_0^{\text{rank}}$  since Assumption 1.7 holds but is close to failure along such sequence. This implies a *larger* critical value is needed to detect weak identification. As choice of valid critical values for testing weak identification is beyond the scope of this paper, I do not pursue this goal but instead comparing the test statistics and the critical value for testing  $H_0^{\text{rank}}$ . The model may be weakly identified if the test statistics are close to the corresponding critical values.

The test for  $H_0^{\text{rank}}$  considered here is based on the analytical bootstrap test from Chen and Fang, 2019. Their test focuses exactly on the null hypothesis  $H_0^{\text{rank}}$  rather than testing

Table A.2: Testing Weak Identification of Additive MP RTE at  $\gamma = 10\%$  and  $\alpha = 5\%$

Note: Valid Test is based on I. Andrews (2018) first-stage pretesting procedure under 10% size distortion.  $F$ -statistics are computed by regressing treatment on a set of saturated instruments, and the Rule of Thumb (RoT) suggests strong identification if the  $F$ -statistics exceed 10.

Court	MTE Poly.	Valid Test	Wald (95%)	MLC (85%)	F-stat	RoT
SBO	Linear MTE	Possibly Strong	[-0.17, -0.02]	[-0.14, -0.06]	77.90	Strong
SBO	Quadratic MTE	Weak	[-0.17, 0.18]	[-0.18, 0.20]		
SBO	Cubic MTE	Weak	[-0.40, 0.45]	[-0.67, 0.55]		
SBO	Quartic MTE	Weak	[-0.64, 0.46]	[-0.76, 0.75]		
EBOS	Linear MTE	Possibly Strong	[-0.24, -0.10]	[-0.23, -0.12]	50.01	Strong
EBOS	Quadratic MTE	Possibly Strong	[-0.22, -0.06]	[-0.20, -0.08]		
EBOS	Cubic MTE	Weak	[-0.34, -0.04]	[-0.45, 0.03]		
EBOS	Quartic MTE	Weak	[-0.36, 0.04]	[-0.63, 0.37]		
WROX	Linear MTE	Possibly Strong	[-0.32, -0.14]	[-0.30, -0.16]	28.15	Strong
WROX	Quadratic MTE	Possibly Strong	[-0.33, -0.12]	[-0.31, -0.13]		
WROX	Cubic MTE	Weak	[-0.33, -0.13]	[-0.40, 0.05]		
WROX	Quartic MTE	Weak	[-0.33, -0.12]	[-0.89, 0.44]		
BMC	Linear MTE	Possibly Strong	[-0.24, 0.00]	[-0.16, -0.07]	31.47	Strong
BMC	Quadratic MTE	Weak	[-0.28, -0.02]	[-0.21, 0.04]		
BMC	Cubic MTE	Weak	[-0.28, 0.00]	[-0.35, 0.12]		
BMC	Quartic MTE	Weak	[-0.37, -0.07]	[-0.30, 0.18]		
ROX	Linear MTE	Possibly Strong	[-0.03, 0.23]	[0.06, 0.17]	34.68	Strong
ROX	Quadratic MTE	Possibly Strong	[-0.02, 0.28]	[0.08, 0.23]		
ROX	Cubic MTE	Weak	[-0.02, 0.34]	[0.05, 0.38]		
ROX	Quartic MTE	Possibly Strong	[-0.10, 0.31]	[0.13, 0.24]		
DOR	Linear MTE	Possibly Strong	[-0.42, -0.18]	[-0.37, -0.25]	43.60	Strong
DOR	Quadratic MTE	Possibly Strong	[-0.60, -0.26]	[-0.56, -0.36]		
DOR	Cubic MTE	Weak	[-0.60, -0.26]	[-0.56, -0.16]		
DOR	Quartic MTE	Weak	[-0.67, -0.29]	[-0.79, -0.08]		

the equality  $\text{rank}(A) = 2M + 1$  as considered by previous work (Kleibergen & Paap, 2006). They show this difference reveals large size distortion when using KP's approach to test for  $H_0^{\text{rank}}$ . Since testing IV relevance at population level is directly related to testing  $H_0^{\text{rank}}$  in our setting, I employ their approach to study this problem.

The test results are collected in Table A.3. From this table we see that the *population* IV relevance condition holds for all courts and models we consider except for the SBO court based on quartic MTE models. As argued above, this result does not necessarily show evidence of strong identification. When looking at the ratios of test statistics and critical

values, we observe that these quantities are close to one under cubic and quartic settings, indicating potential weak identification in such scenarios.

Table A.3: Analytical Bootstrap Test for  $H_0^{\text{rank}} : \text{rank}(A) \leq 2M + 1$

Court	Linear	Quadratic	Cubic	Quartic
Ratio: Test Statistics/Critical Values				
SBO	20.27	1.44	1.83	0.57
EBOS	17.47	3.01	2.19	1.11
WROX	23.29	11.41	3.68	2.46
BMC	16.47	7.54	3.66	1.20
ROX	16.73	6.08	2.70	1.01
DOR	27.58	7.72	2.90	1.32
Test Statistics				
SBO	469.57	1.02	$9.02 \times 10^{-3}$	$3.18 \times 10^{-5}$
EBOS	506.15	1.74	$8.49 \times 10^{-3}$	$4.80 \times 10^{-5}$
WROX	454.48	3.05	$16.68 \times 10^{-3}$	$12.18 \times 10^{-5}$
BMC	243.39	1.05	$3.56 \times 10^{-3}$	$0.86 \times 10^{-5}$
ROX	205.75	0.39	$0.82 \times 10^{-3}$	$0.12 \times 10^{-5}$
DOR	418.09	1.71	$6.08 \times 10^{-3}$	$1.50 \times 10^{-5}$
Critical Values				
SBO	23.17	0.71	$4.92 \times 10^{-3}$	$5.60 \times 10^{-5}$
EBOS	28.97	0.58	$3.88 \times 10^{-3}$	$4.34 \times 10^{-5}$
WROX	19.52	0.27	$4.54 \times 10^{-3}$	$4.94 \times 10^{-5}$
BMC	14.78	0.14	$0.97 \times 10^{-3}$	$0.72 \times 10^{-5}$
ROX	12.30	0.06	$0.31 \times 10^{-3}$	$0.12 \times 10^{-5}$
DOR	15.16	0.22	$2.09 \times 10^{-3}$	$1.14 \times 10^{-5}$

## A.6 Numerical Example of Additive Separability Bias

As alluded in section 2.5.1, the bias of ATE would depend both on the heterogeneity of the coefficient  $\rho_1(W) - \rho_0(W)$  as well as how  $W$  varies the selection into treatment. Consider a linear MTE model as follows:

$$P = \frac{\exp(bW + Z)}{1 + \exp(bW + Z)}$$

$$Y_d = \mu_d(W) + \rho_d(W) \left( U - \frac{1}{2} \right) + e_d$$

$$D = \mathbb{1}[P \leq U]$$

where  $\mathbb{P}(W = 1) = \mathbb{P}(W = 0) = 0.5$  and  $F_Z(z) = (1 + e^{-z})^{-1}$ . Let  $U \perp\!\!\!\perp (Z, W)$  and  $Z \perp\!\!\!\perp W$ . For the parameter specification, I assume

- $\mu_1(w) = 0.5$  and  $\mu_0(w) = 0$  for  $w = 0, 1$ .
- $\rho_1(0) = 1, \rho_1(1) = 1 + \Delta_\rho$ .
- $\rho_0(0) = -1, \rho_0(1) = -1 + \Delta_\rho$ .
- $(e_1, e_0) \sim \mathcal{N}(\mathbf{0}_{2 \times 1}, I_2)$ .
- $(b, \Delta_\rho) \in [-5, 5]^2$ .

When  $b = 0$ , the covariate  $W$  does not shift the probability of being treated. When  $\Delta_\rho = 0$ , there is no heterogeneous treatment effects of covariates varying with selection unobservable (i.e., the additive separability condition holds). Following the intuition provided by Theorem 2.5.1, we would expect bias of ATE estimand when  $b \neq 0$  and  $\Delta_\rho \neq 0$ .

The ATE estimand under additive separability and its bias (defined in section 2.5.1) are shown in the Figure A.2. This figure demonstrates that bias of ATE estimand is pronounced when both  $b$  and  $\Delta_\rho$  are large, and there is no bias on ATE estimand if either  $b$  or  $\Delta_\rho$  equals zero, which coincides the prediction by Theorem 2.5.1. Importantly, this bias has the potential to change the sign of estimand, thereby potentially resulting in misguided policy recommendations.

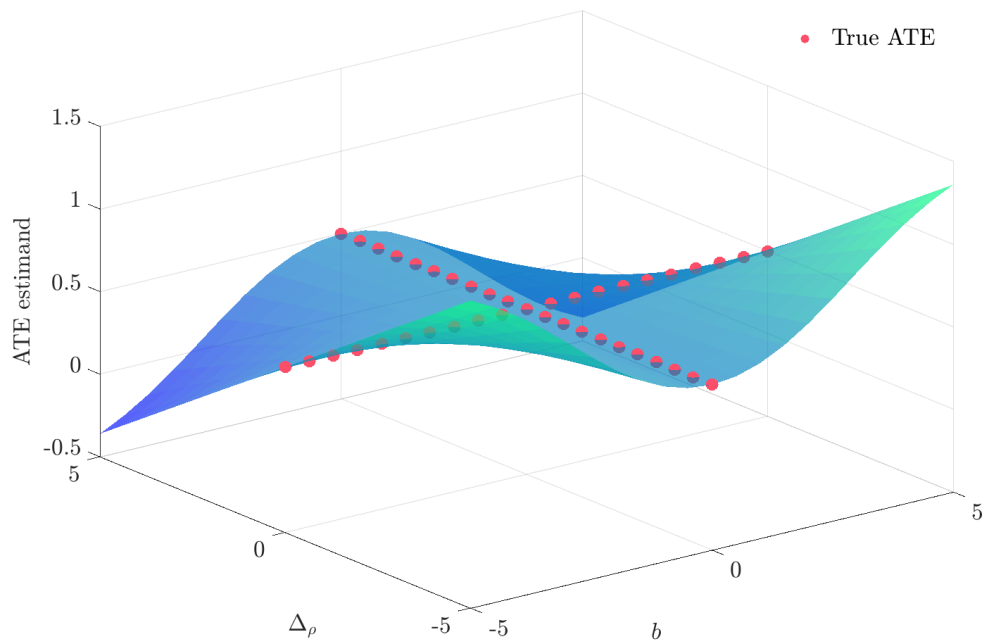
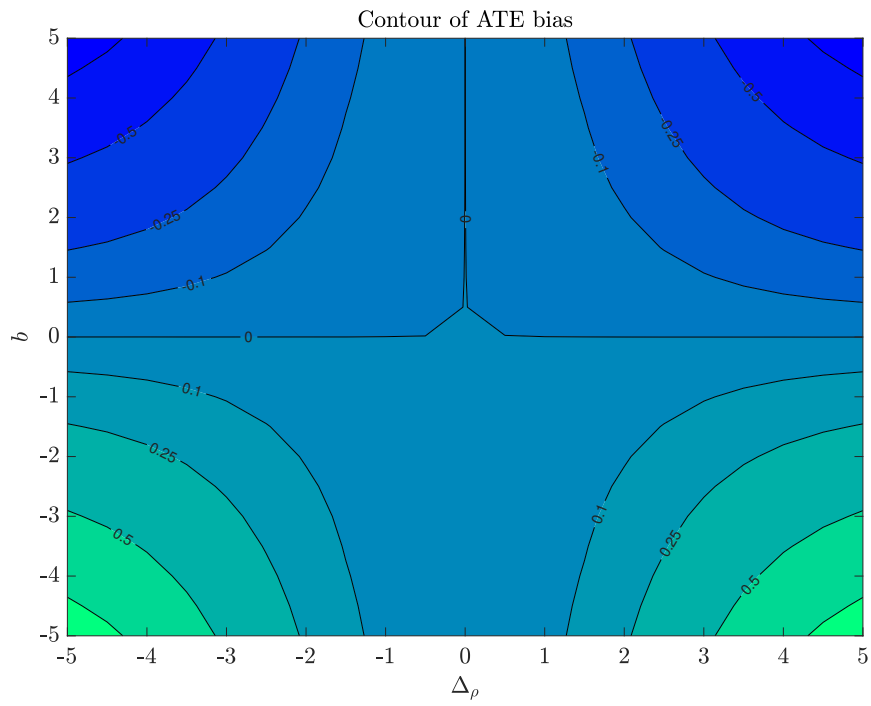
## ***A.7 Additional Literature Discussion***

### **A.7.1 Other approaches to inference on subvectors or functions of parameters**

D. W. Andrews and Guggenberger (2019, Section 12 in supplementary appendix) and D. W. Andrews (2017, Section 2) give thorough literature reviews on inference for weakly identified models. Next I briefly discuss several other main approaches for inference on subvectors or functions of parameters and explain why they cannot be applied to MTE models considered in this paper.

1. Concentrating out strongly identified nuisance parameters

Sometimes researchers assume certain parameters are strongly identified under the null hypothesis. When this holds, identification-robust tests can be constructed by



Top: Contour plot of ATE bias for  $(b, \Delta\rho) \in [-5, 5]^2$ . Bottom: 3D plot of ATE estimand (under additive separability) over the same range, with scatter points showing where the estimand equals true ATE.

FIGURE A.2: Bias of ATE Estimand under Additive Separability

either concentrating out these strongly identified parameters or substituting them with consistent, asymptotically normal estimators. This approach has been adopted in many papers, including several influential work by Stock and Wright, 2000, Kleibergen, 2005, I. Andrews and Mikusheva, 2016b, and D. W. Andrews and Guggenberger, 2019. The resulting confidence sets achieve asymptotic similarity. However, a key limitation is that this method cannot accommodate weakly identified nuisance parameters. This limitation is particularly relevant for the MTE model considered here, where elements of the parameter vector  $\theta$  may be weakly identified when propensity scores exhibit limited variation (see footnote 4).

2. Least-favorable critical values and identification-category-selection method

D. W. Andrews and Cheng (2012, 2013, 2014) and Han and McCloskey (2019) propose uniformly valid inference results for inference on functions of weakly identified parameters based on conventional test statistics such as quasi-likelihood ratio and Wald statistics. These methods are developed for models for which the identification status is determined by whether a vector of strongly identified parameters is zero. When this vector equals zero, the sample objective function does not involve the weakly identified parameters (D. W. Andrews & Cheng, 2012, Assumption A) or the Jacobian matrix of the objective function is approximately a singular matrix (Han & McCloskey, 2019, Assumption ID). For MTE models studied in this paper, the identification status is determined by the vector of propensity scores, with under-identification occurring if this propensity score vector approaches a set where Assumption 1.7 fails, rather than a single point. While some studies have attempted to generalize the framework of D. W. Andrews and Cheng, 2012, they do not yet cover MTE models studied in this paper. For example, Cheng, 2015 considers nonlinear regression models with mixed identification strength where each structural parameter has a one-to-one correspondence with its identification strength parameter. Cox, 2022 develops identification-robust inference for a class of minimum-distance models with restricted parameter spaces, with a primary focus on low-dimensional factor models

under non-negativity restrictions of variances.

### 3. Profiling-based Anderson-Rubin test

Guggenberger et al., 2012 and Guggenberger et al., 2019 propose using profiled Anderson-Rubin test to study inference on a subvector of endogenous coefficients in linear IV models. Their tests rely on the assumption that structural errors are homoskedastic, which is proved to be a key assumption for maintaining asymptotic validity by J. Lee, 2015. In another subsequent work, Guggenberger et al., 2023 relax the homoskedasticity and develop an adaptive procedure by switching into the general subvector test in D. W. Andrews, 2017 if the structural residuals do not follow the approximate homoskedastic structure. Despite the simplicity and convenience of their tests, the assumption of homoskedasticity and linear IV structure do not apply to MTE models. In another related work by I. Andrews and Mikusheva, 2016a, the subvector test is constructed by profiling Anderson-Rubin-type statistics in a minimum distance model. However, as they note (p. 1260), their method reduces to conservative projection inference in linear IV models. A similar limitation likely applies to MTE models, given the linear parametric structure of the moment function.

### **Asymptotic similarity of the subvector inference**

Asymptotic similarity guarantees that, in large samples, the confidence set coverage equals  $1 - \alpha$ , or equivalently, the size of the test equals  $\alpha$ , regardless of identification strength (for a formal definition of similarity, see D. W. Andrews et al., 2020). This property means that confidence sets are not conservative, i.e., unnecessarily wide with coverage exceeding the nominal level, in large samples. While D. W. Andrews and Guggenberger, 2017 has established this property for many identification-robust inference methods on the full parameter vector, it rarely holds for subvector inference with weakly identified nuisance parameters. This failure occurs because weakly identified nuisance parameters usually cannot be consistently estimated under weak identification. Below, I review the main approaches in this literature and present evidence of their inability to achieve asymptotic

similarity.

Methods following D. W. Andrews and Cheng, 2012, 2013, 2014 determine critical values by considering the least-favorable case across all sequences of DGPs. By construction, this robust approach can yield critical values that exceed the true quantile of the test statistic's limiting distribution. Consequently, these tests are not asymptotically similar, as evidenced by the over-coverage probabilities in D. W. Andrews and Cheng (2012, Figure 7).

For linear IV models with homoskedastic errors, the methods of Guggenberger et al., 2012 and Guggenberger et al., 2019 also lack asymptotic similarity. As shown in Guggenberger et al. (2019, Figure 4), null rejection probabilities fall below the significance level when the endogenous coefficient not under testing is weakly identified.

The methods of Chaudhuri and Zivot, 2011, D. W. Andrews, 2017, and I. Andrews, 2018 also fail to achieve asymptotic similarity. These approaches minimize an identification-robust test statistic over a set of weakly identified nuisance parameters (either a confidence set or the full parameter domain), but derive critical values using the true parameter values. Under weak identification, the minimizing nuisance parameters may not converge to their true values, causing the profiled test statistics' distribution to be stochastically dominated by the test statistics evaluated at the true value. This leads to null rejection probabilities falling below the significance level, as demonstrated by the under-rejection of the null hypothesis in both Chaudhuri and Zivot (2011, Table 2) and D. W. Andrews (2017, Table I) for AR/QLR1 tests.

### **A.7.2 Covariates in weakly identified models**

The covariates are often ignored in the literature of weakly identified models, partly because the majority of studies has been focused on the weak identification issues in linear IV models. That is,

$$Y = X\beta + W\eta_Y + U$$

$$X = Z\pi + W\eta_X + V.$$

where  $W$  denotes a set of covariates that possibly include a constant term,  $X$  denotes a vector of endogenous variables, and  $Z$  denotes a vector of instrument variables. In such case we can project out the linear effects of covariates  $(W\eta_Y, W\eta_X)$  by the standard FrischWaughLovell theorem so it reduces to the standard IV model without covariates:

$$Y^{\perp W} = X^{\perp W}\beta + U^{\perp W}$$

$$X^{\perp W} = Z^{\perp W}\pi + V^{\perp W},$$

where  $Y^{\perp W}$  denotes the residual of  $Y$  from regressing on  $W$ . This enable us to discuss the weak identification-robust inference on  $\beta$  by making assumptions on the joint distribution of  $(Y^{\perp W}, X^{\perp W}, Z^{\perp W})$  instead of the distribution of  $(Y, X, Z, W)$ . This argument has been used in a sequence of literature.<sup>6</sup> With a set of high-dimensional covariates, Ma, 2023 considers identification-robust inference on the doubly-robust estimand of LATE.

In nonlinear models subject to weak identification, partialing out the effects of covariates becomes significantly more challenging. Specifically, inference on weakly identified target parameters becomes more complex due to the nonstandard asymptotic behavior of estimators for covariate coefficients. To address identification-robust inference in models with covariates, the existing literature often assumes that the estimation of covariate effects is consistent and asymptotically normal (CAN) when the values of weakly identified coefficients are fixed under the null hypothesis. In other words, this means that covariate effects are assumed to be "strongly identified," while target parameters may be weakly identified. This property usually aligns with the assumption that covariates are "exogenous", in the sense that covariates are independent of structural unobservables but are not necessarily excluded from the outcome equation.

Building on this framework, several studies have imposed the exogeneity of covariates to facilitate identification-robust inference in weakly identified models. For instance, in the consumption capital asset pricing model with constant relative risk aversion preferences,

---

<sup>6</sup> For full vector inference, see Kleibergen (2002), Moreira (2003), and Staiger and Stock (1997), albeit Moreira (2003) only considers the projection of the instrument  $Z$  on exogenous covariates  $W$ . For subvector inference, see Guggenberger et al. (2012, 2019).

Stock and Wright, 2000 treat the discount factor as strongly identified, while the utility parameter is potentially weakly identified. In the case of IV quantile regressions (Chernozhukov & Hansen, 2005), I. Andrews and Mikusheva, 2016b assume that the estimation of covariate effects is CAN once the endogenous coefficients are known, allowing for a null-imposed estimator of covariate coefficients to be used. Similarly, in the endogenous Probit model,<sup>7</sup> D. W. Andrews and Guggenberger, 2019 explore robust inference for the scalar endogenous coefficient, assuming that other covariates are exogenous and their effects can be consistently estimated under the null hypothesis. In the context of MTE models, this assumption leads to an additive separable structure on the outcome equation  $\mathbb{E}[Y_d | U, W]$  and a parametric specification for the first-stage regression. Relatedly, Han and McCloskey, 2019 analyze models with nearly singular Jacobians, which include the Normal MTE model (Björklund & Moffitt, 1987) as a special case.<sup>8</sup> While the Normal MTE model is not their primary focus, they also impose exogeneity and parametric assumptions to facilitate their analysis.

## ***A.8 Two-stage Regression Approach***

In this appendix, I show that the moment function derived from the commonly used two-stage regression approach leads to singular variance under the failure of identification. Non-singularity of moment covariance matrix is crucial for conducting uniformly valid identification-robust inference on structural parameters. This assumption has been commonly imposed by various important strands of literature, for example, the conditional inference procedure (I. Andrews & Mikusheva, 2016b, Assumption 2), the least-favorable critical value approach (D. W. Andrews & Cheng, 2014, Assumption GMM 4\*), and the well-known Robust LM and CLR tests (D. W. Andrews & Guggenberger, 2017, Eq. (3.3)).

---

<sup>7</sup> In the same context, D. W. Andrews and Cheng, 2014 develop an identification-robust inference method by adjusting critical values for conventional tests (such as Wald and QLR). This approach also assumes that covariate effects can be consistently estimated under the null values of weakly identified parameters.

<sup>8</sup> However, the moment conditions formed by two-stage regressions (Han & McCloskey, 2019) are not recommended, as they lead to a singular asymptotic variance matrix shown in Appendix A.8.

Consider a parametric MTE model proposed by Kline and Walters, 2019:

$$\begin{aligned}\mathbb{E}[Y_d | U = u] &= \mu_d + \rho_d h(u) \\ D &= \mathbb{1}[U \leq p(Z)],\end{aligned}$$

where  $U$  is normalized to be uniformly distributed on the unit interval  $[0, 1]$ ,  $Z$  is a discrete exogenous instrument independent of both  $U$  and  $Y_d$ , and  $h(u)$  is a strictly increasing continuous function with  $\mathbb{E}[h(U)] = 0$ . This specification incorporates linear MTE model (Brinch et al., 2017), Heckman normal MTE model (Björklund & Moffitt, 1987), and Logit selection model (Dubin & McFadden, 1984) as special examples.

Under the MTE assumptions, we have

$$\mathbb{E}[Y | Z, D = d] = \mu_d + \rho_d \lambda_d(p(Z)) \tag{A.45}$$

where  $\lambda_0 : (0, 1) \rightarrow \mathbb{R}$  and  $\lambda_1 : (0, 1) \rightarrow \mathbb{R}$  are the control functions giving the means of unobserved heterogeneous components of potential outcomes for the units whose  $U$  (the unobserved cost of getting treatment) is below or above at  $p \in (0, 1)$ :

$$\lambda_1(p) = \mathbb{E}[h(U) | U \leq p], \quad \lambda_0(p) = \mathbb{E}[h(U) | U > p].$$

By regression (A.45), it is obvious to see that  $\mu_d$  and  $\rho_d$  are point identified once  $p(Z)$  has nontrivial variation for both treated and control groups. Therefore, weak identification occurs when propensity scores  $\{p(z) : z = z_0, \dots, z_K\}$  collapse to a single point under a sequence of DGPs.

Typically people estimate propensity score on the first stage and run the second-stage regression (A.45) to obtain estimators of  $\mu_d$  and  $\rho_d$ . Inference on those parameters can be achieved based on the following moment conditions:

$$\mathbb{E}_{\theta^*}[g^{\text{TS}}(W, \theta)] = 0_{(K+5) \times 1} \quad \text{if } \theta = \theta^*,$$

where  $\theta = (\mu_0, \mu_1, \rho_0, \rho_1, \{p(z) : z = z_0, \dots, z_K\})$  is an arbitrary vector of parameters,  $\theta^*$

denotes the true value of parameters, and

$$g^{\text{TS}}(W, \theta) \equiv \begin{pmatrix} [Y - \mu_0 - \rho_0 \lambda_0(p(Z))] \times \mathbb{1}[D = 0] \\ [Y - \mu_0 - \rho_0 \lambda_0(p(Z))] \times \mathbb{1}[D = 0] \times \lambda_0(p(Z)) \\ [Y - \mu_1 - \rho_1 \lambda_1(p(Z))] \times \mathbb{1}[D = 1] \\ [Y - \mu_1 - \rho_1 \lambda_1(p(Z))] \times \mathbb{1}[D = 1] \times \lambda_1(p(Z)) \\ \mathbb{1}[Z = z_0](p(z_0) - \mathbb{1}[D = 1]) \\ \vdots \\ \mathbb{1}[Z = z_K](p(z_K) - \mathbb{1}[D = 1]) \end{pmatrix}. \quad (\text{A.46})$$

**Proposition A.8.1.** *Suppose the parameter  $\theta$  satisfies  $p(z_0) = p(z_1) = \dots = p(z_K)$ , then*

$$\text{var}_{\theta^*}(g^{\text{TS}}(W, \theta)) \equiv \mathbb{E}_{\theta^*} \left[ \left\{ g^{\text{TS}}(W, \theta) - \mathbb{E}_{\theta^*}(g^{\text{TS}}(W, \theta)) \right\} \left\{ g^{\text{TS}}(W, \theta) - \mathbb{E}_{\theta^*}(g^{\text{TS}}(W, \theta)) \right\}' \right]$$

is singular for all  $\theta^*$  in the parameter space.

*Proof.* Since propensity scores do not vary, this implies the first moment condition equals

$$g_1^{\text{TS}}(W, \theta) \equiv \mathbb{1}[D = 1] \times [Y - \mu_0 - \rho_0 \lambda_0(p(Z))] = \mathbb{1}[D = 1] \times [Y - \mu_0 - \rho_0 \lambda_0(p(z_0))]$$

while the second moment condition equals

$$g_2^{\text{TS}}(W, \theta) \equiv \mathbb{1}[D = 1] \times [Y - \mu_0 - \rho_0(p(Z))] \times \lambda_0(p(Z)) = g_1(W, \theta) \lambda_0(p(z_0)).$$

Note that  $g_1(W, \theta) \lambda_0(p(z_0)) = g_2(W, \theta)$ . The moment function  $g^{\text{TS}}(W, \theta)$  is multicollinear for each  $\theta^*$  in the parameter space. Therefore, its variance is singular.  $\square$

The singularity of moment variance for two-stage regressions also appears after reparametrizing the model along the lines of Han and McCloskey, 2019<sup>9</sup>. Applying their reparametriza-

---

<sup>9</sup> This paper derives the a useful reparametrization technique for models with nearly singular Jacobian, which includes the Heckman selection models and the parametric MTE models studied here (see their Example 2.1 and 2.2). The reparametrized model follows the structure posited by D. W. Andrews and Cheng, 2012, 2013, 2014 and therefore allowing powerful inference on functions of model parameters. However, as argued in Appendix A.7.1, their analysis cannot be extended to polynomial MTE models with higher-order polynomials (or multiple selection coefficients).

tion technique to the moment condition  $g^{\text{TS}}$  above gives the reparametrized moment:

$$g^{\text{RP}}(W, \tilde{\theta}) = \begin{pmatrix} [Y - \alpha_0 - \rho_0 \{\lambda_0(p(Z)) - \lambda_0(p(z_0))\}] \times \mathbb{1}[D = 0] \\ [Y - \alpha_0 - \rho_0 \{\lambda_0(p(Z)) - \lambda_0(p(z_0))\}] \times \mathbb{1}[D = 0] \times \lambda_0(p(Z)) \\ [Y - \alpha_1 - \rho_1 \{\lambda_1(p(Z)) - \lambda_1(p(z_0))\}] \times \mathbb{1}[D = 1] \\ [Y - \alpha_1 - \rho_1 \{\lambda_1(p(Z)) - \lambda_1(p(z_0))\}] \times \mathbb{1}[D = 1] \times \lambda_1(p(Z)) \\ \mathbb{1}[Z = z_0](p(z_0) - \mathbb{1}[D = 1]) \\ \vdots \\ \mathbb{1}[Z = z_K](p(z_K) - \mathbb{1}[D = 1]) \end{pmatrix} \quad (\text{A.47})$$

with the reparametrized parameter  $\tilde{\theta} = (\alpha_0, \alpha_1, \rho_0, \rho_1, \{p(z) : z = z_0, \dots, z_K\})$  and  $\alpha_d = \mu_d - \rho_d \lambda_d(p(z_0))$  for  $d = 0, 1$ . It is clear that  $\theta \mapsto \tilde{\theta}$  is a one-to-one mapping, and the reparametrized version of mean potential outcome  $\alpha_d$  is strongly identified from this model for all possible values of propensity scores  $\{p(z) : z = z_0, z_1, \dots, z_K\}$ , whereas  $(\rho_0, \rho_1)$  become weakly identified under limited variation of propensity scores. We can partial out the strongly-identified parameters (see Appendix A.7.1 for the discussion on this approach) to conduct inference on  $(\rho_0, \rho_1)$  based on the reparameterized moment (A.47). However, the validity of such procedure requires the variance of moment condition  $g^{\text{RP}}$  to be nonsingular, which does not hold as this moment condition is constructed by two-stage regressions:

**Corollary A.8.1.** *Suppose the parameter  $\tilde{\theta}$  satisfies  $p(z_0) = p(z_1) = \dots = p(z_K)$ , then*

$$\text{var}_{\tilde{\theta}^*}(g^{\text{RP}}(W, \tilde{\theta})) \equiv \mathbb{E}_{\tilde{\theta}^*} \left[ \left\{ g^{\text{RP}}(W, \tilde{\theta}) - \mathbb{E}_{\tilde{\theta}^*}(g^{\text{RP}}(W, \tilde{\theta})) \right\} \left\{ g^{\text{RP}}(W, \tilde{\theta}) - \mathbb{E}_{\tilde{\theta}^*}(g^{\text{RP}}(W, \tilde{\theta})) \right\}' \right]$$

*is singular for all values of  $\tilde{\theta}^*$  in the parameter space.*

*Proof.* Similar to the proof of Proposition A.8.1, we note that the second moment condition is a factor  $\lambda_0(p(z_0))$  of the first moment condition. Therefore, the singularity follows by the multicollinearity.  $\square$

Although the current setup focuses on discrete instrument without covariates, it is noteworthy that such singularity continues to exist even if instrument is continuous and we estimate this model under additive separability. The source of singularity arises from

the zero variation of control function variables  $\lambda_d(p(Z))$  under the failure of identification, regardless of how  $p(z)$  is estimated from the data. In order to achieve robustness against weak instruments, we need to base on other valid moment conditions derived from the model rather than using the ones derived by two-step regression approach. Recently, D. W. Andrews and Guggenberger, 2019 extend the robustness towards the singularity by selecting and rotating moment conditions that correspond to positive eigenvalues of covariance matrix. However, a simpler solution to address singularity in this setting is to replace the factor  $\lambda_d(p(Z))$  multiplied by the residual of the second stage regressions in (A.46) and (A.47) with some other functions of instrument  $Z$  that has positive variation under identification failure. This is employed in the moment conditions (2.10) constructed in this paper so that singularity problem does not occur.

## Appendix B. Additional Results for Chapter 3

### B.1 Proofs

Throughout this appendix, we use RHS, LLN, CMT, CLT, PSD, PD, and to abbreviate “right-hand side”, “law of large numbers”, “continuous mapping theorem”, “central limit theorem”, “positive semi-definite”, and “positive definite”, respectively. We also define the sequence of random vectors  $\{Z_i\}_{i=1}^n$  with

$$Z_i \equiv \left\{ \sqrt{P(u_k)} [1(X_{i,t} \leq u_k) - 1(X_{i,t+1} \leq u_k)] : t = 1, \dots, T-1, k = 1, \dots, K \right\} \in \mathbb{R}^{(T-1)K}, \quad (\text{B.1})$$

where  $P(u_k)$  is the aggregated probability given in (3.9).

Let  $\Omega \in \mathbb{R}^{(T-1)K \times (T-1)K}$  denote a PSD matrix with a vector of eigenvalues  $\ell \in \mathbb{R}^{(T-1)K}$ . By the Principal-Axis Theorem Scheffe, 1959, pp. 397, 418,  $\epsilon' \Omega \epsilon$  with  $\epsilon \sim N(0_{(T-1)K \times 1}, \Omega)$ , has the same distribution as  $S(\ell)$  defined in Section 3.3.1. That is,  $\epsilon' \Omega \epsilon$  has a generalized chi-square distribution of weights equal to  $\ell$ , unit vector of degrees of freedom, zero vector of non-centrality parameters, and no constant or normal terms. Throughout this appendix, we use  $c^\chi(1 - \alpha; \Omega)$  to denote the  $(1 - \alpha)$ -quantile of  $\epsilon' \Omega \epsilon$ .

*Proof of Theorem 3.2.1. Part (a).* Under Assumption 9 and  $H_0$ ,  $\{Z_i\}_{i=1}^n$  are i.i.d. with  $E[Z_i] = 0_{1 \times (T-1)K}$  and  $V[Z_i] = \Sigma_Z$  as defined in (3.8). By the CLT,

$$\tilde{Z} \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i' \xrightarrow{d} \xi \sim N(0_{(T-1)K \times 1}, \Sigma_Z). \quad (\text{B.2})$$

For any  $t = 1, \dots, T-1$  and  $k = 1, \dots, K$ , the  $((t-1)K + k)$ -component of  $\hat{\delta} \equiv \hat{Z} - \tilde{Z}$  satisfies

$$\begin{aligned} & \sqrt{n \hat{P}(u_k)} [\hat{F}_t(u_k) - \hat{F}_{t+1}(u_k)] - \sqrt{n P(u_k)} [\hat{F}_t(u_k) - \hat{F}_{t+1}(u_k)] \\ & \stackrel{(1)}{=} \sqrt{n} [(\hat{F}_t(u_k) - F_t(u_k)) - (\hat{F}_{t+1}(u_k) - F_{t+1}(u_k))] (\sqrt{\hat{P}(u_k)} - \sqrt{P(u_k)}) \\ & \stackrel{(2)}{=} O_p(1) o_p(1) = o_p(1), \end{aligned} \quad (\text{B.3})$$

where (1) holds by  $F_t(u_k) = F_{t+1}(u_k)$ , which is implied by  $H_0$ , and (2) by  $\sqrt{n}[(\hat{F}_t(u_k) - F_t(u_k)) - (\hat{F}_{t+1}(u_k) - F_{t+1}(u_k))] = O_p(1)$  and  $\hat{P}(u_k) - P(u_k) = o_p(1)$ , which are implied by the CLT and LLN, respectively.

Then, consider the following derivation.

$$S_n \stackrel{(1)}{=} \tilde{Z}'\tilde{Z} + 2\hat{\delta}'\tilde{Z} + \hat{\delta}'\hat{\delta} \stackrel{(2)}{\rightarrow} \xi'\xi. \quad (\text{B.4})$$

where (1) holds by  $\hat{\delta} = \hat{Z} - \tilde{Z}$ , and (2) by (B.2) and (B.3).

To complete the proof, it then suffices to show that RHS of (B.4) can be expressed as (3.12). To this end, consider an orthogonal decomposition of  $\Sigma_Z = H'\Lambda H$ , where  $H \in \mathbb{R}^{(T-1)K \times (T-1)K}$  is an orthogonal matrix (i.e.,  $HH' = I_{(T-1)K \times (T-1)K}$ ) and  $\Lambda = \text{diag}\{\{\lambda_j\}_{j=1}^{(T-1)K}\}$  is the diagonal matrix of eigenvalues of  $\Sigma_Z$ . The desired result then holds by the following derivation:

$$\xi'\xi \stackrel{(1)}{=} \epsilon'\epsilon \stackrel{(2)}{=} \sum_{j=1}^{(T-1)K} \lambda_j \zeta_j^2, \quad (\text{B.5})$$

where (1) holds for  $\epsilon \equiv H\xi$  which implies  $\epsilon'\epsilon = \xi'H'H\xi = \xi'\xi$ , and (2) by  $\epsilon \sim N(0_{(T-1)K \times 1}, \Lambda)$ , and so  $\epsilon = \{\lambda_j^{1/2}\zeta_j\}_{j=1}^{(T-1)K}$  for  $\zeta \sim N(0_{(T-1)K}, I_{(T-1)K \times (T-1)K})$ .

**Part (b).** Under Assumption 9 and the LLN,  $\hat{\Sigma}_Z = \Sigma_Z + o_p(1)$ . Then, Assumption 10 and the CMT imply that  $\hat{\Sigma}_Z^- = \Sigma_Z^{-1} + o_p(1)$ . We can then repeat the arguments in part (a) to get

$$\hat{\delta} = o_p(1) \quad \text{and} \quad \tilde{Z}'\hat{\Sigma}_Z^-\tilde{Z} \stackrel{d}{\rightarrow} \xi'\Sigma_Z^{-1}\xi \sim \chi_{(T-1)K}^2. \quad (\text{B.6})$$

From here, the desired result follows from the next derivation:

$$\bar{S}_n \stackrel{(1)}{=} \tilde{Z}'\hat{\Sigma}_Z^-\tilde{Z} + 2\hat{\delta}'\hat{\Sigma}_Z^-\tilde{Z} + \hat{\delta}'\hat{\Sigma}_Z^-\hat{\delta} \stackrel{d}{\rightarrow} \xi'\Sigma_Z^{-1}\xi \stackrel{(3)}{\sim} \chi_{(T-1)K}^2,$$

where (1) holds by  $\hat{\delta} = \hat{Z} - \tilde{Z}$ , (2) by (B.6), and (3) by  $\xi \sim N(0_{(T-1)K}, \Sigma_Z)$ .  $\square$

*Proof of Theorem 3.3.1. Part (a).* We divide the argument into two cases.

**Case 1:**  $\Sigma_Z \neq 0_{(T-1)K \times (T-1)K}$ . Let  $\lambda \in \mathbb{R}^{(T-1)K}$  denote the vector of eigenvalues of  $\Sigma_Z$ . Since  $\lambda$  is a continuous function of  $\Sigma_Z$  and  $\hat{\Sigma}_Z \xrightarrow{p} \Sigma_Z$  holds, the CMT implies that  $\hat{\lambda}_n \xrightarrow{p} \lambda$ . Since  $\Sigma_Z$  is positive semi-definite,  $\Sigma_Z \neq 0_{(T-1)K \times (T-1)K}$  implies that  $\lambda \neq 0_{(T-1)K \times 1}$ .

Note that  $G(x, \ell)$  is continuous in  $\ell$  for any  $x \in \mathbb{R}$  and  $\ell \in \mathbb{R}^{(T-1)K} \setminus 0_{(T-1)K \times 1}$ . To see why, consider an arbitrary  $x \in \mathbb{R}$  and sequence  $\ell^{(n)} \rightarrow \ell \in \mathbb{R}^{(T-1)K} \setminus 0_{(T-1)K \times 1}$ . The characteristic function of  $S(\ell)$  is

$$\varphi(t, \ell) = \prod_{j=1}^{(T-1)K} (1 - it\ell_j)^{-1/2}$$

and satisfies  $\varphi(t, \ell^{(n)}) \rightarrow \varphi(t, \ell)$  for all  $t \in \mathbb{R}$ . From this and Levy's Continuity Theorem (e.g., see Davidson (1994, Theorem 22.17)) we deduce that  $S(\ell^{(n)}) \xrightarrow{d} S(\ell)$ . This is equivalent to  $G(x, \ell^{(n)}) \rightarrow G(x, \ell)$  since  $S(\ell)$  is continuously distributed (as  $\ell \neq 0_{(T-1)K \times 1}$ ). Since the choices of  $x \in \mathbb{R}$  and  $\ell \in \mathbb{R}^{(T-1)K} \setminus 0_{(T-1)K \times 1}$  were arbitrary, the desired result follows.

Given the continuity of  $G(x, \cdot)$  for all  $x \in \mathbb{R}$ , the CMT gives  $G(x, \hat{\lambda}_n) \xrightarrow{p} G(x, \lambda) = P(S \leq x)$ . In turn, since  $G(x, \hat{\lambda}_n)$  and  $P(S \leq x)$  are weakly increasing and bounded, and  $P(S \leq x)$  is continuous (as  $\lambda \neq 0_{(T-1)K \times 1}$ ), an argument along the lines of A. van der Vaart (1998, Lemma 2.11) implies that

$$\sup_{x \in \mathbb{R}} |G(x, \hat{\lambda}_n) - P(S \leq x)| \xrightarrow{p} 0. \quad (\text{B.7})$$

We now show that

$$c_n^A(1 - \alpha) \xrightarrow{p} c^\chi(1 - \alpha; \Sigma_Z). \quad (\text{B.8})$$

Fix  $\varepsilon > 0$  arbitrarily. Since the CDF of  $S$  is continuous and strictly increasing at  $c^\chi(1 - \alpha; \Sigma_Z) > 0$ ,  $\exists \delta = \delta(\varepsilon) > 0$  such that

$$c^\chi(1 - \alpha; \Sigma_Z) - c^\chi(1 - \alpha - \delta; \Sigma_Z) \leq \varepsilon \quad \text{and} \quad c^\chi(1 - \alpha + \delta; \Sigma_Z) + \delta - c^\chi(1 - \alpha; \Sigma_Z) \leq \varepsilon. \quad (\text{B.9})$$

Then, let  $E_n$  be defined by

$$E_n = \left\{ \sup_{x \in \mathbb{R}} |G(x, \hat{\lambda}_n) - P(S \leq x)| < \delta \right\}.$$

Under  $E_n$ , we have

$$\delta \stackrel{(1)}{\geq} G(c_n^A(1 - \alpha), \hat{\lambda}_n) - P(S \leq c_n^A(1 - \alpha)) \stackrel{(2)}{\geq} (1 - \alpha) - P(S \leq c_n^A(1 - \alpha)),$$

where (1) holds by  $E_n$ , and (2) by (3.14), as it implies  $G(c_n^A(1 - \alpha), \hat{\lambda}_n) \geq 1 - \alpha$ . This yields

$$P(S \leq c_n^A(1 - \alpha)) \geq (1 - \alpha) - \delta. \quad (\text{B.10})$$

From here, we can get

$$c_n^A(1 - \alpha) \stackrel{(1)}{\geq} c^X(1 - \alpha - \delta; \Sigma_Z) \stackrel{(2)}{\geq} c^X(1 - \alpha; \Sigma_Z) - \varepsilon, \quad (\text{B.11})$$

where (1) holds by (B.10) and (2) by the first condition in (B.9). Also under  $E_n$ , we have

$$-\delta \stackrel{(1)}{\leq} G(c_n^A(1 - \alpha) - \delta, \hat{\lambda}_n) - P(S \leq c_n^A(1 - \alpha) - \delta) \stackrel{(2)}{<} (1 - \alpha) - P(S \leq c_n^A(1 - \alpha) - \delta),$$

where (1) holds by  $E_n$ , and (2) by  $G(c_n^A(1 - \alpha) - \delta, \hat{\lambda}_n) < 1 - \alpha$ . This implies that

$$P(S \leq c_n^A(1 - \alpha) - \delta) < (1 - \alpha) + \delta. \quad (\text{B.12})$$

From here, we can get

$$c_n^A(1 - \alpha) \stackrel{(1)}{\leq} c^X((1 - \alpha) + \delta; \Sigma_Z) + \delta \stackrel{(2)}{\leq} c^X(1 - \alpha; \Sigma_Z) + \varepsilon, \quad (\text{B.13})$$

where (1) holds by (B.12) and (2) by the second condition in (B.9). By combining (B.11) and (B.13), we conclude that  $|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| \leq \varepsilon$ . From this argument, we deduce that

$$P(E_n) \leq P(|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| \leq \varepsilon). \quad (\text{B.14})$$

Since  $P(E_n) \rightarrow 1$  by (B.7), we conclude from (B.14) that  $P(|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| \leq \varepsilon) \rightarrow 1$ . Since the choice of  $\varepsilon > 0$  was arbitrary, (B.8) follows.

For any  $\varepsilon \in (0, c^X(1 - \alpha; \Sigma_Z))$ , consider the following argument.

$$\begin{aligned}
& P(\{S_n < c^X(1 - \alpha; \Sigma_Z) - \varepsilon\}) + P(\{|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| \leq \varepsilon\}) - 1 \\
& \leq P(\{S_n < c^X(1 - \alpha; \Sigma_Z) - \varepsilon\} \cap \{|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| \leq \varepsilon\}) \\
& \leq \left\{ \begin{array}{l} P(\{S_n < c^X(1 - \alpha; \Sigma_Z) - \varepsilon\} \cap \{|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| \leq \varepsilon\}) \\ + P(\{S_n < c_n^A(1 - \alpha)\} \cap \{|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| > \varepsilon\}) \end{array} \right\} \\
& \leq \left\{ \begin{array}{l} P(\{S_n < c_n^A(1 - \alpha)\} \cap \{|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| \leq \varepsilon\}) + \\ P(\{S_n < c_n^A(1 - \alpha)\} \cap \{|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| > \varepsilon\}) \end{array} \right\} \\
& = P(S_n < c_n^A(1 - \alpha)) \\
& \stackrel{(1)}{\leq} 1 - E[\phi_n^A(\alpha)] \\
& \stackrel{(2)}{\leq} P(S_n \leq c_n^A(1 - \alpha)) \\
& = \left\{ \begin{array}{l} P(\{S_n \leq c_n^A(1 - \alpha)\} \cap \{|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| \leq \varepsilon\}) + \\ P(\{S_n \leq c_n^A(1 - \alpha)\} \cap \{|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| > \varepsilon\}) \end{array} \right\} \\
& \leq P(S_n \leq c^X(1 - \alpha; \Sigma_Z) + \varepsilon) + P(|c_n^A(1 - \alpha) - c^X(1 - \alpha; \Sigma_Z)| > \varepsilon), \tag{B.15}
\end{aligned}$$

where (1) and (2) hold by  $\{S_n < c_n^A(1 - \alpha)\} \subseteq \{\phi_n^A(\alpha) = 0\} \subseteq \{S_n \leq c_n^A(1 - \alpha)\}$ . By taking sequential limits on (B.15) as  $n \rightarrow \infty$  and  $\varepsilon \rightarrow 0$ , and combined with (B.8), Theorem 3.2.1(a), and the fact that  $S$  is continuously distributed, we conclude that  $E[\phi_n^A(\alpha)] \rightarrow P(S > c^X(1 - \alpha; \Sigma_Z)) = \alpha$ , as desired.

**Case 2:**  $\Sigma_Z = 0_{(T-1)K \times (T-1)K}$ . Under  $H_0$  in (3.3), we also have  $E[Z_i] = 0_{(T-1)K \times 1}$ , and so  $Z_i = 0_{(T-1)K \times 1}$  a.s. By (B.1), this implies that  $1(X_{i,t} \leq u_k) = 1(X_{i,t+1} \leq u_k)$  a.s. for all  $t = 1, \dots, T - 1$  and  $k = 1, \dots, K$ . In turn, this gives that  $\hat{Z} = 0_{(T-1)K \times 1}$  a.s., with  $\hat{Z}$  defined as in (3.7). Then,  $S_n = 0$  holds a.s. By this and  $c_n^A(1 - \alpha) \geq 0$ , we get  $E[\phi_n^A(\alpha)] = P(S_n > c_n^A(1 - \alpha)) = 0 < \alpha$ . From here, the desired result holds by taking limits as  $n \rightarrow \infty$ .

**Part (b)** The conclusion follows directly from combining (3.16), Theorem 3.2.1(b), and that the CDF of  $\chi_{(T-1)K}^2$  is continuous at  $c^X(1 - \alpha; I_{(T-1)K \times (T-1)K}) > 0$ .  $\square$

*Proof of Theorem 3.3.2. Part (a).* We divide the argument into two cases.

**Case 1:**  $\Sigma_Z \neq 0_{(T-1)K \times (T-1)K}$ . For each  $i = 1, \dots, n$ ,  $t = 1, \dots, T-1$ , and  $k = 1, \dots, K$ , let

$$Z_{i,(t-1)K+k}^* \equiv \sqrt{P(u_k)}[1(X_{i,t}^* \leq u_k) - \hat{F}_t(u_k) - (1(X_{i,t+1}^* \leq u_k) - \hat{F}_{t+1}(u_k))],$$

and let  $Z_i^* = \{Z_{i,(t-1)K+k}^* : k = 1, \dots, K, t = 1, \dots, T-1\} \in \mathbb{R}^{(T-1)K}$  and  $\tilde{Z}^* \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i^{* \prime}$ . By A. van der Vaart and Wellner (1996, Theorem 3.6.2),

$$\{\tilde{Z}^* | \mathbf{X}_n\} \xrightarrow{d} \xi \sim N(0_{(T-1)K \times 1}, \Sigma_Z) \quad \text{a.s.} \quad (\text{B.16})$$

Let  $\hat{\delta}^* \equiv \hat{Z}^* - \tilde{Z}^*$  with  $\hat{Z}^*$  as in (3.18). For any  $t = 1, \dots, T-1$  and  $k = 1, \dots, K$ , then conditional on  $\mathbf{X}_n$ , the  $((t-1)K+k)$ -component of  $\hat{\delta}^*$  satisfies

$$\begin{aligned} & \sqrt{n\hat{P}(u_k)}[\hat{F}_t^*(u_k)\hat{F}_t(u_k) - (\hat{F}_{t+1}^*(u_k) - \hat{F}_{t+1}(u_k))] \\ & - \sqrt{nP(u_k)}[\hat{F}_t^*(u_k) - \hat{F}_t(u_k) - (\hat{F}_{t+1}^*(u_k) - \hat{F}_{t+1}(u_k))] \\ & = \sqrt{n}[\hat{F}_t^*(u_k) - \hat{F}_t(u_k) - (\hat{F}_{t+1}^*(u_k) - \hat{F}_{t+1}(u_k))](\sqrt{\hat{P}(u_k)} - \sqrt{P(u_k)}) \\ & \stackrel{(1)}{=} O_p(1)o(1) = o_p(1), \quad \text{w.p.a.1,} \end{aligned} \quad (\text{B.17})$$

where (1) holds by A. van der Vaart and Wellner (1996, Theorem 3.6.2) (which implies that  $\{\sqrt{n}[\hat{F}_t^*(u_k) - \hat{F}_t(u_k) - (\hat{F}_{t+1}^*(u_k) - \hat{F}_{t+1}(u_k))]| \mathbf{X}_n\} = O_p(1)$  a.s.) and that  $\hat{P}(u_k) - P(u_k) = o_p(1)$  by LLN.

Then, consider the following derivation. Conditional on  $\mathbf{X}_n$ ,

$$S_n^* \stackrel{(1)}{=} (\tilde{Z}^*)'(\tilde{Z}^*) + 2(\hat{\delta}^*)'(\tilde{Z}^*) + (\hat{\delta}^*)'(\hat{\delta}^*) \stackrel{(2)}{\xrightarrow{d}} S, \quad \text{w.p.a.1,} \quad (\text{B.18})$$

where (1) holds by  $\hat{\delta}^* = \hat{Z}^* - \tilde{Z}^*$  and (2) by (B.16) and (B.17). As a corollary of (B.18) and also by the condition that  $\Sigma_Z$  is a nonzero matrix, we deduce that  $P(S_n^* \leq x | \mathbf{X}_n) \xrightarrow{p} P(S \leq x)$  for all points  $x \in \mathbb{R}$ . From this point onward, the rest of the proof is identical to that of part (a) in Theorem 3.3.1.

**Case 2:**  $\Sigma_Z = 0_{(T-1)K \times (T-1)K}$ . This result holds by the same argument as in Theorem 3.3.1, except that  $\phi_n^A(\alpha)$  and  $c_n^A(1-\alpha)$  are replaced by  $\phi_n^B(\alpha)$  and  $c_n^B(1-\alpha)$ , respectively.

**Part (b).** By the proof of part (b) in Theorem 3.3.1,  $\hat{\Sigma}_Z^- = \Sigma_Z^{-1} + o_p(1)$ . Then, conditional on  $\mathbf{X}_n$ ,

$$\hat{\Sigma}_Z^- \rightarrow \Sigma_Z^{-1} \text{ w.p.a.1.} \quad (\text{B.19})$$

We can then repeat the arguments in part (a) to get that, conditional on  $\mathbf{X}_n$ ,

$$\hat{\delta}^* = o_p(1) \quad \text{and} \quad (\tilde{Z}^*)' \hat{\Sigma}_Z^- (\tilde{Z}^*) \xrightarrow{d} \xi' \Sigma_Z^{-1} \xi \sim \chi_{(T-1)K}^2 \text{ w.p.a.1.} \quad (\text{B.20})$$

From here, the desired result follows from the next derivation. Conditional on  $\mathbf{X}_n$ ,

$$\bar{S}_n^* \stackrel{(1)}{=} (\tilde{Z}^*)' \hat{\Sigma}_Z^- (\tilde{Z}^*) + 2(\hat{\delta}^*)' \hat{\Sigma}_Z^- (\tilde{Z}^*) + (\hat{\delta}^*)' \hat{\Sigma}_Z^- (\hat{\delta}^*) \stackrel{(2)}{\xrightarrow{d}} \chi_{(T-1)K}^2, \text{ w.p.a.1,} \quad (\text{B.21})$$

where (1) holds by  $\hat{\delta}^* = \hat{Z}^* - \tilde{Z}^*$ , and (2) by (B.19) and (B.20). As a corollary of (B.21), we have that  $P(\bar{S}_n^* \leq x | \mathbf{X}_n) \xrightarrow{p} P(\chi_{(T-1)K}^2 \leq x)$  for all  $x \in \mathbb{R}$ . From this point onward, the rest of the proof follows from arguments in part (a) in Theorem 3.3.1.  $\square$

**Theorem B.1.1.** *Let Assumption 9 hold, and let  $\hat{Z}^\pi$  be as in (3.24), where  $\pi$  is a uniformly chosen random permutation in  $\mathcal{M}^n$ . Then,*

$$P(\hat{Z}^\pi \leq x | \mathbf{X}_n) \xrightarrow{p} P(\xi \leq x) \quad (\text{B.22})$$

for all  $x$  such that  $P(\xi \leq x)$  is continuous, where  $\xi \sim N(0_{(T-1)K \times 1}, \Omega_Z^\pi)$  with

$$\Omega_Z^\pi \equiv \frac{1}{T!} \sum_{\pi \in \mathcal{M}} B(\pi) \Omega_Z B(\pi)', \quad (\text{B.23})$$

$\Omega_Z = E[Z_i Z_i']$  with  $Z_i$  as in (B.1), and  $\{B(\pi) \in \{-1, 0, 1\}^{(T-1)K \times (T-1)K} : \pi \in \mathcal{M}\}$  are known matrices defined within in the proof.

*Proof.* We divide the proof into several steps.

**Step 1.** Introduce suitable notation.

For each  $i = 1, \dots, n$  and  $k = 1, \dots, K$ , let

$$V_{i,k} \equiv \begin{bmatrix} 1(X_{i,1} \leq u_k) - 1(X_{i,2} \leq u_k) \\ \vdots \\ 1(X_{i,T-1} \leq u_k) - 1(X_{i,T} \leq u_k) \end{bmatrix} \quad \text{and} \quad V_i \equiv \begin{bmatrix} V_{i,1} \\ \vdots \\ V_{i,K} \end{bmatrix} \in \{-1, 0, 1\}^{(T-1)K \times 1}. \quad (\text{B.24})$$

Also, let

$$\begin{aligned} M &\equiv \text{diag}\{\sqrt{P(u_1)}, \dots, \sqrt{P(u_K)}\} \otimes I_{T-1} \in [0, 1]^{(T-1)K \times (T-1)K} \\ \hat{M} &\equiv \text{diag}\{\sqrt{\hat{P}(u_1)}, \dots, \sqrt{\hat{P}(u_K)}\} \otimes I_{T-1} \in [0, 1]^{(T-1)K \times (T-1)K}, \end{aligned} \quad (\text{B.25})$$

where  $\otimes$  denotes the Kronecker product. Note that  $Z_i = M \times V_i$  for all  $i = 1, \dots, n$ .

Let  $\pi^n = \{\pi_i : i = 1, \dots, n\} \in \mathcal{M}^n$  denote a fixed permutation. For any  $i = 1, \dots, n$  and  $k = 1, \dots, K$ , the  $\pi$ -permutation analogs of  $V_{i,k}$  and  $V_i$

$$V_{i,k}^\pi \equiv \begin{bmatrix} 1(X_{i,\pi_i(1)} \leq u_k) - 1(X_{i,\pi_i(2)} \leq u_k) \\ \vdots \\ 1(X_{i,\pi_i(T-1)} \leq u_k) - 1(X_{i,\pi_i(T)} \leq u_k) \end{bmatrix} \quad \text{and} \quad V_i^\pi \equiv \begin{bmatrix} V_{i,1}^\pi \\ \vdots \\ V_{i,K}^\pi \end{bmatrix} \in \{-1, 0, 1\}^{K(T-1) \times 1}. \quad (\text{B.26})$$

We note that  $\hat{M}$  is invariant to  $\pi^n$ . To see why, note that for each  $k = 1, \dots, K$ , and  $t = 1, \dots, T$ ,

$$\hat{P}^\pi(u_k) = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T 1(u_{k-1} < X_{i,\pi_i(t)} \leq u_k) \stackrel{(1)}{=} \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T 1(u_{k-1} < X_{i,t} \leq u_k) = \hat{P}(u_k), \quad (\text{B.27})$$

where (1) holds because the sum over  $t = 1, \dots, T$  is invariant across the permutation.

**Step 2.** For any  $i = 1, \dots, n$  and  $\pi_i \in \mathcal{M}$ , we define the matrix  $B(\pi_i)$  described in the statement and establish the following representation:

$$V_i^\pi = B(\pi_i) \times V_i. \quad (\text{B.28})$$

This result follows from expressing  $V_{i,k}^\pi$  as a particular linear combination of  $V_{i,k}$ . To see why, fix  $t = 1, \dots, T$  arbitrarily. If  $\pi_i(t) < \pi_i(t+1)$ , then we have

$$\begin{aligned} 1(X_{i,\pi_i(t)} \leq u_k) - 1(X_{i,\pi_i(t+1)} \leq u_k) &= \begin{bmatrix} 1(X_{i,\pi_i(t)} \leq u_k) - 1(X_{i,\pi(t)+1} \leq u_k) + \dots \\ + 1(X_{i,\pi_i(t+1)-1} \leq u_k) - 1(X_{i,\pi(t+1)} \leq u_k) \end{bmatrix} \\ &= (0, \dots, 0, 1, \dots, 1, 0, \dots, 0) \times V_{i,k}, \end{aligned}$$

where the ones are located at time periods corresponding to  $\pi_i(t), \dots, \pi_i(t+1) - 1$ . Con-

versely, if  $\pi_i(t) > \pi_i(t+1)$ , then we have

$$\begin{aligned} 1(X_{i,\pi_i(t)} \leq u_k) - 1(X_{i,\pi_i(t+1)} \leq u_k) &= \begin{bmatrix} -(1(X_{i,\pi_i(t)} \leq u_k) - 1(X_{i,\pi(t)+1} \leq u_k)) - \dots \\ -(1(X_{i,\pi_i(t+1)-1} \leq u_k) - 1(X_{i,\pi(t+1)} \leq u_k)) \end{bmatrix} \\ &= (0, \dots, 0, -1, \dots, -1, 0, \dots, 0) \times V_{i,k}, \end{aligned}$$

where the minus ones are located at time periods corresponding to  $\pi_i(t), \dots, \pi_i(t+1) - 1$ . Since  $\pi_i(t)$  and  $\pi_i(t+1)$  were arbitrarily chosen, we can define a matrix  $\tilde{B}(\pi_i) \in \{-1, 0, 1\}^{(T-1) \times (T-1)}$  such that  $V_{i,k}^\pi = \tilde{B}(\pi_i) \times V_{i,k}$ . By collecting results for  $k = 1, \dots, K$ , and setting

$$B(\pi_i) \equiv \text{diag}\{\tilde{B}(\pi_i), \dots, \tilde{B}(\pi_i)\}, \quad (\text{B.29})$$

(B.28) follows. Finally, by repeating this operation for all  $\pi = \pi_i \in \mathcal{M}$ , we define the collection of matrices  $\{B(\pi) \in \{-1, 0, 1\}^{(T-1)K \times (T-1)K} : \pi \in \mathcal{M}\}$ .

**Step 3.** Establish the Hoeffding's condition for  $\frac{1}{\sqrt{n}} \sum_{i=1}^n B(\pi_i) Z_i$ , where  $\boldsymbol{\pi}^n = \{\pi_i : i = 1, \dots, n\}$  denotes a randomly chosen permutation in  $\mathcal{M}^n$ . That is,

$$\left( \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\pi_i) Z_i, \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\tilde{\pi}_i) Z_i \right) \xrightarrow{d} (\xi, \tilde{\xi}), \quad (\text{B.30})$$

where  $\boldsymbol{\pi}^n = \{\pi_i : i = 1, \dots, n\}$  and  $\tilde{\boldsymbol{\pi}}^n = \{\tilde{\pi}_i : i = 1, \dots, n\}$  denote two mutually independent random permutations chosen uniformly from  $\mathcal{M}^n$  and independent of the data, and  $\xi$  and  $\tilde{\xi}$  are i.i.d.  $N(0_{(T-1)K \times 1}, \Omega_Z^\pi)$ .

We establish (B.30) using the Cramér-Wold device. That is, for arbitrary  $\lambda, \nu \in \mathbb{R}^{(T-1)K \times 1}$ , (B.30) follows from showing that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\lambda' B(\pi_i) + \nu' B(\tilde{\pi}_i)) Z_i \xrightarrow{d} N(0_{(T-1)K \times 1}, \lambda' \Omega_Z^\pi \lambda + \nu' \Omega_Z^\pi \nu). \quad (\text{B.31})$$

We begin the argument by showing that  $\{(\lambda' B(\pi_i) + \nu' B(\tilde{\pi}_i)) Z_i\}_{i=1}^n$  is an i.i.d. sequence. To see why, note that  $\{Z_i\}_{i=1}^n$  is i.i.d. by Assumption 9. Also, since  $\boldsymbol{\pi}^n = \{\pi_i : i = 1, \dots, n\}$  and  $\tilde{\boldsymbol{\pi}}^n = \{\tilde{\pi}_i : i = 1, \dots, n\}$  are defined as i.i.d. sequences, we conclude that  $\{B(\pi_i)\}_{i=1}^n$  and

$\{B(\tilde{\boldsymbol{\pi}}_i)\}_{i=1}^n$  are also i.i.d. By combining these facts, we get that  $\{(\lambda' B(\boldsymbol{\pi}_i) + \nu' B(\tilde{\boldsymbol{\pi}}_i)) Z_i\}_{i=1}^n$  is an i.i.d. sequence.

As a next step, we now show that

$$\sum_{\pi \in \mathcal{M}} B(\pi) = 0_{(T-1)K \times (T-1)K}. \quad (\text{B.32})$$

Since  $B(\pi) \equiv \text{diag}\{\tilde{B}(\pi), \dots, \tilde{B}(\pi)\}$ , it suffices to show that  $\sum_{\pi \in \mathcal{M}} \tilde{B}(\pi) = 0_{(T-1) \times (T-1)}$ . For each  $t = 1, \dots, T-1$ , denote the  $t$ 'th row of  $\tilde{B}(\pi)$ , denoted  $\tilde{B}_t(\pi)$ , can be expressed as follows:

$$\tilde{B}_t(\pi) = (0, \dots, 0, 1, \dots, 1, 0, \dots, 0) \times (1[\pi(t) < \pi(t+1)] - 1[\pi(t) > \pi(t+1)]),$$

where the sequence of ones appears in the positions  $\min\{\pi(t), \pi(t+1)\}$  through  $\max\{\pi(t), \pi(t+1)\} - 1$ . From here, we get that  $\sum_{\pi \in \mathcal{M}} \tilde{B}_t(\pi) = 0_{1 \times (T-1)}$ , as the occurrence of the vector  $(0, \dots, 0, 1, \dots, 1, 0, \dots, 0)$  in the sum over  $\mathcal{M}$  cancels with the corresponding vector  $(0, \dots, 0, -1, \dots, -1, 0, \dots, 0)$  when  $\pi(t)$  and  $\pi(t+1)$  are reversed.

Next, we show that  $E[B(\boldsymbol{\pi}_i)] = E[B(\tilde{\boldsymbol{\pi}}_i)] = 0_{(T-1)K \times (T-1)K}$  for all  $i = 1, \dots, n$ . To see why, fix  $i = 1, \dots, n$  arbitrarily and note that

$$E[B(\boldsymbol{\pi}_i)] \stackrel{(1)}{=} E[B(\tilde{\boldsymbol{\pi}}_i)] \stackrel{(2)}{=} \frac{1}{T!} \sum_{\pi \in \mathcal{M}} B(\pi) \stackrel{(3)}{=} 0_{(T-1)K \times (T-1)K}, \quad (\text{B.33})$$

where (1) holds because  $B(\boldsymbol{\pi}_i)$  and  $B(\tilde{\boldsymbol{\pi}}_i)$  are equally distributed, (2) because there are  $|\mathcal{M}| = (T!)$  possible permutations of  $\{1, 2, \dots, T\}$ , all equally likely, and (3) by (B.32).

Then, for all  $i = 1, \dots, n$ ,

$$E[(\lambda' B(\boldsymbol{\pi}_i) + \nu' B(\tilde{\boldsymbol{\pi}}_i)) Z_i] \stackrel{(1)}{=} E[(\lambda' E[B(\boldsymbol{\pi}_i)] + \nu' E[B(\tilde{\boldsymbol{\pi}}_i)]) Z_i] \stackrel{(2)}{=} 0, \quad (\text{B.34})$$

where (1) holds by  $Z_i \perp (B(\boldsymbol{\pi}_i), B(\tilde{\boldsymbol{\pi}}_i))$  and (2) by (B.33).

From here, note that for all  $i = 1, \dots, n$ ,

$$\begin{aligned}
V[(\lambda' B(\boldsymbol{\pi}_i) + \nu' B(\tilde{\boldsymbol{\pi}}_i)) Z_i] &= E[(\lambda' B(\boldsymbol{\pi}_i) + \nu' B(\tilde{\boldsymbol{\pi}}_i)) Z_i Z_i' (B(\boldsymbol{\pi}_i)' \lambda + B(\tilde{\boldsymbol{\pi}}_i)' \nu)] \\
&\stackrel{(1)}{=} E[(\lambda' B(\boldsymbol{\pi}_i) + \nu' B(\tilde{\boldsymbol{\pi}}_i)) \Omega_Z (B(\boldsymbol{\pi}_i)' \lambda + B(\tilde{\boldsymbol{\pi}}_i)' \nu)] \\
&\stackrel{(2)}{=} \lambda' E[B(\boldsymbol{\pi}_i) \Omega_Z B(\boldsymbol{\pi}_i)'] \lambda + \nu' E[B(\tilde{\boldsymbol{\pi}}_i) \Omega_Z B(\tilde{\boldsymbol{\pi}}_i)'] \nu, \\
&\stackrel{(3)}{=} \lambda' \Omega_Z^\pi \lambda + \nu' \Omega_Z^\pi \nu, \tag{B.35}
\end{aligned}$$

where (1) holds by  $Z_i \perp (B(\boldsymbol{\pi}_i), B(\tilde{\boldsymbol{\pi}}_i))$  and  $E[Z_i Z_i'] = \Omega_Z$ , (2) by (B.33) and that  $B(\boldsymbol{\pi}_i) \perp B(\tilde{\boldsymbol{\pi}}_i)$ , and (3) by (B.23) and that there are  $|\mathcal{M}| = (T!)$  possible permutations of  $\{1, 2, \dots, T\}$ , and all are equally likely.

To conclude the step, note that (B.31) follows from the CLT, as  $\{(\lambda' B(\boldsymbol{\pi}_i) + \nu' B(\tilde{\boldsymbol{\pi}}_i)) Z_i\}_{i=1}^n$  was shown to be an i.i.d. sequence that satisfies (B.34) and (B.35).

**Step 4.** Use the previous steps to conclude the proof.

By Chung and Romano (2016, Lemma A.1), (B.22) is equivalent to showing that  $\hat{Z}^\pi$  satisfies the following Hoeffding condition:

$$(\hat{Z}^\pi, \hat{Z}^{\tilde{\pi}}) \xrightarrow{d} (\xi, \tilde{\xi}), \tag{B.36}$$

where  $\hat{Z}^\pi$  and  $\hat{Z}^{\tilde{\pi}}$  are permuted according to  $\boldsymbol{\pi}^n = \{\boldsymbol{\pi}_i : i = 1, \dots, n\}$  and  $\tilde{\boldsymbol{\pi}}^n = \{\tilde{\boldsymbol{\pi}}_i : i = 1, \dots, n\}$ , respectively, which are two mutually independent random permutations chosen uniformly from  $\mathcal{M}^n$  and independent of the data, and  $\xi$  and  $\tilde{\xi}$  are i.i.d. according to  $N(0_{(T-1)K \times 1}, \Omega_Z^\pi)$ .

Before proving the desired result, we establish three preliminary results. First, by repeating the arguments in step 3 but with  $M$  replaced by  $I \in \mathbb{R}^{K(T-1) \times K(T-1)}$ , we have that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i = O_p(1) \quad \text{and} \quad \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\tilde{\boldsymbol{\pi}}_i) V_i = O_p(1). \tag{B.37}$$

Second, note that Assumption 9, the LLN, and the CMT imply that

$$\hat{\delta} \equiv \hat{M} - M = o_p(1). \tag{B.38}$$

Third, note that for any permutation  $\pi \in \mathcal{M}$ , we have

$$\begin{aligned}
M \times B(\pi) &= \text{diag}\{\sqrt{P(u_1)}\tilde{B}(\pi), \dots, \sqrt{P(u_K)}\tilde{B}(\pi)\} \\
&= \{\text{diag}\{\tilde{B}(\pi), \dots, \tilde{B}(\pi)\}\} \times (\text{diag}\{\sqrt{P(u_1)}, \dots, \sqrt{P(u_K)}\} \otimes I_{T-1}) \\
&= B(\pi) \times M.
\end{aligned} \tag{B.39}$$

The desired result follows from the next derivation.

$$\begin{aligned}
(\hat{Z}^\pi, \hat{Z}^{\tilde{\pi}}) &\stackrel{(1)}{=} \left( \frac{\hat{M}}{\sqrt{n}} \sum_{i=1}^n V_i^\pi, \frac{\hat{M}}{\sqrt{n}} \sum_{i=1}^n V_i^{\tilde{\pi}} \right) \\
&\stackrel{(2)}{=} \left( \frac{\hat{M}}{\sqrt{n}} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i, \frac{\hat{M}}{\sqrt{n}} \sum_{i=1}^n B(\tilde{\boldsymbol{\pi}}_i) V_i \right) \\
&\stackrel{(3)}{=} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\boldsymbol{\pi}_i) Z_i + \hat{\delta} \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i, \right. \\
&\quad \left. \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\tilde{\boldsymbol{\pi}}_i) Z_i + \hat{\delta} \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\tilde{\boldsymbol{\pi}}_i) V_i \right) \\
&\stackrel{(4)}{=} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\boldsymbol{\pi}_i) Z_i, \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\tilde{\boldsymbol{\pi}}_i) Z_i \right) + o_p(1) \\
&\stackrel{(5)}{\xrightarrow{d}} (\xi, \tilde{\xi}),
\end{aligned}$$

as desired, where (1) holds by (3.24), (B.26), and (B.27); (2) holds by (B.28); (3) by (B.39) and  $Z_i = MV_i$  for all  $i = 1, \dots, n$ ; (4) by (B.37) and (B.38); and (5) by (B.30).  $\square$

*Proof of Theorem 3.3.3.* Throughout the proof, we continuously invoke the results and notations from Theorem B.1.1. Recall that this result derives the asymptotic distribution of  $\hat{Z}^\pi$  as in (3.24), where  $\boldsymbol{\pi}$  is a uniformly chosen random permutation in  $\mathcal{M}^n$ . Recall from this theorem that  $\Omega_Z \equiv E[Z_i Z_i']$ , which equals  $\Sigma_Z \equiv \text{var}(Z_i)$  under  $H_0$  in (3.3);  $\Omega_Z^\pi$ , defined in (B.23), represents the asymptotic variance of randomization distribution;  $B(\pi)$ , defined below (B.29) for  $\pi \in \mathcal{M}$ , denotes a known matrix taking values in  $\{-1, 0, 1\}$ . With these notations in mind, we present the proof below.

**Part (a).** We divide the argument into two cases.

**Case 1:**  $\Sigma_Z \neq 0_{T(K-1) \times T(K-1)}$ . By  $T = 2$ , there are  $T! = 2$  permutations of  $(1, 2)$ , hence  $\mathcal{M} = \{(1, 2), (2, 1)\}$ . Following the construction in step 2 of Theorem B.1.1,  $B((1, 2)) =$

$I_{K \times K}$  and  $B((2, 1)) = -I_{K \times K}$ . Therefore,

$$\Omega_Z^\pi \stackrel{(1)}{=} \frac{1}{2}B(1, 2)\Omega_Z B(1, 2)' + \frac{1}{2}B((2, 1))\Omega_Z B((2, 1))' \stackrel{(2)}{=} \Omega_Z \stackrel{(3)}{=} \Sigma_Z,$$

where (1) holds by the definition of  $\Omega_Z^\pi$ , (2) holds by the definition of  $B(\pi)$ , and (3) follows by  $\Sigma_Z \equiv \text{var}(Z_i) = E[Z_i Z_i'] \equiv \Omega_Z$  under  $H_0$  in (3.3).

Theorem B.1.1 then implies that  $P(\hat{Z}^\pi \leq x | \mathbf{X}_n) \xrightarrow{P} P(\xi \leq x)$  for all  $x$  such that  $P(\xi \leq x)$  is continuous, where  $\xi \sim N(0_{(T-1)K \times 1}, \Sigma_Z)$ . Then, the continuous mapping theorem from Chung and Romano (2016, Lemma A.6) implies that for non-studentized statistic:

$$P(S_n^\pi \leq x | \mathbf{X}_n) \xrightarrow{P} P(S \leq x), \quad (\text{B.40})$$

for all  $x \in \mathbb{R}$ , where  $S$  is as in (3.12). This convergence relies on the fact that  $\Sigma_Z \neq 0_{T(K-1) \times T(K-1)}$ , which implies  $P(S \leq x)$  is continuous for all  $x \in \mathbb{R}$ . From this point onward, the rest of the proof follows from arguments in part (a) of Theorem 3.3.1.

**Case 2:**  $\Sigma_Z = 0_{T(K-1) \times T(K-1)}$ . By the same arguments as in Theorem 3.3.1, we have that  $1(X_{i,t} \leq u_k) = 1(X_{i,t+1} \leq u_k)$  a.s. for all  $t = 1, \dots, T-1$  and  $k = 1, \dots, K$ , and so  $S_n = 0$  a.s. Furthermore, for all  $\pi \in \mathcal{M}$ , we have that  $1(X_{i,\pi(t)} \leq u_k) = 1(X_{i,\pi(t+1)} \leq u_k)$  a.s. for all  $t = 1, \dots, T-1$ ,  $k = 1, \dots, K$ . This implies that  $S_n^\pi = 0$  a.s., and so  $S_n = c_n^\pi(1 - \alpha) = 0$ . The desired result follows from this and the construction of the test in (3.26).

**Part (b).** We construct an example with  $T = 3$ . For  $\tau_1, \tau_2 \in (0, 1)$ , we focus on a Markov chain with two states,  $s_1 = 0$  and  $s_2 = 1$ , and a transition matrix given by

$$\begin{pmatrix} P(X_{i,t+1} = 0 | X_{i,t} = 0) & P(X_{i,t+1} = 1 | X_{i,t} = 0) \\ P(X_{i,t+1} = 0 | X_{i,t} = 1) & P(X_{i,t+1} = 1 | X_{i,t} = 1) \end{pmatrix} = \begin{pmatrix} \tau_1 & 1 - \tau_1 \\ 1 - \tau_2 & \tau_2 \end{pmatrix}.$$

In the steady state, the marginal distribution is such that for all  $t = 1, 2, 3$ ,

$$P(X_{i,t} = 0) = \frac{1 - \tau_2}{2 - \tau_1 - \tau_2} \quad \text{and} \quad P(X_{i,t} = 1) = \frac{1 - \tau_1}{2 - \tau_1 - \tau_2}. \quad (\text{B.41})$$

To assess the marginal homogeneity of this Markov chain on only two support points, it suffices to test the hypothesis at one of the two points (as the other is just its complement).

For this reason, we construct our test statistic with  $K = 1$  and  $u_1 = 0$ . It follows then

$$\Omega_Z = \frac{(1 - \tau_1)(1 - \tau_2)^2}{(2 - \tau_1 - \tau_2)^2} \begin{bmatrix} 2 & -(2 - (\tau_1 + \tau_2)) \\ -(2 - (\tau_1 + \tau_2)) & 2 \end{bmatrix}.$$

By  $T = 3$ , the  $T! = 6$  permutations of  $(1, 2, 3)$  are

$$\mathcal{M} = \{(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)\}.$$

Following the construction in step 2 of Theorem B.1.1, we have

$$\begin{aligned} B((1, 2, 3)) &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & B((1, 3, 2)) &= \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}, & B((2, 1, 3)) &= \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \\ B((2, 3, 1)) &= \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}, & B((3, 1, 2)) &= \begin{bmatrix} -1 & -1 \\ 1 & 0 \end{bmatrix}, & B((3, 2, 1)) &= \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}, \end{aligned}$$

and

$$\Omega_Z^\pi = \frac{1}{6} \sum_{\pi \in \mathcal{M}} B(\pi) \Omega_Z B(\pi)'$$

It is not hard to verify that  $\Omega_Z^\pi$  is PD and  $\Omega_Z^\pi \neq \Omega_Z$ . By the same arguments as in part (a), we conclude that

$$P(S_n^\pi \leq x | \mathbf{X}_n) \xrightarrow{P} P(S^\pi \leq x), \quad (\text{B.42})$$

for all  $x \in \mathbb{R}$ , where  $S^\pi = \sum_{j=1}^2 \lambda_j^\pi \zeta_j^2$  with  $\{\zeta_j\}_{j=1}^2$  being i.i.d.  $N(0, 1)$ , and  $\{\lambda_j^\pi\}_{j=1}^2$  are the eigenvalues of  $\Omega_Z^\pi$ . From this point onward, we can repeat arguments in part (a) of Theorem 3.3.1 to show that

$$c_n^\pi(1 - \alpha) \xrightarrow{P} c^\chi(1 - \alpha; \Omega_Z^\pi) \quad \text{and} \quad E_P[\phi_n^\pi(\alpha)] \rightarrow P[S \leq c^\chi(1 - \alpha; \Omega_Z^\pi)], \quad (\text{B.43})$$

where  $S = \sum_{j=1}^2 \lambda_j \zeta_j^2$  with  $\{\lambda_j\}_{j=1}^2$  equal to the eigenvalues of  $\Sigma_Z = \Omega_Z$ . Since  $\Omega_Z^\pi \neq \Omega_Z$ , we have that  $c^\chi(1 - \alpha; \Omega_Z^\pi) \neq c^\chi(1 - \alpha; \Omega_Z)$  and therefore  $P[S \leq c^\chi(1 - \alpha; \Omega_Z^\pi)] \neq \alpha$ . To show the asymptotic overrejection, it suffices to find examples of parameters in which  $P[S \leq c^\chi(1 - \alpha; \Omega_Z^\pi)] > \alpha$ . For instance, by choosing  $\tau_1 = \tau_2 = 0.1$ , we obtain

$$c^\chi(0.9; \Omega_Z^\pi) = 1.56 < c^\chi(0.9; \Sigma_Z) = 2.36,$$

$$c^\chi(0.95; \Omega_Z^\pi) = 2.12 < c^\chi(0.95; \Sigma_Z) = 3.33,$$

$$c^\chi(0.99; \Omega_Z^\pi) = 3.49 < c^\chi(0.99; \Sigma_Z) = 5.72,$$

with the following asymptotic overrejection:

$$E_P[\phi_n^\pi(0.1)] \rightarrow 0.1828 > 0.1, \quad E_P[\phi_n^\pi(0.05)] \rightarrow 0.1194 > 0.05, \quad E_P[\phi_n^\pi(0.01)] \rightarrow 0.0446 > 0.01.$$

**Part (c).** Under  $H_0$ , we have  $\Omega_Z = \Sigma_Z$ . Then, Assumption 10 implies  $\Omega_Z$  is PD. Thus,  $B(\pi)\Omega_Z B(\pi)'$  is PSD for all  $\pi \in \mathcal{M}$ , where  $B(\pi)$  is defined in the proof of Theorem B.1.1. Furthermore, by choosing  $\pi = (1, \dots, T)$ , step 2 of Theorem B.1.1 implies that  $B((1, \dots, T)) = I_{(T-1)K \times (T-1)K}$ , and so  $B((1, \dots, T))\Omega_Z B((1, \dots, T))' = \Omega_Z$ . Then,  $\Omega_Z^\pi = \frac{1}{T!} \sum_{\pi \in \mathcal{M}} B(\pi)\Omega_Z B(\pi)'$  is PD, as it is the sum of PSD matrices with at least one PD matrix.

By Theorem B.1.1, Lemma B.1.1, and the fact that  $\Omega_Z^\pi$  is PD, the Slutsky's theorem from Chung and Romano (2016, Lemma A.5) implies that  $P(\bar{S}_n^\pi \leq x | \mathbf{X}_n) \xrightarrow{P} P(\chi_{(T-1)K}^2 \leq x)$  for all  $x \in \mathbb{R}$ . From this point onward, the rest of the proof follows from arguments in part (a) in Theorem 3.3.1.  $\square$

*Proof of Theorem 3.3.4. Part (a).* Fix  $(i, t, s) \in \{1, \dots, n\} \times \{1, \dots, T\} \times \{1, \dots, T\}$  arbitrarily, and let  $\pi \in \mathcal{M}$  be any permutation that interchanges  $t$  and  $s$ . For any  $x \in \mathbb{R}$ ,

$$\begin{aligned} F_t(x) &= P(X_{i,t} \leq x) \\ &= \lim_{\{u_j\}_{j \neq t} \rightarrow \infty} P(\{X_{i,1}, \dots, X_{i,t}, \dots, X_{i,s}, \dots, X_{i,T}\} \leq (u_1, \dots, x, \dots, u_s, \dots, u_T)) \\ &\stackrel{(1)}{=} \lim_{\{u_j\}_{j \neq t} \rightarrow \infty} P(\{X_{i,\pi(1)}, \dots, X_{i,\pi(t)}, \dots, X_{i,\pi(s)}, \dots, X_{i,\pi(T)}\} \leq (u_1, \dots, x, \dots, u_s, \dots, u_T)) \\ &\stackrel{(2)}{=} \lim_{\{u_j\}_{j \neq t} \rightarrow \infty} P(\{X_{i,\pi(1)}, \dots, X_{i,s}, \dots, X_{i,t}, \dots, X_{i,\pi(T)}\} \leq (u_1, \dots, x, \dots, u_s, \dots, u_T)) \\ &= P(X_{i,s} \leq x) = F_s(x), \end{aligned}$$

where (1) holds by  $P \in \Omega_{\text{TE}}$  and (2) by the specification of  $\pi$ . Since  $x \in \mathbb{R}$  and  $(t, s) \in \{1, \dots, T\} \times \{1, \dots, T\}$  were arbitrary,  $H_0$  in (3.3) holds.

To see that the reverse implication fails, consider the following example:

$$\mathbf{X}_n = \{(X_{i,1}, X_{i,2}, X_{i,3})\}_{i=1}^n \text{ i.i.d. with } X_{i,1} = X_{i,2} \perp X_{i,3}$$

and  $X_{i,t} \sim N(0, 1)$ . It is not hard to verify that this distribution satisfies Assumption 9 but does not belong to  $\Omega_{\text{TE}}$ .

**Part (b).** Let  $\mathbf{X}_n^\pi \equiv \{\{X_{i,\pi_i(t)}\}_{t=1}^T\}_{i=1}^n$  denote the sample permuted according to an arbitrary permutation  $\pi^n = \{\pi_i\}_{i=1}^n \in \mathcal{M}^n$ . Then,

$$F_{\mathbf{X}_n} \stackrel{(1)}{=} \prod_{i=1}^n F_{X_{i,1}, \dots, X_{i,T}} \stackrel{(2)}{=} \prod_{i=1}^n F_{X_{i,\pi_i(1)}, \dots, X_{i,\pi_i(T)}} \stackrel{(3)}{=} F_{\mathbf{X}_n^\pi}, \quad (\text{B.44})$$

where (1) and (3) hold by Assumption 9, and (2) by  $P \in \Omega_{\text{TE}}$ . We note that (B.44) implies that the randomization hypothesis (i.e., Lehmann and Romano (2022, Definition 17.2.1)) holds. From here, Lehmann and Romano (2022, Theorem 17.2.1) implies that the permutation test described in Lehmann and Romano (2022, Section 17.2.1) satisfies (3.28) with equality. In turn, this implies that our permutation test (i.e., the non-random version of the test in Lehmann and Romano (2022, Section 17.2.1)) satisfies (3.28).  $\square$

**Lemma B.1.1.** *Under Assumption 9,*

$$\hat{\Sigma}_{Z^\pi} \xrightarrow{P} \Omega_Z^\pi.$$

*Proof.* This proof relies on notation and arguments in the proof of Theorem B.1.1. First, we show that

$$\frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) Z_i Z_i' B(\boldsymbol{\pi}_i)' \xrightarrow{P} \Omega_Z^\pi. \quad (\text{B.45})$$

By similar arguments as in step 3 of Theorem B.1.1, we note that  $\{B(\boldsymbol{\pi}_i) Z_i Z_i' B(\boldsymbol{\pi}_i)'\}_{i=1}^n$  is i.i.d. with

$$E[B(\boldsymbol{\pi}_i) Z_i Z_i' B(\boldsymbol{\pi}_i)'] \stackrel{(1)}{=} E[B(\boldsymbol{\pi}_i) \Omega_Z B(\boldsymbol{\pi}_i)'] \stackrel{(2)}{=} \Omega_Z^\pi.$$

where (1) holds by  $\{B(\boldsymbol{\pi}_i)\}_{i=1}^n \perp \{Z_i\}_{i=1}^n$  and  $E[Z_i Z_i'] = \Omega_Z$ , and (2) by (B.23) and the fact that  $\boldsymbol{\pi}_i$  is uniformly distributed in  $\mathcal{M}$ . From these observations and the LLN, (B.45) follows.

Second, we note that

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) Z_i V_i' B(\boldsymbol{\pi}_i)' &= O_p(1) \\
\frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i Z_i' B(\boldsymbol{\pi}_i)' &= O_p(1) \\
\frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i V_i' B(\boldsymbol{\pi}_i)' &= O_p(1).
\end{aligned} \tag{B.46}$$

These can be shown by the arguments that yield (B.45), except that  $Z_i Z_i'$  is replaced by  $Z_i V_i'$ ,  $V_i Z_i'$ , and  $V_i V_i'$ , respectively.

To conclude the proof, consider the following derivation.

$$\begin{aligned}
\hat{\Sigma}_{Z\boldsymbol{\pi}} &\stackrel{(1)}{=} \frac{1}{n} \sum_{i=1}^n \hat{M} V_i \boldsymbol{\pi}' (\hat{M} V_i \boldsymbol{\pi})' - \left( \frac{1}{n} \sum_{i=1}^n \hat{M} V_i \boldsymbol{\pi}' \right) \left( \frac{1}{n} \sum_{i=1}^n \hat{M} V_i \boldsymbol{\pi}' \right)' \\
&\stackrel{(2)}{=} \hat{M} \left( \frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i V_i' B(\boldsymbol{\pi}_i)' \right) \hat{M}' - \hat{M} \left( \frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i \right) \left( \frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i \right)' \hat{M}' \\
&\stackrel{(3)}{=} (M + o_p(1)) \left( \frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i V_i' B(\boldsymbol{\pi}_i)' \right) (M + o_p(1))' + o_p(1) \\
&\stackrel{(4)}{=} \left[ \begin{aligned} &\frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) Z_i Z_i' B(\boldsymbol{\pi}_i)' + \left( \frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) Z_i V_i' B(\boldsymbol{\pi}_i)' \right) o_p(1) + \\ &o_p(1) \left( \frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i V_i' B(\boldsymbol{\pi}_i)' \right) + o_p(1) \left( \frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i Z_i' B(\boldsymbol{\pi}_i)' \right) o_p(1) \end{aligned} \right] + o_p(1) \\
&\stackrel{(5)}{=} \Omega_Z^\boldsymbol{\pi} + o_p(1),
\end{aligned}$$

as desired, where (1) holds by (B.25), (B.26), and (B.27), (2) by (B.28), (3) by (B.38) and  $\frac{1}{n} \sum_{i=1}^n B(\boldsymbol{\pi}_i) V_i = o_p(1)$  (implied by (B.37)), (4) by  $Z_i = M V_i$  for all  $i = 1, \dots, n$  and (B.39), and (5) by (B.45) and (B.46).  $\square$

**Lemma B.1.2.** *Let  $\Omega_{\text{FE}}$  denote the class of distributions that are “fully” exchangeable over units and time periods, i.e.,  $\mathbf{X}_n = \{\{X_{i,t}\}_{i=1}^n\}_{t=1}^T$  has the same distribution as  $\{\{X_{\lambda(i,t)}\}_{i=1}^n\}_{t=1}^T$  where  $\lambda$  denotes an arbitrary permutation of units and time periods. Under Assumption 9, the following statements hold:*

(a) If  $n = 1$ ,  $\Omega_{\text{FE}} = \Omega_{\text{TE}}$ .

(b) If  $n > 1$ ,  $\Omega_{\text{FE}} \subsetneq \Omega_{\text{TE}}$ .

*Proof.* Part (a) is straightforward, so we prove part (b). We begin by showing a useful intermediate result:  $\mathbf{X}_n \sim P \in \Omega_{\text{FE}}$  implies that its CDF can be written as

$$P(\mathbf{X}_n \leq y) = \prod_{i=1}^n \prod_{t=1}^T F(y_{i,t}) \quad \text{for all } y \in \mathbb{R}^{nT}, \quad (\text{B.47})$$

where  $F$  is the CDF of  $X_{1,1}$ . Since Assumption 9 already implies independence across units:

$$P(\mathbf{X}_n \leq y) = \prod_{i=1}^n \tilde{F}(\{y_{i,t}\}_{t=1}^T), \quad (\text{B.48})$$

where  $\tilde{F}$  is the CDF of the vector  $\{X_{1,t}\}_{t=1}^T$ . Then the desired result (B.47) follows immediately from (B.48) provided that  $\{X_{1,t}\}_{t=1}^T$  are i.i.d. with marginal CDF  $F$ . We now establish this result in two steps.

First, we show  $\{X_{1,t}\}_{t=1}^T$  is an independent sequence. To this end, fix  $\{x_t\}_{t=1}^T \in \mathbb{R}^T$  arbitrarily. For any  $s = 1, \dots, T-1$ , consider the following permutation:  $\lambda(1, t) = (1, t)$  for  $t \leq s$  and  $\lambda(1, t) = (2, t)$  for  $t > s$ ,  $\lambda(2, t) = (2, t)$  for  $t \leq s$  and  $\lambda(2, t) = (1, t)$  for  $t > s$ , and  $\lambda(j, t) = (j, t)$  for all  $j > 2$  and  $t = 1, \dots, T$ . Then,

$$\begin{aligned} & P(X_{1,1} \leq x_1, \dots, X_{1,s} \leq x_s, X_{1,s+1} \leq x_{s+1}, \dots, X_{1,T} \leq x_T) \\ &= \lim_{u \rightarrow \infty} P \left( \begin{array}{l} X_{1,1} \leq x_1, \dots, X_{1,s} \leq x_s, X_{1,s+1} \leq x_{s+1}, \dots, X_{1,T} \leq x_T \\ X_{j,t} \leq u \text{ for all } j \geq 2 \text{ and } t = 1, \dots, T \end{array} \right) \\ &\stackrel{(1)}{=} \lim_{u \rightarrow \infty} P \left( \begin{array}{l} X_{\lambda(1,1)} \leq x_1, \dots, X_{\lambda(1,s)} \leq x_s, X_{\lambda(1,s+1)} \leq x_{s+1}, \dots, X_{\lambda(1,T)} \leq x_T \\ X_{\lambda(j,t)} \leq u \text{ for all } j \geq 2 \text{ and } t = 1, \dots, T \end{array} \right) \\ &\stackrel{(2)}{=} \lim_{u \rightarrow \infty} P \left( \begin{array}{l} X_{1,1} \leq x_1, \dots, X_{1,s} \leq x_s, X_{2,s+1} \leq x_{s+1}, \dots, X_{2,T} \leq x_T \\ X_{\lambda(j,t)} \leq u \text{ for all } j \geq 2 \text{ and } t = 1, \dots, T \end{array} \right) \\ &= P(X_{1,1} \leq x_1, \dots, X_{1,s} \leq x_s, X_{2,s+1} \leq x_{s+1}, \dots, X_{2,T} \leq x_T) \\ &\stackrel{(3)}{=} P(X_{1,1} \leq x_1, \dots, X_{1,s} \leq x_s) P(X_{2,s+1} \leq x_{s+1}, \dots, X_{2,T} \leq x_T), \end{aligned} \quad (\text{B.49})$$

where (1) holds by  $P \in \Omega_{\text{FE}}$ , (2) by the specification of  $\lambda$ , and (3) by Assumption 9. By taking limits of (B.49) as  $x_1, \dots, x_s \rightarrow \infty$ , we get

$$P(X_{1,s+1} \leq x_{s+1}, \dots, X_{1,T} \leq x_T) = P(X_{2,s+1} \leq x_{s+1}, \dots, X_{2,T} \leq x_T). \quad (\text{B.50})$$

Since (B.50) holds for all  $\{x_t\}_{t=1}^T \in \mathbb{R}^T$ , we can combine it with (B.49) to get

$$(X_{1,1}, \dots, X_{1,s}) \perp (X_{1,s+1}, \dots, X_{1,T}) \text{ for any } s = 1, \dots, T-1. \quad (\text{B.51})$$

The desired result follows by considering (B.51) sequentially for  $s = 1$ ,  $s = 2$ , and so on.

Second, we show that  $\{X_{1,t}\}_{t=1}^T$  is an identically distributed sequence. To this end, fix  $x \in \mathbb{R}$  arbitrarily. For any  $s \neq 1$ , consider the following permutation:  $\lambda(1,1) = (1,s)$ ,  $\lambda(1,s) = \lambda(1,1)$ , and  $\lambda(i,t) = (i,t)$  otherwise. Then,

$$\begin{aligned} P(X_{1,1} \leq x) &= \lim_{u \rightarrow \infty} P(X_{1,1} \leq x, X_{i,t} \leq u \text{ for all } (i,t) \neq (1,1)) \\ &\stackrel{(1)}{=} \lim_{u \rightarrow \infty} P(X_{\lambda(1,1)} \leq x, X_{\lambda(i,t)} \leq u \text{ for all } (i,t) \neq (1,1)) \\ &\stackrel{(2)}{=} \lim_{u \rightarrow \infty} P(X_{1,s} \leq x, X_{\lambda(i,t)} \leq u \text{ for all } (i,t) \neq (1,1)) \\ &= P(X_{1,s} \leq x), \end{aligned} \quad (\text{B.52})$$

where (1) holds by  $P \in \Omega_{\text{FE}}$  and (2) by the specification of  $\lambda$ . Since (B.52) holds for all  $x \in \mathbb{R}$ ,  $X_{1,1}$  and  $X_{1,s}$  have the same distribution. Since the choice of  $s = 2, \dots, T$  was arbitrary, the desired result follows.

Finally, we conclude the proof by finding a distribution  $P \in \Omega_{\text{TE}}$  but  $P \notin \Omega_{\text{FE}}$  so the inclusion is strict. Consider the following example:  $\mathbf{X}_n = \{\{X_{i,t}\}_{i=1}^n\}_{t=1}^T$  with  $n = 2$ ,  $T = 2$ , where  $X_{1,1} = X_{1,2} = Z_1$ ,  $X_{2,1} = X_{2,2} = Z_2$ , and  $\{Z_1, Z_2\}$  are i.i.d.  $N(0,1)$ . It is trivial to see that this distribution satisfies Assumption 9 and  $P \in \Omega_{\text{TE}}$ , however  $P \notin \Omega_{\text{FE}}$ .  $\square$

## Appendix C. Additional Results for Chapter 4

### C.1 Bound Expressions

In this section we summarize all the bounds that will be used in the proofs of lemmas and main results. These expressions are given for an arbitrary values of  $(y, w) \in \mathbb{R} \times \text{supp}(W)$ . Denote  $\underline{c} := \underline{c}(w, \eta)$  and  $\bar{c} := \bar{c}(w, \eta)$ , where the dependence on  $w$  and  $\eta$  is implicitly understood.

#### C.1.1 Lower bound on $Y_1$

Let

$$\underline{\tau}_1 = \frac{(p_{1|w} - \underline{c})\bar{c}}{p_{1|w}(\bar{c} - \underline{c})} \quad \text{and} \quad \underline{Q}_1 = Q_{Y|X,W}(\underline{\tau}_1 | 1, w)$$

and

$$\begin{aligned} & \underline{E}_{Y_1|W}(y | w) \\ &= \max \left\{ F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\bar{c}}, \frac{\underline{c} - p_{1|w}}{\underline{c}} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\underline{c}} \right\} \\ &= F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\bar{c}} \mathbb{1}(y < \underline{Q}_1) + \left( \frac{\underline{c} - p_{1|w}}{\underline{c}} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\underline{c}} \right) \mathbb{1}(y \geq \underline{Q}_1) \end{aligned}$$

$$\begin{aligned} & \underline{E}_{Y_1|X,W}(y | 0, w) \\ &= \max \left\{ F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \bar{c})}{p_{0|w}\bar{c}}, \frac{\underline{c} - p_{1|w}}{\underline{c}p_{0|w}} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \underline{c})}{p_{0|w}\underline{c}} \right\} \\ &= F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \bar{c})}{p_{0|w}\bar{c}} \mathbb{1}(y < \underline{Q}_1) + \left( \frac{\underline{c} - p_{1|w}}{\underline{c}p_{0|w}} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \underline{c})}{p_{0|w}\underline{c}} \right) \mathbb{1}(y \geq \underline{Q}_1) \end{aligned}$$

$$\underline{E}_{Y_1|X,W}(y | 1, w) = F_{Y|X,W}(y | 1, w).$$

Define

$$\underline{p}_1(y, w) = \bar{c} \mathbb{1}(y < \underline{Q}_1) + \underline{A}_1 \mathbb{1}(y = \underline{Q}_1) + \underline{c} \mathbb{1}(y > \underline{Q}_1)$$

where

$$\underline{A}_1 = \frac{\mathbb{P}(Y = \underline{Q}_1, X = 1 | W = w)}{\left( \frac{\underline{c} - p_{1|w}}{\underline{c}} + F_{Y|X,W}(\underline{Q}_1 | 1, w) \frac{p_{1|w}}{\underline{c}} \right) - \frac{p_{1|w}}{\bar{c}} \mathbb{P}(Y < \underline{Q}_1 | X = 1, W = w)}.$$

If the denominator is 0, set  $\underline{A}_1 = p_{1|w}$ . We can also derive the associated quantiles function bounds for  $\tau \in (0, 1)$ :

$$\bar{Q}_{Y_1|W}(\tau | w) := \underline{F}_{Y_1|W}^{-1}(\tau | w) = Q_{Y|X,W} \left( \min \left\{ \frac{\bar{c}}{p_{1|w}} \tau, \frac{p_{1|w} - \underline{c}}{p_{1|w}} + \frac{\underline{c}}{p_{1|w}} \tau \right\} \mid 1, w \right)$$

$$\bar{Q}_{Y_1|X,W}(\tau | 0, w) := \underline{F}_{Y_1|X}^{-1}(\tau | 0, w) = Q_{Y|X,W} \left( \min \left\{ \frac{\bar{c}p_{0|w}}{p_{1|w}(1 - \bar{c})} \tau, \frac{p_{1|w} - \underline{c}}{p_{1|w}(1 - \underline{c})} + \frac{\underline{c}p_{0|w}}{p_{1|w}(1 - \underline{c})} \tau \right\} \mid 1, w \right).$$

### C.1.2 Upper bound on $Y_1$

Let

$$\bar{\tau}_1 = 1 - \underline{\tau}_1 \quad \text{and} \quad \bar{Q}_1 = Q_{Y|X,W}(\bar{\tau}_1 | 1, w)$$

and

$$\begin{aligned} \bar{F}_{Y_1|W}(y | w) &= \min \left\{ F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\underline{c}}, \frac{\bar{c} - p_{1|w}}{\bar{c}} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\bar{c}} \right\} \\ &= F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\underline{c}} \mathbb{1}(y < \bar{Q}_1) + \left( \frac{\bar{c} - p_{1|w}}{\bar{c}} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}}{\bar{c}} \right) \mathbb{1}(y \geq \bar{Q}_1) \end{aligned}$$

$$\begin{aligned} \bar{F}_{Y_1|X,W}(y | 0, w) &= \min \left\{ F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \underline{c})}{p_{0|w}\underline{c}}, \frac{\bar{c} - p_{1|w}}{\bar{c}p_{0|w}} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \bar{c})}{p_{0|w}\bar{c}} \right\} \\ &= F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \underline{c})}{(1 - p_{1|w})\underline{c}} \mathbb{1}(y < \bar{Q}_1) + \left( \frac{\bar{c} - p_{1|w}}{\bar{c}p_{0|w}} + F_{Y|X,W}(y | 1) \frac{p_{1|w}(1 - \bar{c})}{p_{0|w}\bar{c}} \right) \mathbb{1}(y \geq \bar{Q}_1) \end{aligned}$$

$$\bar{F}_{Y_1|X,W}(y | 1, w) = F_{Y|X,W}(y | 1, w).$$

Define

$$\bar{p}_1(y, w) = \underline{c} \mathbb{1}(y < \bar{Q}_1) + \bar{A}_1 \mathbb{1}(y = \bar{Q}_1) + \bar{c} \mathbb{1}(y > \bar{Q}_1)$$

where

$$\bar{A}_1 = \frac{\mathbb{P}(Y = \bar{Q}_1, X = 1 | W = w)}{\left( \frac{\bar{c} - p_{1|w}}{\bar{c}} + F_{Y|X,W}(\bar{Q}_1 | 1, w) \frac{p_{1|w}}{\bar{c}} \right) - \frac{p_{1|w}}{\underline{c}} \mathbb{P}(Y < \bar{Q}_1 | X = 1, W = w)}.$$

If the denominator is 0, set  $\bar{A}_1 = p_{1|w}$ . We can also derive the associated quantiles function bounds for  $\tau \in (0, 1)$ :

$$\underline{Q}_{Y_1|W}(\tau | w) := \bar{F}_{Y_1|W}^{-1}(\tau | w) = Q_{Y|X,W} \left( \max \left\{ \frac{\underline{c}}{p_{1|w}} \tau, \frac{p_{1|w} - \bar{c}}{p_{1|w}} + \frac{\bar{c}}{p_{1|w}} \tau \right\} \mid 1, w \right)$$

$$\underline{Q}_{Y_1|X,W}(\tau | 0, w) := \bar{F}_{Y_1|X,W}^{-1}(\tau | 0, w) = Q_{Y|X,W} \left( \max \left\{ \frac{\underline{c}p_{0|w}}{p_{1|w}(1 - \underline{c})} \tau, \frac{p_{1|w} - \bar{c}}{p_{1|w}(1 - \bar{c})} + \frac{\bar{c}p_{0|w}}{p_{1|w}(1 - \bar{c})} \tau \right\} \mid 1, w \right).$$

### C.1.3 Lower bound on $Y_0$

Under  $c$ -dependence, we have that  $\mathbb{P}(X = 0 | Y_0, W = w) = 1 - \mathbb{P}(X = 1 | Y_0, W = w) \in [1 - \bar{c}(w, \eta), 1 - \underline{c}(w, \eta)]$ . So we can take the bound expressions for  $F_{Y_1|W}$ , swap  $p_{1|w}$  and  $p_{0|w}$ , and swap  $(\underline{c}, \bar{c})$  and  $(1 - \bar{c}, 1 - \underline{c})$  and get the correct expressions for the bounds for  $F_{Y_0|W}$ . Here are all the expressions.

Let

$$\underline{\tau}_0 = \frac{(\bar{c} - p_{1|w})(1 - \underline{c})}{p_{0|w}(\bar{c} - \underline{c})} \quad \text{and} \quad \underline{Q}_0 = Q_{Y|X,W}(\underline{\tau}_0 | 0, w)$$

and

$$\begin{aligned} & \underline{F}_{Y_0|W}(y | w) \\ &= \max \left\{ F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1 - \underline{c}}, \frac{p_{1|w} - \bar{c}}{1 - \bar{c}} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1 - \bar{c}} \right\} \\ &= F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1 - \underline{c}} \mathbb{1}(y < \underline{Q}_0) + \left( \frac{p_{1|w} - \bar{c}}{1 - \bar{c}} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1 - \bar{c}} \right) \mathbb{1}(y \geq \underline{Q}_0) \end{aligned}$$

$$\begin{aligned} & \underline{F}_{Y_0|X,W}(y | 1, w) \\ &= \max \left\{ F_{Y|X,W}(y | 0, w) \frac{p_{0|w}\underline{c}}{p_{1|w}(1 - \underline{c})}, \frac{p_{1|w} - \bar{c}}{(1 - \bar{c})p_{1|w}} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}\bar{c}}{p_{1|w}(1 - \bar{c})} \right\} \\ &= F_{Y|X,W}(y | 0, w) \frac{p_{0|w}\underline{c}}{p_{1|w}(1 - \underline{c})} \mathbb{1}(y < \underline{Q}_0) + \left( \frac{p_{1|w} - \bar{c}}{(1 - \bar{c})p_{1|w}} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}\bar{c}}{p_{1|w}(1 - \bar{c})} \right) \mathbb{1}(y \geq \underline{Q}_0) \end{aligned}$$

$$\underline{F}_{Y_0|X,W}(y | 0, w) = F_{Y|X}(y | 0, w).$$

Define

$$\underline{p}_0(y, w) = \underline{c} \mathbb{1}(y < \underline{Q}_0) + (1 - \underline{A}_0) \mathbb{1}(y = \underline{Q}_0) + \bar{c} \mathbb{1}(y > \underline{Q}_0)$$

where

$$\underline{A}_0 = \frac{p_{0|w}(F_{Y|X,W}(\underline{Q}_0|0, w) - F_{Y|X,W}(\underline{Q}_0 - |0, w))}{\underline{F}_{Y_0|W}(\underline{Q}_0|w) - \underline{F}_{Y_0|W}(\underline{Q}_0 - |w)}.$$

If the denominator is 0, set  $\underline{A}_0 = p_{0|w}$ . We can also derive the associated quantiles function bounds for  $\tau \in (0, 1)$ :

$$\bar{Q}_{Y_0|W}(\tau | w) := \underline{F}_{Y_0|W}^{-1}(\tau) = Q_{Y|X,W} \left( \min \left\{ \frac{1-\underline{c}}{p_{0|w}}\tau, \frac{p_{0|w} - (1-\bar{c})}{p_{0|w}} + \frac{1-\bar{c}}{p_{0|w}}\tau \right\} | 0, w \right)$$

$$\bar{Q}_{Y_0|X,W}(\tau | 1, w) := \underline{F}_{Y_0|X}^{-1}(\tau | 1) = Q_{Y|X,W} \left( \min \left\{ \frac{(1-\underline{c})p_{1|w}}{p_{0|w}\underline{c}}\tau, \frac{p_{0|w} - (1-\bar{c})}{p_{0|w}\bar{c}} + \frac{p_{1|w}(1-\bar{c})}{p_{0|w}\bar{c}}\tau \right\} | 0, w \right).$$

### C.1.4 Upper bound for $Y_0$

Let

$$\bar{\tau}_0 = 1 - \tau_0 \quad \text{and} \quad \bar{Q}_0 = Q_{Y|X,W}(\bar{\tau}_0|0, w)$$

and

$$\bar{F}_{Y_0|W}(y | w)$$

$$= \min \left\{ F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1-\bar{c}}, \frac{p_{1|w} - \underline{c}}{1-\underline{c}} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1-\underline{c}} \right\}$$

$$= F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1-\bar{c}} \mathbb{1}(y < \bar{Q}_0) + \left( \frac{p_{1|w} - \underline{c}}{1-\underline{c}} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}}{1-\underline{c}} \right) \mathbb{1}(y \geq \bar{Q}_0)$$

$$\bar{F}_{Y_0|X,W}(y | 1, w)$$

$$= \min \left\{ F_{Y|X,W}(y | 0, w) \frac{p_{0|w}\bar{c}}{p_{1|w}(1-\bar{c})}, \frac{p_{1|w} - \underline{c}}{(1-\underline{c})p_{1|w}} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}\underline{c}}{p_{1|w}(1-\underline{c})} \right\}$$

$$= F_{Y|X,W}(y | 0, w) \frac{p_{0|w}\bar{c}}{p_{1|w}(1-\bar{c})} \mathbb{1}(y < \bar{Q}_0) + \left( \frac{p_{1|w} - \underline{c}}{(1-\underline{c})p_{1|w}} + F_{Y|X,W}(y | 0, w) \frac{p_{0|w}\underline{c}}{p_{1|w}(1-\underline{c})} \right) \mathbb{1}(y \geq \bar{Q}_0).$$

Define

$$\bar{p}_0(y, w) = \bar{c} \mathbb{1}(y < \bar{Q}_0) + (1 - \bar{A}_0) \mathbb{1}(y = \bar{Q}_0) + \underline{c} \mathbb{1}(y > \bar{Q}_0)$$

where

$$\bar{A}_0 = \frac{\mathbb{P}(Y = \bar{Q}_0, X = 0 | W = w)}{\left( \frac{p_{1|w} - \underline{c}}{1-\underline{c}} + F_{Y|X,W}(\bar{Q}_0|0, w) \frac{p_{0|w}}{1-\underline{c}} \right) - \frac{p_{0|w}}{1-\bar{c}} \mathbb{P}(Y < \bar{Q}_0 | X = 0, W = w)}.$$

If the denominator is 0, set  $\bar{A}_0 = p_{0|w}$ . We can also derive the associated quantiles function bounds for  $\tau \in (0, 1)$ :

$$\begin{aligned} \underline{Q}_{Y_0|W}(\tau | w) &:= \bar{F}_{Y_0|W}^{-1}(\tau | w) \\ &= Q_{Y|X,W} \left( \max \left\{ \frac{1 - \bar{c}}{p_{0|w}} \tau, \frac{p_{0|w} - (1 - \underline{c})}{p_{0|w}} + \frac{1 - \underline{c}}{p_{0|w}} \tau \right\} \mid 0, w \right) \end{aligned}$$

$$\begin{aligned} \underline{Q}_{Y_0|X,W}(\tau | 1, w) &:= \bar{F}_{Y_0|X,W}^{-1}(\tau | 1, w) \\ &= Q_{Y|X,W} \left( \max \left\{ \frac{(1 - \bar{c})p_{1|w}}{p_{0|w}\bar{c}} \tau, \frac{p_{0|w} - (1 - \underline{c})}{p_{0|w}\underline{c}} + \frac{(1 - \underline{c})p_{1|w}}{p_{0|w}\underline{c}} \tau \right\} \mid 0, w \right). \end{aligned}$$

## C.2 Proofs for Section 4.3

*Proof of Lemma 4.3.1.* Let  $x \in \{0, 1\}$  and fix  $w \in \text{supp}(W)$ . By the law of iterated expectations,  $\mathbb{E}[X | Y_x, W = w] = \mathbb{E}[\mathbb{E}[X | Y_1, Y_0, W = w] | Y_x, W = w]$ . Since  $\mathbb{E}[X | Y_1, Y_0, W = w] \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$  almost surely, we then have that  $\mathbb{E}[X | Y_x, W = w] \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$  almost surely as well.  $\square$

*Proof of Proposition 4.3.1. Part 1:* Suppose marginal  $c$ -dependence holds with bound functions  $[\underline{c}(w, \eta), \bar{c}(w, \eta)]$ . Fix  $(x, w) \in \{0, 1\} \times \text{supp}(W)$ . We have that

$$R_x(Y_x, w) = \frac{p_x(Y_x, w)}{1 - p_x(Y_x, w)} \Big/ \frac{p_{1|w}}{1 - p_{1|w}} \in \left[ \frac{\underline{c}(w, \eta)}{1 - \underline{c}(w, \eta)} \Big/ \frac{p_{1|w}}{1 - p_{1|w}}, \frac{\bar{c}(w, \eta)}{1 - \bar{c}(w, \eta)} \Big/ \frac{p_{1|w}}{1 - p_{1|w}} \right], \quad (\text{C.1})$$

where the inclusion holds from the mapping  $a \mapsto a/(1 - a)$  being strictly increasing over  $a \in (0, 1)$  and from  $p_x(Y_x, w) \in [\underline{c}(w, \eta), \bar{c}(w, \eta)] \subset (0, 1)$  almost surely. We note that

$$\frac{\bar{c}(w, \eta)}{1 - \bar{c}(w, \eta)} \Big/ \frac{p_{1|w}}{1 - p_{1|w}} \in \left[ \frac{p_{1|w}}{1 - p_{1|w}} \Big/ \frac{p_{1|w}}{1 - p_{1|w}}, +\infty \right) = [1, +\infty)$$

where the inclusion holds from  $\bar{c}(w, \eta) \in [p_{1|w}, 1)$ . Similarly,

$$\frac{\underline{c}(w, \eta)}{1 - \underline{c}(w, \eta)} \Big/ \frac{p_{1|w}}{1 - p_{1|w}} \in \left( \frac{0}{1 - 0} \Big/ \frac{p_{1|w}}{1 - p_{1|w}}, \frac{p_{1|w}}{1 - p_{1|w}} \Big/ \frac{p_{1|w}}{1 - p_{1|w}} \right] = (0, 1]$$

from  $\underline{c}(w, \eta) \in (0, p_{1|w}]$ . We conclude that the GSM holds with the bound functions from equation (4.4). Replacing marginal  $c$ -dependence with joint  $c$ -dependence delivers the same bounds for the GJSM.

**Part 2:** Suppose the GSM holds with bound functions  $[\underline{\Lambda}(w, \eta), \bar{\Lambda}(w, \eta)]$ . Fix  $(x, w) \in \{0, 1\} \times \text{supp}(W)$ . We have that

$$p_x(Y_x, w) = \frac{p_{1|w} R_x(Y_x, w)}{p_{0|w} + p_{1|w} R_x(Y_x, w)} \in \left[ \frac{p_{1|w} \underline{\Lambda}(w, \eta)}{p_{0|w} + p_{1|w} \underline{\Lambda}(w, \eta)}, \frac{p_{1|w} \bar{\Lambda}(w, \eta)}{p_{0|w} + p_{1|w} \bar{\Lambda}(w, \eta)} \right]. \quad (\text{C.2})$$

The equality holds from inverting the equation  $R_x(Y_x, w) = \frac{p_x(Y_x, w)}{1 - p_x(Y_x, w)} / \frac{p_{1|w}}{1 - p_{1|w}}$  in  $p_x(Y_x, w)$ .

The inclusion holds from the mapping  $a \mapsto a/(1+a)$  being strictly increasing for  $a \in [0, +\infty)$  and from  $R_x(Y_x, w) \in [\underline{\Lambda}(w, \eta), \bar{\Lambda}(w, \eta)] \subset (0, +\infty)$  almost surely. We note that

$$\frac{p_{1|w} \underline{\Lambda}(w, \eta)}{p_{0|w} + p_{1|w} \underline{\Lambda}(w, \eta)} \in \left( \frac{p_{1|w} \cdot 0}{p_{0|w} + p_{1|w} \cdot 0}, \frac{p_{1|w} \cdot 1}{p_{0|w} + p_{1|w} \cdot 1} \right] = (0, p_{1|w}]$$

by  $\underline{\Lambda}(w, \eta) \in (0, 1]$ . Similarly,

$$\frac{p_{1|w} \bar{\Lambda}(w, \eta)}{p_{0|w} + p_{1|w} \bar{\Lambda}(w, \eta)} \in \left[ \frac{p_{1|w} \cdot 1}{p_{0|w} + p_{1|w} \cdot 1}, 1 \right) = [p_{1|w}, 1)$$

by  $\bar{\Lambda}(w, \eta) \in [1, +\infty)$ . We conclude that marginal  $c$ -dependence holds with the bound functions from equation (1). Replacing the GSM with the GJSM delivers the same bounds for joint  $c$ -dependence.  $\square$

### C.3 Proofs for Section 4.4

*Proof of Lemma 4.4.1.* By Lemma 4.3.1, it suffices to show the desired results under Assumption 12. Let  $y \in \mathbb{R}$  and  $w \in \text{supp}(W)$  be fixed. Note that

$$\begin{aligned} \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)}{p_1(Y, w)} \mid X = 1, W = w \right] p_{1|w} &= \mathbb{E} \left[ \frac{\mathbb{1}(Y_1 \leq y) X}{p_1(Y_1, w)} \mid W = w \right] \\ &= \mathbb{E} \left[ \frac{\mathbb{1}(Y_1 \leq y) \mathbb{E}[X \mid Y_1, W = w]}{p_1(Y_1, w)} \mid W = w \right] \\ &= \mathbb{E} [\mathbb{1}(Y_1 \leq y) \mid W = w] \\ &= F_{Y_1|W}(y \mid w), \end{aligned}$$

where the second equality follows from the law of iterated expectations and the third from  $p_1(Y_1, w) \geq \underline{c}(w, \eta) > 0$  almost surely by Assumption 12. Likewise, we have

$$\mathbb{E} \left[ \frac{\mathbb{1}(Y > y)}{p_1(Y, w)} \mid X = 1, W = w \right] p_{1|w} = 1 - F_{Y_1|W}(y \mid w).$$

Therefore,

$$\begin{aligned} F_{Y_1|W}(y \mid w) &= \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)}{p_1(Y, w)} \mid X = 1, W = w \right] p_{1|w} \\ &\leq \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)}{\underline{c}(w, \eta)} \mid X = 1, W = w \right] p_{1|w} \\ &= F_{Y|X,W}(y \mid 1, w) \frac{p_{1|w}}{\underline{c}(w, \eta)} \end{aligned}$$

and

$$\begin{aligned} F_{Y_1|W}(y \mid w) &= 1 - \mathbb{P}(Y_1 > y \mid W = w) \\ &= 1 - \mathbb{E} \left[ \frac{\mathbb{1}(Y > y)}{p_1(Y, w)} \mid X = 1, W = w \right] p_{1|w} \\ &\leq 1 - \mathbb{E} \left[ \frac{\mathbb{1}(Y > y)}{\bar{c}(w, \eta)} \mid X = 1, W = w \right] p_{1|w} \\ &= \frac{\bar{c}(w, \eta) - p_{1|w}}{\bar{c}(w, \eta)} + F_{Y|X,W}(y \mid 1, w) \frac{p_{1|w}}{\bar{c}(w, \eta)}. \end{aligned}$$

The inequalities follow from  $p_1(Y_1, w)^{-1} \in [\bar{c}(w, \eta)^{-1}, \underline{c}(w, \eta)^{-1}]$  almost surely. By these two inequalities,

$$\begin{aligned} F_{Y_1|W}(y \mid w) &\leq \min \left\{ F_{Y|X,W}(y \mid 1, w) \frac{p_{1|w}}{\underline{c}(w, \eta)}, \frac{\bar{c}(w, \eta) - p_{1|w}}{\bar{c}(w, \eta)} + F_{Y|X,W}(y \mid 1, w) \frac{p_{1|w}}{\bar{c}(w, \eta)} \right\} \\ &= \bar{F}_{Y_1|W}(y \mid w). \end{aligned}$$

Similarly,

$$\begin{aligned} F_{Y_1|W}(y \mid w) &= \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)}{p_1(Y, w)} \mid X = 1, W = w \right] p_{1|w} \\ &\geq \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)}{\bar{c}(w, \eta)} \mid X = 1, W = w \right] p_{1|w} \\ &= F_{Y|X,W}(y \mid 1, w) \frac{p_{1|w}}{\bar{c}(w, \eta)} \end{aligned}$$

and

$$\begin{aligned}
F_{Y_1|W}(y | w) &= 1 - \mathbb{P}(Y_1 > y | W = w) \\
&= 1 - \mathbb{E} \left[ \frac{\mathbb{1}(Y > y)}{p_1(Y, w)} \mid X = 1, W \right] p_{1|w} \\
&\geq 1 - \mathbb{E} \left[ \frac{\mathbb{1}(Y > y)}{\underline{c}(w, \eta)} \mid X = 1, W = w \right] p_{1|w} \\
&= \frac{\underline{c}(w, \eta) - p_{1|w}}{\underline{c}(w, \eta)} + F_{Y|X, W}(y | 1, w) \frac{p_{1|w}}{\underline{c}(w, \eta)}.
\end{aligned}$$

By these two inequalities,

$$\begin{aligned}
F_{Y_1|W}(y | w) &\geq \max \left\{ F_{Y|X, W}(y | 1, w) \frac{p_{1|W}}{\bar{c}(w, \eta)}, \frac{\underline{c}(w, \eta) - p_{1|w}}{\underline{c}(w, \eta)} + F_{Y|X, W}(y | 1, w) \frac{p_{1|w}}{\underline{c}(w, \eta)} \right\} \\
&= \underline{F}_{Y_1|W}(y | w).
\end{aligned}$$

Therefore we have established  $F_{Y_1|W}(y | w) \in [\underline{F}_{Y_1|W}(y | w), \bar{F}_{Y_1|W}(y | w)]$ , as desired.

Next we establish the bounds for  $F_{Y_0|W}(y | w)$ . We also have that

$$\mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)}{1 - p_0(Y, w)} \mid X = 0, W = w \right] p_{0|w} = F_{Y_0|W}(y | w).$$

Furthermore, note that  $\mathbb{E}[(1 - X) | Y_0, W = w] = \mathbb{P}(X = 0 | Y_0, W = w) \in [1 - \bar{c}(w, \eta), 1 - \underline{c}(w, \eta)]$ . Changing  $X = 1$  to  $X = 0$ , and  $(\underline{c}(w, \eta), \bar{c}(w, \eta))$  to  $(1 - \bar{c}(w, \eta), 1 - \underline{c}(w, \eta))$  yields

$$\begin{aligned}
F_{Y_0|W}(y | w) &\leq F_{Y|X, W}(y | 0, w) \frac{p_{0|w}}{1 - \bar{c}(w, \eta)} \\
F_{Y_0|W}(y | w) &\leq \frac{p_{1|w} - \underline{c}(w, \eta)}{1 - \underline{c}(w, \eta)} + F_{Y|X, W}(y | 0, w) \frac{p_{0|w}}{1 - \underline{c}(w, \eta)} \\
F_{Y_0|W}(y | w) &\geq F_{Y|X, W}(y | 0, w) \frac{p_{0|w}}{1 - \underline{c}(w, \eta)} \\
F_{Y_0|W}(y | w) &\geq \frac{p_{1|w} - \bar{c}(w, \eta)}{1 - \bar{c}(w, \eta)} + F_{Y|X}(y | 0) \frac{p_{0|w}}{1 - \bar{c}(w, \eta)}.
\end{aligned}$$

almost surely. Therefore,  $F_{Y_0|W}(y | w) \in [\underline{F}_{Y_0|W}(y | w), \bar{F}_{Y_0|W}(y | w)]$ .  $\square$

### C.3.1 Proof of Theorem 4.4.1

We provide and show a number of preliminary lemmas that are used to prove Theorem 4.4.1. This first lemma establishes some properties of cdf bounds for  $Y_x$  given  $(X, W)$ .

**Lemma C.3.1** (Bounds of CDFs). *Let assumptions 11 and 12 hold. Then, for  $x \in \{0, 1\}$  and  $w \in \text{supp}(W)$ ,*

1. *The functions  $\underline{F}_{Y_x|X,w}(\cdot | 1 - x, w)$  and  $\overline{F}_{Y_x|X,W}(\cdot | 1 - x, w)$ , which are defined in Appendix C.1, are cdfs;*
2. *For all  $y \in \mathbb{R}$ ,*

$$\begin{aligned} \underline{F}_{Y_x|X,W}(y | 1 - x, w)p_{1-x|w} + F_{Y|X,W}(Y | X, w)p_{x|w} &= \underline{F}_{Y_x|W}(y | w) \\ \overline{F}_{Y_x|X,W}(y | 1 - x, w)p_{1-x|w} + F_{Y|X,W}(Y | X, w)p_{x|w} &= \overline{F}_{Y_x|W}(y | w). \end{aligned}$$

*Proof of Lemma C.3.1. Proof of Part 1:* We show that  $\underline{F}_{Y_1|X,W}(y | 0, w)$  is a cdf by showing it is nondecreasing, has limits  $(0, 1)$  when  $y$  approaches  $(-\infty, +\infty)$ , and is right-continuous. The same arguments can be used to deduce that  $\overline{F}_{Y_1|X,W}(y | 0, w)$ ,  $\underline{F}_{Y_0|X,W}(y | 1, w)$ , and  $\overline{F}_{Y_0|X,W}(y | 1, w)$  are also cdfs.

The function

$$\underline{F}_{Y_1|X,W}(y | 0, w) = \max \left\{ F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \bar{c})}{p_{0|w}\bar{c}}, \frac{\underline{c} - p_{1|w}}{\underline{c}p_{0|w}} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \underline{c})}{\underline{c}p_{0|w}} \right\}$$

is nondecreasing in  $y$  since each of its two arguments is nondecreasing in  $y$ , due to  $F_{Y|X,W}(\cdot | 1, w)$  being a cdf. Then note that

$$\lim_{y \rightarrow \infty} \underline{F}_{Y_1|X,W}(y | 0, w) = \max \left\{ \frac{p_{1|w}(1 - \bar{c})}{p_{0|w}\bar{c}}, 1 \right\} = \max \left\{ \frac{p_{1|w} - p_{1|w}\bar{c}}{\bar{c} - p_{1|w}\bar{c}}, 1 \right\} = 1,$$

where the last equality follows by  $p_{1|w} \leq \bar{c}$ . Also note that

$$\lim_{y \rightarrow -\infty} \underline{F}_{Y_1|X,W}(y | 0, w) = \max \left\{ 0, \frac{\underline{c} - p_{1|w}}{\underline{c}p_{0|w}} \right\} = 0,$$

where the last equality follows by  $\underline{c} \leq p_{1|w}$ . Finally, we can see that  $\underline{F}_{Y_1|X,W}(y | 0, w)$  is right-continuous with respect to  $y$  since  $F_{Y|X}(y | 1, w)$  is right-continuous and by the

continuity of affine transformations and of the maximum function. Therefore,  $\underline{F}_{Y_1|X,W}(y | 0, w)$  is a cdf.

**Proof of Part 2:** We show the first equality with  $x = 1$ , and the same arguments can be used to establish the equality for other cases. For  $y \in \mathbb{R}$ , the desired result follows by the following derivations:

$$\begin{aligned}
& \underline{F}_{Y_1|X,W}(y | 0, w)p_{0|w} + F_{Y|X,W}(y | 1, w)p_{1|w} \\
&= \max \left\{ F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \bar{c})}{p_{0|w}\bar{c}}, \frac{\underline{c} - p_{1|w}}{\underline{c}p_{0|w}} + F_{Y|X,W}(y | 1, w) \frac{p_{1|w}(1 - \underline{c})}{p_{0|w}\underline{c}} \right\} p_{0|w} \\
&\quad + F_{Y|X,W}(y | 1, w)p_{1|w} \\
&= \max \left\{ F_{Y|X,W}(y | 1, w)p_{1|w} \left( \frac{1 - \bar{c}}{\bar{c}} + 1 \right), \frac{\underline{c} - p_{1|w}}{\underline{c}} + F_{Y|X,W}(y | 1, w)p_{1|w} \left( \frac{1 - \underline{c}}{\underline{c}} + 1 \right) \right\} \\
&= \max \left\{ \frac{F_{Y|X,W}(y | 1, w)p_{1|w}}{\bar{c}}, \frac{\underline{c} - p_{1|w}}{\underline{c}} + \frac{F_{Y|X,W}(y | 1, w)p_{1|w}}{\underline{c}} \right\} \\
&= \underline{F}_{Y_1|W}(y | w).
\end{aligned}$$

Thus the proof is complete.  $\square$

**Lemma C.3.2.** *Let  $x \in \{0, 1\}$  and  $w \in \text{supp}(W)$ . Suppose  $m(\cdot)$  is a Borel measurable function and  $\mathbb{P}(m(Y_x) \geq \delta | W = w) = 1$  for some  $\delta > 0$ . The following statements are equivalent:*

1. *Conditional on  $W = w$ , the following statement holds almost surely:*

$$m(Y_x) = \mathbb{P}(X = x | Y_x, W = w). \quad (\text{C.3})$$

2. *For all  $y \in \mathbb{R}$ , the following equality holds:*

$$\mathbb{E} \left[ \frac{\mathbb{1}(Y_x \leq y) \mathbb{1}(X = x)}{m(Y_x)} | W = w \right] = \mathbb{P}(Y_x \leq y | W = w). \quad (\text{C.4})$$

*Proof of Lemma C.3.2.* We first prove that (C.3) implies (C.4). This follows from the law

of iterated expectations:

$$\begin{aligned}
\mathbb{E} \left[ \frac{\mathbb{1}(Y_x \leq y) \mathbb{1}(X = x)}{m(Y_x)} \mid W = w \right] &= \mathbb{E} \left[ \frac{\mathbb{1}(Y_x \leq y) \mathbb{E}[\mathbb{1}(X = x) \mid Y_x, W = w]}{m(Y_x)} \mid W = w \right] \\
&= \mathbb{E} \left[ \frac{\mathbb{1}(Y_x \leq y) \mathbb{P}(X = x \mid Y_x, W = w)}{m(Y_x)} \mid W = w \right] \\
&= \mathbb{E}[\mathbb{1}(Y_x \leq y) \mid W = w] \\
&= \mathbb{P}(Y_x \leq y \mid W = w),
\end{aligned}$$

where we use (C.3) and the assumption  $m(Y_x) \geq \delta > 0$  for the third equality.

Next, we prove that (C.4) implies (C.3). To show this result, we first establish a few key facts:

1. By the law of iterated expectations, (C.4) implies

$$\mathbb{E} \left[ \mathbb{1}(Y_x \leq y) \frac{\mathbb{P}(X = x \mid Y_x, W = w) - m(Y_x)}{m(Y_x)} \mid W = w \right] = 0$$

for all  $y \in \mathbb{R}^2$ .

2. For  $y \in \mathbb{R}$ , define the preimage from a half-space on  $\mathbb{R}^2$  as

$$I_y = \{\omega \in \Omega : Y_x(\omega) \leq y\},$$

where  $\Omega$  denotes  $Y_x$ 's sample space. Let  $\mathcal{A} = \{I_y : y \in \mathbb{R}\}$ . We then note that  $\mathcal{A}$  is closed under intersection since

$$I_y \cap I_{y'} = I_{\min\{y, y'\}} \in \mathcal{A} \quad \text{for any } y, y' \in \mathbb{R}.$$

This, combined with the non-emptiness of  $\mathcal{A}$ , implies that  $\mathcal{A}$  is a  $\pi$ -system.

3. The sample space can be written as a countable union of sets in  $\mathcal{A}$  since

$$\Omega = \{\omega \in \Omega : Y_x(\omega) < \infty\} = \bigcup_{n=1}^{\infty} I_n.$$

4. The random variable  $[\mathbb{P}(X = x \mid Y_x, W = w) - m(Y_x)]/m(Y_x)$  is measurable with respect to the  $\sigma$ -algebra generated by  $Y_x$  due to the Borel measurability of  $m(\cdot)$ , and

it is integrable since

$$\left| \frac{\mathbb{P}(X = x | Y_x, W = w) - m(Y_x)}{m(Y_x)} \right| \leq \frac{\mathbb{P}(X = x | Y_x, W = w)}{m(Y_x)} + 1 \leq \frac{1}{\delta} + 1 < +\infty,$$

where the first inequality follows by triangle inequality, and the second inequality follows by the assumption that  $m(Y_x) \geq \delta > 0$  almost surely.

Given the above facts, it follows by Billingsley (1995, Theorem 34.1) that

$$\frac{\mathbb{P}(X = x | Y_x, W = w) - m(Y_x)}{m(Y_x)} = 0 \quad \text{with probability one conditional on } W = w.$$

From this equality we conclude that  $\mathbb{P}(X = x | Y_x, W = w) = m(Y_x)$  with probability one conditional on  $W = w$  since  $m(Y_x) \geq \delta > 0$  almost surely. So the desired result has been established.  $\square$

The following lemma is a subset of Lemma 21.1 in A. W. van der Vaart, 2000, so we omit its proof.

**Lemma C.3.3** (Properties of CDFs and Quantiles). *Let  $p \in (0, 1)$  and  $x \in \mathbb{R}$ . Let  $F$  be a cdf and  $Q(p) = \inf\{z \in \mathbb{R} : F(z) \geq p\}$  be its quantile function. Then,*

1.  $Q(p) \leq x$  if and only if  $p \leq F(x)$ ;
2.  $F(Q(p)) \geq p$  where equality can fail only if  $F$  is discontinuous at  $Q(p)$ ;
3.  $F(Q(p)-) \leq p$ .

This next lemma is a compendium of properties of the cdf bounds. Its results are used throughout our proofs for the main theorems.

**Lemma C.3.4** (Preliminary Results). *Let  $w \in \text{supp}(W)$  and suppose assumptions 11 and 12 hold with  $\underline{c}(w, \eta) < p_{1|w} < \bar{c}(w, \eta)$ . Then, for  $x \in \{0, 1\}$ ,*

1.  $\underline{\tau}_x, \bar{\tau}_x \in (0, 1)$ ;
2.  $\bar{F}_{Y_x|W}(y | w)$  is continuous at  $y = \bar{Q}_x$  if and only if  $\mathbb{P}(Y = \bar{Q}_x | X = x, W = w) = 0$ , and  $\underline{F}_{Y_x|W}(y | w)$  is continuous at  $y = \underline{Q}_x$  if and only if  $\mathbb{P}(Y = \underline{Q}_x | X = x, W = w) = 0$ ;

3.  $\bar{A}_1, \underline{A}_1, 1 - \bar{A}_0, 1 - \underline{A}_0 \in [\underline{c}, \bar{c}]$ ;
4.  $\bar{p}_x(Y_x, w), \underline{p}_x(Y_x, w) \in [\underline{c}, \bar{c}]$  almost surely;
5. For all  $y \in \mathbb{R}$ ,

$$\begin{aligned} \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)X}{\underline{p}_1(Y, w)} \middle| W = w \right] &= \underline{F}_{Y_1|W}(y | w) \\ \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)X}{\bar{p}_1(Y, w)} \middle| W = w \right] &= \bar{F}_{Y_1|W}(y | w), \\ \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)(1 - X)}{1 - \underline{p}_0(Y, w)} \middle| W = w \right] &= \underline{F}_{Y_0|W}(y | w) \\ \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)(1 - X)}{1 - \bar{p}_0(Y, w)} \middle| W = w \right] &= \bar{F}_{Y_0|W}(y | w). \end{aligned}$$

6. For all  $(y_1, y_0) \in \mathbb{R}^2$ , the following inequalities are equivalent:

- (a)  $\bar{F}_{Y_x|W}(y_x|w) \leq \underline{F}_{Y_{1-x}|W}(y_{1-x}|w)$ ;
- (b)  $F_{Y|X,W}(y_x|x, w) \leq \underline{F}_{Y_{1-x}|X,W}(y_{1-x}|x, w)$ ;
- (c)  $\bar{F}_{Y_x|X,W}(y_x | 1 - x, w) \leq F_{Y|X,W}(y_{1-x} | 1 - x, w)$ .

Also, the following inequalities are equivalent:

- (d)  $\bar{F}_{Y_x|W}(y_x|w) \geq \underline{F}_{Y_{1-x}|W}(y_{1-x}|w)$ ;
- (e)  $F_{Y|X,W}(y_x|x, w) \geq \underline{F}_{Y_{1-x}|X,W}(y_{1-x}|x, w)$ ;
- (f)  $\bar{F}_{Y_x|X,W}(y_x | 1 - x, w) \geq F_{Y|X,W}(y_{1-x} | 1 - x, w)$ .

7. The following inequalities are equivalent:

- (a)  $F_{Y|X,W}(\bar{Q}_1 - | 1, w) \leq \underline{F}_{Y_0|X,W}(\underline{Q}_0 - | 1, w)$ ;
- (b)  $\bar{F}_{Y_1|X,W}(\bar{Q}_1 - | 0, w) \leq F_{Y|X,W}(\underline{Q}_0 - | 0, w)$ ;
- (c)  $\bar{F}_{Y_1|W}(\bar{Q}_1 - | w) \leq \underline{F}_{Y_0|W}(\underline{Q}_0 - | w)$ .

Also, the following inequalities are equivalent:

- (d)  $F_{Y|X,W}(\underline{Q}_1 - | 1, w) \leq \bar{F}_{Y_0|X,W}(\bar{Q}_0 - | 1, w)$ ;
- (e)  $\underline{F}_{Y_1|X,W}(\underline{Q}_1 - | 0, w) \leq F_{Y|X,W}(\bar{Q}_0 - | 0, w)$ ;
- (f)  $\underline{F}_{Y_1|W}(\underline{Q}_1 - | w) \leq \bar{F}_{Y_0|W}(\bar{Q}_0 - | w)$ .

8. The following inequalities hold:

$$\begin{aligned} & \max \left\{ \overline{F}_{Y_1|W}(\overline{Q}_1 - |w), \underline{F}_{Y_0|W}(\underline{Q}_0 - |w) \right\} \\ & \leq \frac{\overline{c} - p_{1|w}}{\overline{c} - \underline{c}} \leq \min \left\{ \overline{F}_{Y_1|W}(\overline{Q}_1 | w), \underline{F}_{Y_0|W}(\underline{Q}_0 | w) \right\} \end{aligned}$$

and

$$\begin{aligned} & \max \left\{ \overline{F}_{Y_0|W}(\overline{Q}_0 - |w), \underline{F}_{Y_1|W}(\underline{Q}_1 - |w) \right\} \\ & \leq \frac{p_{1|w} - \underline{c}}{\overline{c} - \underline{c}} \leq \min \left\{ \overline{F}_{Y_0|W}(\overline{Q}_0 | w), \underline{F}_{Y_1|W}(\underline{Q}_1 | w) \right\}. \end{aligned}$$

9. The following inequalities hold:

$$\begin{aligned} & \max \left\{ F_{Y|X,W}(\overline{Q}_1 - |1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 - |1, w) \right\} \\ & \leq \overline{\tau}_1 \leq \min \left\{ F_{Y|X,W}(\overline{Q}_1 | 1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 | 1, w) \right\} \end{aligned}$$

and

$$\begin{aligned} & \max \left\{ \overline{F}_{Y_0|X,W}(\overline{Q}_0 - |1, w), F_{Y|X,W}(\underline{Q}_1 - |1, w) \right\} \\ & \leq \underline{\tau}_1 \leq \min \left\{ \overline{F}_{Y_0|X,W}(\overline{Q}_0 | 1, w), F_{Y|X,W}(\underline{Q}_1 | 1, w) \right\}. \end{aligned}$$

10. For all  $(y_1, y_0) \in \mathbb{R}^2$ , the following inequalities are equivalent:

- (a)  $\overline{F}_{Y_x|W}(y_x|w) + \overline{F}_{Y_{1-x}|W}(y_{1-x}|w) \geq 1$ ;
- (b)  $F_{Y|X,W}(y_x|x, w) + \overline{F}_{Y_{1-x}|X,W}(y_{1-x}|x, w) \geq 1$ ;
- (c)  $\overline{F}_{Y_x|X,W}(y_x | 1 - x, w) + F_{Y|X,W}(y_{1-x} | 1 - x, w) \geq 1$ .

Also, the following inequalities are equivalent:

- (d)  $\underline{F}_{Y_x|W}(y_x|w) + \underline{F}_{Y_{1-x}|W}(y_{1-x}|w) \geq 1$ ;
- (e)  $F_{Y|X,W}(y_x|x, w) + \underline{F}_{Y_{1-x}|X,W}(y_{1-x}|x, w) \geq 1$ ;
- (f)  $\underline{F}_{Y_x|X,W}(y_x | 1 - x, w) + F_{Y|X,W}(y_{1-x} | 1 - x, w) \geq 1$ .

11. The following inequalities are equivalent:

- (a)  $\overline{F}_{Y_1|W}(\overline{Q}_1 | w) + \overline{F}_{Y_0|W}(\overline{Q}_0 - |w) \geq 1$ ;
- (b)  $F_{Y|X,W}(\overline{Q}_1 | 1, w) + \overline{F}_{Y_0|X,W}(\overline{Q}_0 - |1, w) \geq 1$ ;

$$(c) \bar{F}_{Y_1|X,W}(\bar{Q}_1|0, w) + F_{Y|X,W}(\bar{Q}_0 - |0, w) \geq 1.$$

Also, the following inequalities are equivalent:

$$(d) \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) + \bar{F}_{Y_0|W}(\bar{Q}_0|w) \geq 1;$$

$$(e) F_{Y|X,W}(\bar{Q}_1 - | 1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 | 1, w) \geq 1;$$

$$(f) \bar{F}_{Y_1|X,W}(\bar{Q}_1 - |0, w) + F_{Y|X,W}(\bar{Q}_0|0, w) \geq 1.$$

*Proof of Lemma C.3.4.* For brevity, we omit covariates  $w \in \text{supp}(W)$  and drop notation referring on conditional probability  $\cdot | w$  and  $\cdot | W = w$  from this proof. However, note that our arguments hold when conditioning on  $W = w$  throughout.

**Proof of Part 1:** First consider  $\underline{\tau}_1$ :

$$\underline{\tau}_1 = \frac{p_1 - \underline{c}}{\bar{c} - \underline{c}} \frac{\bar{c}}{p_1} = \frac{p_1 \bar{c} - \underline{c} \bar{c}}{p_1 \bar{c} - \underline{c} p_1} < \frac{p_1 \bar{c} - \underline{c} \bar{c}}{p_1 \bar{c} - \underline{c} \bar{c}} = 1,$$

where the inequality is strict because  $p_1 < \bar{c}$  and  $\underline{c} > 0$ . Similarly,

$$\underline{\tau}_1 = \frac{p_1 - \underline{c}}{\bar{c} - \underline{c}} \frac{\bar{c}}{p_1} > \frac{\underline{c} - \underline{c}}{\bar{c} - \underline{c}} \frac{\bar{c}}{p_1} = 0$$

where the inequality is strict because  $p_1 > \underline{c}$  and  $\bar{c} > 0$ . Thus,  $\underline{\tau}_1 \in (0, 1)$ . Since  $\bar{\tau}_1 = 1 - \underline{\tau}_1$ ,  $\bar{\tau}_1 \in (0, 1)$  as well. The proofs for  $\bar{\tau}_0$  and  $\underline{\tau}_0$  are similar.

**Proof of Part 2:** First consider the statement involving  $\bar{Q}_x$  with  $x = 1$ . We show the following inequality

$$\frac{p_1}{\bar{c}} \mathbb{P}(Y = \bar{Q}_1 | X = 1) \leq \bar{F}_{Y_1}(\bar{Q}_1) - \bar{F}_{Y_1}(\bar{Q}_1 -) \leq \frac{p_1}{\underline{c}} \mathbb{P}(Y = \bar{Q}_1 | X = 1). \quad (\text{C.5})$$

From this inequality and Assumption 12 that  $0 < \underline{c} \leq p_1 \leq \bar{c}$ , we will conclude that  $\bar{F}_{Y_1}(y)$  is continuous at  $y = \bar{Q}_1$  if and only if  $\mathbb{P}(Y = \bar{Q}_1 | X = 1) = 0$ .

To show the lower bound inequality in (C.5), note that

$$\begin{aligned}
\bar{F}_{Y_1}(\bar{Q}_1) - \bar{F}_{Y_1}(\bar{Q}_1-) &= \frac{\bar{c} - p_1}{\bar{c}} + F_{Y|X}(\bar{Q}_1 | 1) \frac{p_1}{\bar{c}} - \frac{p_1}{\underline{c}} F_{Y|X}(\bar{Q}_1- | 1) \\
&= \frac{\bar{c} - p_1}{\bar{c}} - p_1 F_{Y|X}(Q_{Y|X}(\bar{\tau}_1 | 1) - | 1) \frac{\bar{c} - \underline{c}}{\underline{c}\bar{c}} + \frac{p_1}{\bar{c}} \mathbb{P}(Y = \bar{Q}_1 | X = 1) \\
&\geq \frac{\bar{c} - p_1}{\bar{c}} - p_1 \bar{\tau}_1 \frac{\bar{c} - \underline{c}}{\underline{c}\bar{c}} + \frac{p_1}{\bar{c}} \mathbb{P}(Y = \bar{Q}_1 | X = 1) \\
&= \frac{p_1}{\bar{c}} \mathbb{P}(Y = \bar{Q}_1 | X = 1),
\end{aligned}$$

The first line holds by the definition of  $\bar{F}_{Y_1}$ . The third line holds by Lemma C.3.3.3. The last line holds by the definition of  $\bar{\tau}_1$ .

Likewise, we also have the following derivation:

$$\begin{aligned}
\bar{F}_{Y_1}(\bar{Q}_1) - \bar{F}_{Y_1}(\bar{Q}_1-) &= \frac{\bar{c} - p_1}{\bar{c}} + F_{Y|X}(\bar{Q}_1 | 1) \frac{p_1}{\bar{c}} - \frac{p_1}{\underline{c}} F_{Y|X}(\bar{Q}_1- | 1) \\
&= \frac{\bar{c} - p_1}{\bar{c}} - p_1 F_{Y|X}(Q_{Y|X}(\bar{\tau}_1 | 1) | 1) \frac{\bar{c} - \underline{c}}{\underline{c}\bar{c}} + \frac{p_1}{\bar{c}} \mathbb{P}(Y = \bar{Q}_1 | X = 1) \\
&\leq \frac{\bar{c} - p_1}{\bar{c}} - p_1 \bar{\tau}_1 \frac{\bar{c} - \underline{c}}{\underline{c}\bar{c}} + \frac{p_1}{\bar{c}} \mathbb{P}(Y = \bar{Q}_1 | X = 1) \\
&= \frac{p_1}{\bar{c}} \mathbb{P}(Y = \bar{Q}_1 | X = 1),
\end{aligned}$$

where we use Lemma C.3.3.2 in the third line. This establishes the upper bound inequality in (C.5). So the desired result follows. The proofs for the statements involving  $\underline{Q}_1$ ,  $\bar{Q}_0$ , and  $\underline{Q}_0$  are similar by establishing the following bounds:

$$\begin{aligned}
\underline{F}_{Y_1}(\underline{Q}_1) - \underline{F}_{Y_1}(\underline{Q}_1-) &\in \left[ \frac{p_1}{\bar{c}} \mathbb{P}(Y = \underline{Q}_1 | X = 1), \quad \frac{p_1}{\underline{c}} \mathbb{P}(Y = \underline{Q}_1 | X = 1) \right] \\
\bar{F}_{Y_0}(\bar{Q}_0) - \bar{F}_{Y_0}(\bar{Q}_0-) &\in \left[ \frac{p_0}{1 - \underline{c}} \mathbb{P}(Y = \bar{Q}_0 | X = 0), \quad \frac{p_0}{1 - \bar{c}} \mathbb{P}(Y = \bar{Q}_0 | X = 0) \right] \quad (\text{C.6}) \\
\underline{F}_{Y_0}(\underline{Q}_0) - \underline{F}_{Y_0}(\underline{Q}_0-) &\in \left[ \frac{p_0}{1 - \underline{c}} \mathbb{P}(Y = \underline{Q}_0 | X = 0), \quad \frac{p_0}{1 - \bar{c}} \mathbb{P}(Y = \underline{Q}_0 | X = 0) \right],
\end{aligned}$$

which can be derived by similar steps to those above.

**Proof of Part 3:** First consider  $\bar{A}_1$  if the denominator is positive, as  $\bar{A}_1 = p_1$  is trivially bounded in  $[\underline{c}, \bar{c}]$  by Assumption 12 if the denominator becomes zero. By its definition, we have

$$\bar{A}_1 = \frac{p_1 \mathbb{P}(Y = \bar{Q}_1 | X = 1)}{\left(\frac{\bar{c} - p_1}{\bar{c}} + F_{Y|X}(\bar{Q}_1 | 1) \frac{p_1}{\bar{c}}\right) - \frac{p_1}{\bar{c}} \mathbb{P}(Y < \bar{Q}_1 | X = 1)} = \frac{p_1 \mathbb{P}(Y = \bar{Q}_1 | X = 1)}{\bar{F}_{Y_1}(\bar{Q}_1) - \bar{F}_{Y_1}(\bar{Q}_1 -)}.$$

From the inequality in equation (C.5), we deduce that  $\mathbb{P}(Y = \bar{Q}_1 | X = 1) > 0$ , and

$$\frac{p_1 \mathbb{P}(Y = \bar{Q}_1 | X = 1)}{\bar{F}_{Y_1}(\bar{Q}_1) - \bar{F}_{Y_1}(\bar{Q}_1 -)} \in [\underline{c}, \bar{c}].$$

So this concludes that  $\bar{A}_1 \in [\underline{c}, \bar{c}]$ . Similarly, the results for  $\underline{A}_1$ ,  $1 - \bar{A}_0$ , and  $1 - \underline{A}_0$  can be deduced by inequalities (C.6).

**Proof of Part 4:** These propensity scores can only take the values  $\underline{c}, \bar{c}, \bar{A}_1, \underline{A}_1, 1 - \bar{A}_0$ , and  $1 - \underline{A}_0$ . By Part 3, these values all lie in  $[\underline{c}, \bar{c}]$ .

**Proof of Part 5:** First we show that  $\mathbb{E} \left[ \mathbb{1}(Y \leq y) X / \underline{p}_1(Y) \right] = \underline{F}_{Y_1}(y)$  for all  $y \in \mathbb{R}$ . The proof for  $\bar{F}_{Y_1}$  is similar by interchanging  $\underline{c}$  with  $\bar{c}$  and thus omitted.

To prove the desired identity, we split the analysis in three cases depending on the value of  $y \in \mathbb{R}$ . For  $y < \underline{Q}_1$ , we have  $\mathbb{1}(Y \leq y) / \underline{p}_1(Y) = \mathbb{1}(Y \leq y) / \bar{c}$  and thus

$$\mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y) X}{\underline{p}_1(Y)} \right] = \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y) X}{\bar{c}} \right] = \frac{F_{Y|X}(y | 1) p_1}{\bar{c}} = \underline{F}_{Y_1}(y).$$

When  $y = \underline{Q}_1$ , we can write

$$\begin{aligned} \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq \underline{Q}_1) X}{\underline{p}_1(Y)} \right] &= \mathbb{E} \left[ \frac{\mathbb{1}(Y < \underline{Q}_1) X}{\bar{c}} \right] + \mathbb{E} \left[ \frac{\mathbb{1}(Y = \underline{Q}_1) X}{\underline{A}_1} \right] \\ &= \frac{\mathbb{P}(Y < \underline{Q}_1 | X = 1) p_1}{\bar{c}} + \mathbb{P}(Y = \underline{Q}_1 | X = 1) p_1 \left( \frac{\mathbb{P}(Y = \underline{Q}_1 | X = 1) p_1}{\bar{F}_{Y_1}(\underline{Q}_1) - \bar{F}_{Y_1}(\underline{Q}_1 -)} \right)^{-1} \\ &= \frac{\mathbb{P}(Y < \underline{Q}_1 | X = 1) p_1}{\bar{c}} + \bar{F}_{Y_1}(\underline{Q}_1) - \bar{F}_{Y_1}(\underline{Q}_1 -) \\ &= \bar{F}_{Y_1}(\underline{Q}_1 -) + \bar{F}_{Y_1}(\underline{Q}_1) - \bar{F}_{Y_1}(\underline{Q}_1 -) \\ &= \bar{F}_{Y_1}(\underline{Q}_1). \end{aligned}$$

Finally, when  $y > \underline{Q}_1$ , we can write

$$\begin{aligned}
\mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)X}{\underline{p}_1(Y)} \right] &= \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq \underline{Q}_1)X}{\underline{p}_1(Y)} \right] + \mathbb{E} \left[ \frac{\mathbb{1}(\underline{Q}_1 < Y \leq y)X}{\underline{c}} \right] \\
&= \underline{F}_{Y_1}(\underline{Q}_1) + \frac{\left( F_{Y|X}(y | 1) - F_{Y|X}(\underline{Q}_1 | 1) \right) p_1}{\underline{c}} \\
&= \frac{\underline{c} - p_1}{\underline{c}} + F_{Y|X}(\underline{Q}_1 | 1) \frac{p_1}{\underline{c}} + \frac{\left( F_{Y|X}(y | 1) - F_{Y|X}(\underline{Q}_1 | 1) \right) p_1}{\underline{c}} \\
&= \frac{\underline{c} - p_1}{\underline{c}} + F_{Y|X}(y | 1) \frac{p_1}{\underline{c}} \\
&= \underline{F}_{Y_1}(y).
\end{aligned}$$

Thus the desired identity holds for all  $y \in \mathbb{R}$ .

Next we show the identity  $\mathbb{E} \left[ \mathbb{1}(Y \leq y)(1 - X)/(1 - \underline{p}_0(Y)) \right] = \underline{F}_{Y_0}(y)$  for all  $y \in \mathbb{R}$ .

The proof for  $\overline{F}_{Y_0}$  is similar by interchanging  $\underline{c}$  with  $\bar{c}$  and thus omitted. Similar to above, we split the analysis in three cases depending on the value of  $y \in \mathbb{R}$ . For  $y < \underline{Q}_0$ , we have  $\mathbb{1}(Y \leq y)/(1 - \underline{p}_0(Y)) = \mathbb{1}(Y \leq y)/(1 - \underline{c})$  and thus

$$\mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)(1 - X)}{1 - \underline{p}_0(Y)} \right] = \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)(1 - X)}{1 - \underline{c}} \right] = \frac{F_{Y|X}(y | 0)p_0}{1 - \underline{c}} = \underline{F}_{Y_0}(y).$$

When  $y = \underline{Q}_0$ , we can write

$$\begin{aligned}
\mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)(1 - X)}{1 - \underline{p}_0(Y)} \right] &= \mathbb{E} \left[ \frac{\mathbb{1}(Y < \underline{Q}_0)(1 - X)}{1 - \underline{c}} \right] + \mathbb{E} \left[ \frac{\mathbb{1}(Y = \underline{Q}_0)(1 - X)}{\underline{A}_0} \right] \\
&= \frac{\mathbb{P}(Y < \underline{Q}_0 | X = 0)p_0}{1 - \underline{c}} + \mathbb{P}(Y = \underline{Q}_0 | X = 0)p_0 \left( \frac{\mathbb{P}(Y = \underline{Q}_0 | X = 0)p_0}{\underline{F}_{Y_0}(\underline{Q}_0) - \underline{F}_{Y_0}(\underline{Q}_0 -)} \right)^{-1} \\
&= \frac{\mathbb{P}(Y < \underline{Q}_0 | X = 0)p_0}{1 - \underline{c}} + \underline{F}_{Y_0}(\underline{Q}_0) - \underline{F}_{Y_0}(\underline{Q}_0 -) \\
&= \underline{F}_{Y_0}(\underline{Q}_0 -) + \underline{F}_{Y_0}(\underline{Q}_0) - \underline{F}_{Y_0}(\underline{Q}_0 -) \\
&= \underline{F}_{Y_0}(\underline{Q}_0).
\end{aligned}$$

When  $y > \underline{Q}_0$ , we can write

$$\begin{aligned}
\mathbb{E} \left[ \frac{1(Y \leq y)(1-X)}{1-p_0(Y)} \right] &= \mathbb{E} \left[ \frac{1(Y \leq \underline{Q}_0)(1-X)}{1-\underline{p}_0(Y)} \right] + \mathbb{E} \left[ \frac{\mathbb{1}(\underline{Q}_0 < Y \leq y)(1-X)}{1-\bar{c}} \right] \\
&= \underline{F}_{Y_0}(\underline{Q}_0) + \frac{(F_{Y|X}(y|0) - F_{Y|X}(\underline{Q}_0|0)) p_0}{1-\bar{c}} \\
&= \frac{p_1 - \bar{c}}{1-\bar{c}} + F_{Y|X}(\underline{Q}_0|0) \frac{p_0}{1-\bar{c}} + \frac{(F_{Y|X}(y|0) - F_{Y|X}(\underline{Q}_0|0)) p_0}{1-\bar{c}} \\
&= \frac{p_1 - \bar{c}}{1-\bar{c}} + F_{Y|X}(y|0) \frac{p_0}{1-\bar{c}} \\
&= \underline{F}_{Y_0}(y).
\end{aligned}$$

Thus the desired identity has been established for all  $y \in \mathbb{R}$ .

**Proof of Part 6:** We begin by considering the first sequence of equivalences between (a), (b), and (c) for  $x = 1$ . By Lemma C.3.3.1.

$$\begin{aligned}
\bar{F}_{Y_1}(y_1) \leq \underline{E}_{Y_0}(y_0) &\iff \bar{R}_1(y_1) := \bar{Q}_{Y_0}(\bar{F}_{Y_1}(y_1)) \leq y_0 \\
F_{Y|X}(y_1|1) \leq \underline{F}_{Y_0|X}(y_0|1) &\iff \bar{R}_2(y_1) := \bar{Q}_{Y_0|X}(F_{Y|X}(y_1|1)|1) \leq y_0 \\
\bar{F}_{Y_1|X}(y_1|0) \leq F_{Y_0|X}(y_0|0) &\iff \bar{R}_3(y_1) := Q_{Y|X}(\bar{F}_{Y_1|X}(y_1|0)|0) \leq y_0.
\end{aligned}$$

The equivalence relationship for the statements on the left hand side holds if  $\bar{R}_1(y_1) = \bar{R}_2(y_1) = \bar{R}_3(y_1)$  for all  $y_1 \in \mathbb{R}$ . By direct calculation, we can see that

$$\begin{aligned}
\bar{R}_2(y_1) &= \bar{R}_3(y_1) \\
&= Q_{Y|X} \left( \min \left\{ \frac{(1-\underline{c})p_1}{p_0\underline{c}} F_{Y|X}(y_1|1), \frac{p_0 - (1-\bar{c})}{p_0\bar{c}} + \frac{p_1(1-\bar{c})}{p_0\bar{c}} F_{Y|X}(y_1|1) \right\} \mid 0 \right)
\end{aligned}$$

and thus it remains to show that  $\bar{R}_1(y_1) = \bar{R}_2(y_1)$ . Recall that

$$\begin{aligned}
\bar{R}_1(y_1) &= \bar{Q}_{Y_0}(\bar{F}_{Y_1}(y_1)) \\
&= Q_{Y|X} \left( \min \left\{ \frac{1-\underline{c}}{p_0} \bar{F}_{Y_1}(y_1), \frac{p_0 - (1-\bar{c})}{p_0} + \frac{1-\bar{c}}{p_0} \bar{F}_{Y_1}(y_1) \right\} \mid 0 \right).
\end{aligned}$$

We split the proof into two cases. First consider  $y_1 < \bar{Q}_1 = Q_{Y|X}(\bar{\tau}_1 | 1)$ . In such case we have

$$F_{Y|X}(y_1 | 1) < \bar{\tau}_1 = \frac{(\bar{c} - p_1)c}{(\bar{c} - c)p_1} \quad (\text{C.7})$$

by Lemma C.3.3.1, and

$$\bar{F}_{Y_1}(y_1) = F_{Y|X}(y | 1) \frac{p_1}{c}. \quad (\text{C.8})$$

Using equations (C.7) and (C.8), it can be verified that

$$\frac{1 - c}{p_0} \bar{F}_{Y_1}(y_1) < \frac{p_0 - (1 - \bar{c})}{p_0} + \frac{1 - \bar{c}}{p_0} \bar{F}_{Y_1}(y_1).$$

This implies

$$\bar{R}_1(y_1) = Q_{Y|X} \left( \frac{1 - c}{p_0} \bar{F}_{Y_1}(y_1) | 0 \right) = Q_{Y|X} \left( \frac{(1 - c)p_1}{p_0 c} F_{Y|X}(y_1 | 1) | 0 \right) \quad \text{if } y_1 < \bar{Q}_1. \quad (\text{C.9})$$

Next consider  $y \geq \bar{Q}_1$ . Then we have  $F_{Y|X}(y_1 | 1) \geq \bar{\tau}_1$  by Lemma C.3.3.1, and

$$\bar{F}_{Y_1}(y_1) = \frac{\bar{c} - p_1}{\bar{c}} + F_{Y|X}(y | 1) \frac{p_1}{\bar{c}}.$$

These two implications lead to the following inequality

$$\frac{1 - c}{p_0} \bar{F}_{Y_1}(y_1) \geq \frac{p_0 - (1 - \bar{c})}{p_0} + \frac{1 - \bar{c}}{p_0} \bar{F}_{Y_1}(y_1),$$

which further implies

$$\begin{aligned} \bar{R}_1(y_1) &= Q_{Y|X} \left( \frac{p_0 - (1 - \bar{c})}{p_0} + \frac{1 - \bar{c}}{p_0} \bar{F}_{Y_1}(y_1) | 0 \right) \\ &= Q_{Y|X} \left( \frac{p_0 - (1 - \bar{c})}{p_0 \bar{c}} + \frac{p_1(1 - \bar{c})}{p_0 \bar{c}} F_{Y|X}(y_1 | 1) | 0 \right) \quad \text{if } y_1 \geq \bar{Q}_1. \end{aligned} \quad (\text{C.10})$$

By Lemma C.3.3.1,  $y_1 < \bar{Q}_1$  is equivalent to  $F_{Y|X}(y_1 | 1) < \bar{\tau}_1$ , and it is further equivalent to

$$\frac{(1 - c)p_1}{p_0 c} F_{Y|X}(y_1 | 1) < \frac{p_0 - (1 - \bar{c})}{p_0 \bar{c}} + \frac{p_1(1 - \bar{c})}{p_0 \bar{c}} F_{Y|X}(y_1 | 1).$$

From this, (C.9), and (C.10), we deduce that

$$\begin{aligned}
\bar{R}_2(y_1) &= Q_{Y|X} \left( \min \left\{ \frac{(1-\underline{c})p_1}{p_0\underline{c}} F_{Y|X}(y_1 | 1), \frac{p_0 - (1-\bar{c})}{p_0\bar{c}} + \frac{p_1(1-\bar{c})}{p_0\bar{c}} F_{Y|X}(y_1 | 1) \right\} | 0 \right) \\
&= Q_{Y|X} \left( \frac{(1-\underline{c})p_1}{p_0\underline{c}} F_{Y|X}(y_1 | 1) | 0 \right) \mathbb{1}(y < \bar{Q}_1) \\
&\quad + Q_{Y|X} \left( \frac{\bar{c} - p_1}{p_0\bar{c}} + \frac{p_1(1-\bar{c})}{p_0\bar{c}} F_{Y|X}(y_1 | 1) | 0 \right) \mathbb{1}(y \geq \bar{Q}_1) \\
&= \bar{R}_1(y_1).
\end{aligned}$$

Therefore, the desired conclusion for  $x = 1$  has been established. Similar arguments can be used to show that the same conclusion also holds for  $x = 0$ , and to show the second set of equivalences between (d), (e), and (f).

**Proof of Part 7:** We first consider the equivalence of the statement (a), (b), and (c). We can write

$$\Delta_1 := F_{Y|X}(\bar{Q}_1 - | 1) - \underline{F}_{Y_0|X}(\underline{Q}_0 - | 1) = F_{Y|X}(\bar{Q}_1 - | 1) - \frac{p_0\underline{c}F_{Y|X}(\underline{Q}_0 - | 0)}{p_1(1-\underline{c})}$$

From this, we note that

$$\Delta_2 := \bar{F}_{Y_1}(\bar{Q}_1 -) - \underline{F}_{Y_0}(\underline{Q}_0 -) = \frac{p_1 F_{Y|X}(\bar{Q}_1 - | 1)}{\underline{c}} - \frac{p_0 F_{Y|X}(\underline{Q}_0 - | 0)}{1-\underline{c}} = \frac{p_1}{\underline{c}} \Delta_1,$$

and

$$\begin{aligned}
\Delta_3 &:= \bar{F}_{Y_1|X}(\bar{Q}_1 - | 0) - F_{Y|X}(\underline{Q}_0 - | 0) \\
&= \frac{p_1(1-\underline{c})}{p_0\underline{c}} F_{Y|X}(\bar{Q}_1 - | 1) - F_{Y|X}(\underline{Q}_0 - | 0) \\
&= \frac{p_1(1-\underline{c})}{p_0\underline{c}} \Delta_1.
\end{aligned}$$

The desired result follows by noting that  $\Delta_1$ ,  $\Delta_2$ , and  $\Delta_3$  all have the same sign. The proof of the equivalence of statements (d), (e), and (f) is similar and thus omitted.

**Proof of Part 8:** We have that

$$\bar{F}_{Y_1}(\bar{Q}_1 -) = \frac{F_{Y|X}(Q_{Y|X}(\bar{\tau}_1 | 1) - | 1)p_1}{\underline{c}} \leq \bar{\tau}_1 \frac{p_1}{\underline{c}} = \frac{\bar{c} - p_1}{\bar{c} - \underline{c}}$$

by Lemma C.3.3.3. Similarly,

$$\bar{F}_{Y_1}(\bar{Q}_1) = \frac{\bar{c} - p_1}{\bar{c}} + \frac{F_{Y|X}(Q_{Y|X}(\bar{\tau}_1 | 1) | 1)p_1}{\bar{c}} \geq \frac{\bar{c} - p_1}{\bar{c}} + \bar{\tau}_1 \frac{p_1}{\bar{c}} = \frac{\bar{c} - p_1}{\bar{c} - \underline{c}}.$$

by Lemma C.3.3.2. The other inequalities can be shown in a similar manner. Their derivations are thus omitted.

**Proof of Part 9:** We have that

$$F_{Y|X}(\bar{Q}_1 - | 1) = F_{Y|X}(Q_{Y|X}(\bar{\tau}_1 | 1) - | 1) \leq \bar{\tau}_1$$

by Lemma C.3.3.3. Similarly,

$$F_{Y|X}(\bar{Q}_1 | 1) = F_{Y|X}(Q_{Y|X}(\bar{\tau}_1 | 1) | 1) \geq \bar{\tau}_1$$

by Lemma C.3.3.2. The same arguments can be applied to  $F_{Y_0|X}(Q_0 - | 1)$  and  $F_{Y_0|X}(Q_0 | 1)$ .

So we have

$$F_{Y_0|X}(Q_0 - | 1) = F_{Y|X}(Q_{Y|X}(\tau_0 | 0) - | 0) \frac{p_0 \underline{c}}{p_1(1 - \underline{c})} \leq \frac{\tau_0 p_0 \underline{c}}{p_1(1 - \underline{c})} = \frac{\underline{c}(\bar{c} - p_1)}{p_1(\bar{c} - \underline{c})} = \bar{\tau}_1.$$

via Lemma C.3.3.3, and

$$F_{Y_0|X}(Q_0 | 1) = \frac{p_1 - \bar{c}}{(1 - \bar{c})p_1} + F_{Y|X}(Q_{Y|X}(\tau_0 | 0) | 0) \frac{p_0 \bar{c}}{p_1(1 - \bar{c})} \geq \frac{p_1 - \bar{c}}{(1 - \bar{c})p_1} + \frac{\tau_0 p_0 \bar{c}}{p_1(1 - \bar{c})} = \bar{\tau}_1.$$

via Lemma C.3.3.2. The proofs for the other inequalities are similar and thus omitted.

**Proof of Part 10:** We begin by considering the first set of equivalences between (a), (b), and (c) when  $x = 1$ , and the equivalences for  $x = 0$  are identical. By Lemma C.3.3.1, we have the following equivalence relationships:

$$\bar{F}_{Y_1}(y_1) + \bar{F}_{Y_0}(y_0) \geq 1 \iff y_1 \geq \underline{R}_1(y_0) := \underline{Q}_{Y_1}(1 - \bar{F}_{Y_0}(y_0))$$

$$F_{Y|X}(y_1 | 1) + \bar{F}_{Y_0|X}(y_0 | 1) \geq 1 \iff y_1 \geq \underline{R}_2(y_0) := Q_{Y|X}(1 - \bar{F}_{Y_0|X}(y_0 | 1) | 1)$$

$$\bar{F}_{Y_1|X}(y_1 | 0) + F_{Y|X}(y_0 | 0) \geq 1 \iff y_1 \geq \underline{R}_3(y_0) := \underline{Q}_{Y_1|X}(1 - F_{Y|X}(y_0 | 0) | 1).$$

To prove the equivalence of statements on the left, it suffices to show that  $\underline{R}_1(y_0) = \underline{R}_2(y_0) = \underline{R}_3(y_0)$  for all  $y_0 \in \mathbb{R}$ . First, we directly compute  $\underline{R}_2(y_0)$  and  $\underline{R}_3(y_0)$ :

$$\begin{aligned}
\underline{R}_2(y_0) &= Q_{Y|X}(1 - \overline{F}_{Y_0|X}(y_0 | 1) | 1) \\
&= Q_{Y|X} \left( 1 - \min \left\{ \frac{p_0 \bar{c} F_{Y|X}(y_0|0)}{p_1(1 - \bar{c})}, \frac{p_1 - \underline{c} + p_0 \underline{c} F_{Y|X}(y_0|0)}{p_1(1 - \underline{c})} \right\} | 1 \right) \\
&= Q_{Y|X} \left( \max \left\{ 1 - \frac{p_0 \bar{c} F_{Y|X}(y_0|0)}{p_1(1 - \bar{c})}, 1 - \frac{p_1 - \underline{c} + p_0 \underline{c} F_{Y|X}(y_0|0)}{p_1(1 - \underline{c})} \right\} | 1 \right) \\
&= Q_{Y|X} \left( \max \left\{ 1 - \frac{p_0 \bar{c} F_{Y|X}(y_0|0)}{p_1(1 - \bar{c})}, \frac{p_1(1 - \underline{c}) - p_1 + \underline{c} - p_0 \underline{c} F_{Y|X}(y_0|0)}{p_1(1 - \underline{c})} \right\} | 1 \right) \\
&= Q_{Y|X} \left( \max \left\{ 1 - \frac{p_0 \bar{c} F_{Y|X}(y_0|0)}{p_1(1 - \bar{c})}, \frac{\underline{c} p_0(1 - F_{Y|X}(y_0|0))}{p_1(1 - \underline{c})} \right\} | 1 \right),
\end{aligned}$$

and

$$\begin{aligned}
\underline{R}_3(y_0) &= \underline{Q}_{Y_1|X}(1 - F_{Y|X}(y_0|0) | 1) \\
&= Q_{Y|X} \left( \max \left\{ \frac{\underline{c} p_0(1 - F_{Y|X}(y_0|0))}{p_1(1 - \underline{c})}, \frac{p_1 - \bar{c} + \bar{c} p_0(1 - F_{Y|X}(y_0|0))}{p_1(1 - \bar{c})} \right\} | 1 \right) \\
&= Q_{Y|X} \left( \max \left\{ \frac{\underline{c} p_0(1 - F_{Y|X}(y_0|0))}{p_1(1 - \underline{c})}, 1 - \frac{\bar{c} p_0 F_{Y|X}(y_0|0)}{p_1(1 - \bar{c})} \right\} | 1 \right).
\end{aligned}$$

Since  $y_0$  was arbitrary, we conclude that  $\underline{R}_2(y_0) = \underline{R}_3(y_0)$  for all  $y_0 \in \mathbb{R}$ .

We next establish that  $\underline{R}_1(y_0) = \underline{R}_2(y_0)$  for all  $y_0 \in \mathbb{R}$ . Note that

$$\begin{aligned}
\underline{R}_1(y_0) &= \underline{Q}_{Y_1}(1 - \overline{F}_{Y_0}(y_0)) \\
&= Q_{Y|X} \left( \max \left\{ \frac{\underline{c}}{p_1}(1 - \overline{F}_{Y_0}(y_0)), \frac{p_1 - \bar{c}}{p_1} + \frac{\bar{c}}{p_1}(1 - \overline{F}_{Y_0}(y_0)) \right\} | 1 \right) \\
&= Q_{Y|X} \left( \max \left\{ \frac{\underline{c}}{p_1}(1 - \overline{F}_{Y_0}(y_0)), 1 - \frac{\bar{c}}{p_1} \overline{F}_{Y_0}(y_0) \right\} | 1 \right).
\end{aligned}$$

If  $y_0 < \overline{Q}_0$ , Lemma C.3.3.1 implies

$$\overline{F}_{Y_0}(y_0) = \frac{p_0}{1 - \bar{c}} F_{Y|X}(y | 0) \quad \text{and} \quad F_{Y|X}(y_0|0) < \bar{\tau}_0 = \frac{(p_1 - \underline{c})(1 - \bar{c})}{(\bar{c} - \underline{c})p_0}.$$

These two (in)equalities imply that

$$\frac{\underline{c}}{p_1}(1 - \overline{F}_{Y_0}(y_0)) < 1 - \frac{\bar{c}}{p_1} \overline{F}_{Y_0}(y_0).$$

Then it follows that

$$\underline{R}_1(y_0) = Q_{Y|X} \left( 1 - \frac{\bar{c}}{p_1} \bar{F}_{Y_0}(y_0) \mid 1 \right) = Q_{Y|X} \left( 1 - \frac{\bar{c}p_0 F_{Y|X}(y_0|0)}{p_1(1-\bar{c})} \mid 1 \right) \quad \text{if } y_0 < \bar{Q}_0. \quad (\text{C.11})$$

Similarly, it can be verified that

$$\underline{R}_1(y_0) = Q_{Y|X} \left( \frac{\underline{c}}{p_1} (1 - \bar{F}_{Y_0}(y_0)) \mid 1 \right) = Q_{Y|X} \left( \frac{\underline{c}p_0(1 - F_{Y|X}(y_0|0))}{p_1(1-\underline{c})} \mid 1 \right) \quad \text{if } y_0 \geq \bar{Q}_0. \quad (\text{C.12})$$

By Lemma C.3.3.1,  $y_0 < \bar{Q}_0$  is equivalent to  $F_{Y|X}(y_0|0) < \bar{\tau}_0$ , and it is further equivalent to

$$1 - \frac{\bar{c}p_0 F_{Y|X}(y_0|0)}{p_1(1-\bar{c})} > \frac{\underline{c}p_0(1 - F_{Y|X}(y_0|0))}{p_1(1-\underline{c})}.$$

Therefore, we can write

$$\begin{aligned} \underline{R}_2(y_0) &= Q_{Y|X} \left( 1 - \frac{\bar{c}p_0 F_{Y|X}(y_0|0)}{p_1(1-\bar{c})} \mid 1 \right) \mathbb{1}(y_0 < \bar{Q}_0) \\ &\quad + Q_{Y|X} \left( \frac{\underline{c}p_0(1 - F_{Y|X}(y_0|0))}{p_1(1-\underline{c})} \mid 1 \right) \mathbb{1}(y_0 \geq \bar{Q}_0). \end{aligned} \quad (\text{C.13})$$

By combining (C.11), (C.12), and (C.13), we note that  $\underline{R}_2(y_0) = \underline{R}_1(y_0)$  for all  $y_0 \in \mathbb{R}$ , as desired. The proof for the second set of equivalences between (d), (e), and (f) is similar and thus omitted.

**Proof of Part 11:** We show the first set of equivalences between (a), (b), and (c), and the proof for the second set of equivalences between (d), (e), and (f) follows similar arguments and thus omitted. First, we expand

$$\Delta'_1 := F_{Y|X}(\bar{Q}_1 \mid 1) + \bar{F}_{Y_0|X}(\bar{Q}_0 - \mid 1) - 1 = F_{Y|X}(\bar{Q}_1 \mid 1) + F_{Y|X}(\bar{Q}_0 - \mid 0) \frac{p_0 \bar{c}}{p_1(1-\bar{c})} - 1.$$

Next, note that

$$\begin{aligned}
\Delta'_2 &:= \bar{F}_{Y_1}(\bar{Q}_1) + \bar{F}_{Y_0}(\bar{Q}_0 -) - 1 \\
&= \frac{\bar{c} - p_1}{\bar{c}} + F_{Y|X}(\bar{Q}_1 | 1) \frac{p_1}{\bar{c}} + F_{Y|X}(\bar{Q}_0 - | 0) \frac{p_0}{1 - \bar{c}} - 1 \\
&= \frac{p_1}{\bar{c}} \left[ F_{Y|X}(\bar{Q}_1 | 1) + F_{Y|X}(\bar{Q}_0 - | 0) \frac{p_0 \bar{c}}{p_1(1 - \bar{c})} - 1 \right] \\
&= \frac{p_1}{\bar{c}} \Delta'_1,
\end{aligned}$$

and

$$\begin{aligned}
\Delta'_3 &:= \bar{F}_{Y_1|X}(\bar{Q}_1 | 0) + F_{Y|X}(\bar{Q}_0 - | 0) - 1 \\
&= \frac{\bar{c} - p_1}{\bar{c} p_0} + F_{Y|X}(\bar{Q}_1 | 1) \frac{p_1(1 - \bar{c})}{p_0 \bar{c}} + F_{Y|X}(\bar{Q}_0 - | 0) - 1 \\
&= \frac{p_1(1 - \bar{c})}{p_0 \bar{c}} \left[ F_{Y|X}(\bar{Q}_1 | 1) + F_{Y|X}(\bar{Q}_0 - | 0) \frac{p_0 \bar{c}}{p_1(1 - \bar{c})} + \frac{\bar{c} - p_1}{p_1(1 - \bar{c})} - \frac{(1 - p_1)\bar{c}}{p_1(1 - \bar{c})} \right] \\
&= \frac{p_1(1 - \bar{c})}{p_0 \bar{c}} \left[ F_{Y|X}(\bar{Q}_1 | 1) + F_{Y|X}(\bar{Q}_0 - | 0) \frac{p_0 \bar{c}}{p_1(1 - \bar{c})} - 1 \right] \\
&= \frac{p_1(1 - \bar{c})}{p_0 \bar{c}} \Delta'_1.
\end{aligned}$$

Therefore, the desired result follows by noting that  $\Delta'_1, \Delta'_2, \Delta'_3$  all have the same sign.  $\square$

*Proof of Theorem 4.4.1.* Fix a  $w \in \text{supp}(W)$  and  $(\varepsilon, \gamma, C_{1,0|1,w}, C_{1,0|0,w}) \in [0, 1]^2 \times \mathcal{C}^2$ . We prove this theorem by constructing a probability distribution  $\tilde{\mathbb{P}}$  for  $(Y_1, Y_0, X)$  conditional on  $W = w$  such that for all  $y \in \mathbb{R}$ ,  $x \in \{0, 1\}$ , and  $(y_1, y_0) \in \mathbb{R}^2$ , the following conditions hold:

1.  $\tilde{\mathbb{P}}(Y_1 \leq y | W = w) = \varepsilon \underline{F}_{Y_1|W}(y | w) + (1 - \varepsilon) \bar{F}_{Y_1|W}(y | w)$  and  
 $\tilde{\mathbb{P}}(Y_0 \leq y | W = w) = \gamma \underline{F}_{Y_0|W}(y | w) + (1 - \gamma) \bar{F}_{Y_0|W}(y | w)$ ;
2.  $\tilde{\mathbb{P}}(X = x | W = w) = p_{x|w}$ ;
3.  $\tilde{\mathbb{P}}(Y_x \leq y | X = x, W = w) = F_{Y|X,W}(Y | X, w)$ ;

4.  $\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid X = x, W = w) = C_{1,0|x,w}(\tilde{\mathbb{P}}(Y_1 \leq y_1 \mid X = x, W = w), \tilde{\mathbb{P}}(Y_1 \leq y_0 \mid X = x, W = w));$
5.  $\tilde{\mathbb{P}}(X = 1 \mid Y_x, W = w) \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$   $\tilde{\mathbb{P}}$ -almost surely.

Condition 1 requires that an arbitrary convex combination of marginal cdf bounds stated in Theorem 4.4.1 can be achieved by the constructed measure. Condition 4 then states that any bivariate copula  $C_{1,0|x,w}$  is also achievable. Conditions 2 and 3 require the constructed measure generate the same distribution of  $(Y, X)$  as the observed data conditional on  $W = w$ . Finally, Condition 5 requires the marginal  $c$ -dependence Assumption 12 to be satisfied for the constructed measure when conditioning on  $W = w$ . As a result, the constructed measure  $\tilde{\mathbb{P}}$  generates the marginal cdfs and copulas in Theorem 4.4.1 and satisfies all the requirements in the definition of identified set  $\mathcal{I}_0^{\text{marg}}(F_{Y,X,W})$ .

For the conciseness of the proof, we write  $C_{1,0|x,w}$  as  $C_{x,w}$  for  $x \in \{0, 1\}$  so that subscripts of copulas denote the conditioning variables.

Let

$$\begin{aligned} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x \mid W = w) &= xC_{1,w}(F_{Y_1|X,W}(y_1 \mid 1, w), F_0(y_0 \mid 1, w; \gamma))p_{1|w} \\ &\quad + (1-x)C_{0,w}(F_1(y_1 \mid 0, w; \varepsilon), F_{Y_1|X,W}(y_0|0, w))p_{0|w}. \end{aligned} \tag{C.14}$$

where

$$F_0(y_0 \mid 1, w; \gamma) = \gamma \underline{F}_{Y_0|X,W}(y_0 \mid 1, w) + (1-\gamma) \overline{F}_{Y_0|X,W}(y_0 \mid 1, w)$$

$$F_1(y_1 \mid 0, w; \varepsilon) = \varepsilon \underline{F}_{Y_1|X,W}(y_1 \mid 0, w) + (1-\varepsilon) \overline{F}_{Y_1|X,W}(y_1 \mid 0, w).$$

Since convex combinations of cdfs are cdfs, and by Lemma C.3.1.1, both  $F_0(\cdot \mid 1, w; \gamma)$  and  $F_1(\cdot \mid 0, w; \varepsilon)$  are cdfs. By Sklar's Theorem (Nelsen, 2006, Theorem 2.3.3), the expression in (C.14) is a joint distribution function for  $(Y_1, Y_0, X)$  conditional on  $W = w$ .

For the rest of the proof, we verify conditions 1-5 for the constructed measure  $\tilde{\mathbb{P}}$ .

**Verifying Condition 1:** For  $y \in \mathbb{R}$ , we can see that

$$\begin{aligned}
\tilde{\mathbb{P}}(Y_1 \leq y \mid W = w) &= \sum_{x \in \{0,1\}} \lim_{y_0 \rightarrow +\infty} \tilde{\mathbb{P}}(Y_1 \leq y, Y_0 \leq y_0, X = x \mid W = w) \\
&= \lim_{y_0 \rightarrow +\infty} C_{1,w}(F_{Y|X,W}(y_1 \mid 1, w), F_0(y_0 \mid 1, w; \gamma))p_{1|w} \\
&\quad + \lim_{y_0 \rightarrow +\infty} C_{0,w}(F_1(y_1 \mid 0, w; \varepsilon), F_{Y|X,W}(y_0 \mid 0, w))p_{0|w} \\
&= C_{1,w}(F_{Y|X,W}(y_1 \mid 1, w), 1)p_{1|w} + C_{0,w}(F_1(y_1 \mid 0, w; \varepsilon), 1)p_{0|w} \\
&= F_{Y|X,W}(y_1 \mid 1, w)p_{1|w} + F_1(y_1 \mid 0, w; \varepsilon)p_{0|w} \\
&= \varepsilon(F_{Y|X,W}(y_1 \mid 1, w)p_{1|w} + \underline{F}_{Y_1|X,W}(y_1 \mid 0, w)p_{0|w}) \\
&\quad + (1 - \varepsilon)(F_{Y|X,W}(y_1 \mid 1, w)p_{1|w} + \overline{F}_{Y_1|X,W}(y_1 \mid 0, w)p_{0|w}) \\
&= \varepsilon \underline{F}_{Y_1|W}(y \mid w) + (1 - \varepsilon) \overline{F}_{Y_1|W}(y \mid w).
\end{aligned}$$

The third line holds since  $C_{x,w}(1, u) = C_{x,w}(u, 1) = u$  for  $x \in \{0, 1\}$  and  $u \in [0, 1]$ . The last line holds by Lemma C.3.1.2.

Likewise,

$$\begin{aligned}
\tilde{\mathbb{P}}(Y_0 \leq y \mid W = w) &= \sum_{x \in \{0,1\}} \lim_{y_1 \rightarrow +\infty} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y, X = x \mid W = w) \\
&= \gamma \underline{F}_{Y_0|W}(y \mid w) + (1 - \gamma) \overline{F}_{Y_0|W}(y \mid w).
\end{aligned}$$

**Verifying Condition 2:** For  $x \in \{0, 1\}$ , we have that

$$\begin{aligned}
\tilde{\mathbb{P}}(X = x \mid W = w) &= \lim_{y_1, y_0 \rightarrow \infty} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x \mid W = w) \\
&= xC_{1,w}(1, 1)p_{1|w} + (1 - x)C_{0,w}(1, 1)p_{0|w} \\
&= xp_{1|w} + (1 - x)p_{0|w} \\
&= p_{x|w}.
\end{aligned}$$

The third equality uses the fact that  $C_{x,w}(1, 1) = 1$  for  $x \in \{0, 1\}$ .

**Verifying Condition 3:** For  $x \in \{0, 1\}$  and  $y \in \mathbb{R}$ , we have that

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_x \leq y \mid X = x, W = w) \\
&= \lim_{y' \rightarrow +\infty} \frac{\tilde{\mathbb{P}}(Y_x \leq y, Y_{1-x} \leq y', X = x \mid W = w)}{\tilde{\mathbb{P}}(X = x \mid W = w)} \\
&= \frac{xC_{1,w}(F_{Y|X,W}(y \mid 1, w), 1)p_{1|w} + (1-x)C_{0,w}(1, F_{Y|X,W}(y \mid 0, w))p_{0|w}}{p_{x|w}} \\
&= \frac{x F_{Y|X,W}(y \mid 1, w)p_{1|w} + (1-x) F_{Y|X,W}(y \mid 0, w)p_{0|w}}{p_{x|w}} \\
&= \frac{F_{Y|X,W}(Y \mid X, w)p_{x|w}}{p_{x|w}} \\
&= F_{Y|X,W}(Y \mid X, w).
\end{aligned}$$

The third line holds again by  $C_x(1, u) = C_x(u, 1) = u$  for  $x \in \{0, 1\}$  and  $u \in [0, 1]$ . The last line follows by Assumption 11 that  $p_{x|w} > 0$  for  $x \in \{0, 1\}$ .

**Verifying Condition 4:** First, following similar steps for verifying condition 3, we have

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_x \leq y \mid X = 1 - x, W = w) \\
&= \frac{\lim_{y' \rightarrow +\infty} \tilde{\mathbb{P}}(Y_x \leq y, Y_{1-x} \leq y', X = 1 - x \mid W = w)}{\tilde{\mathbb{P}}(X = 1 - x \mid W = w)} \\
&= \frac{(1-x)C_{1,w}(1, F_0(y_0 \mid 1, w; \gamma))p_{1|w} + xC_{0,w}(F_1(y_1 \mid 0, w; \epsilon), 1)p_{0|w}}{p_{1-x|w}} \tag{C.15} \\
&= \frac{(1-x)p_{1|w}F_0(y_0 \mid 1, w; \gamma) + xp_{0|w}F_1(y_1 \mid 0, w; \epsilon)}{p_{1-x|w}} \\
&= (1-x)F_0(y_0 \mid 1, w; \gamma) + xF_1(y_1 \mid 0, w; \epsilon).
\end{aligned}$$

Then for  $(y_1, y_0) \in \mathbb{R}^2$ , it follows that

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 | X = x) \\
&= xC_{1,w}(F_{Y|X,W}(y_1 | 1, w), F_0(y_0 | 1, w; \gamma)) + (1-x)C_{0,w}(F_1(y_1 | 0, w; \epsilon), F_{Y|X,W}(y_0 | 0, w)) \\
&= xC_{1,w}(\tilde{\mathbb{P}}(Y_1 \leq y_1 | X = 1, W = w), \tilde{\mathbb{P}}(Y_0 \leq y_0 | X = 1, W = w)) \\
&\quad + (1-x)C_{0,w}(\tilde{\mathbb{P}}(Y_1 \leq y_1 | X = 0, W = w), \tilde{\mathbb{P}}(Y_0 \leq y_0 | X = 0, W = w)) \\
&= C_{x,w}(\tilde{\mathbb{P}}(Y_1 \leq y_1 | X = x, W = w), \tilde{\mathbb{P}}(Y_0 \leq y_0 | X = x, W = w)).
\end{aligned}$$

The second line holds by Condition 3 and equation (C.15).

**Verifying Condition 5:** In this part, we establish an explicit formula of the propensity score function under  $\tilde{\mathbb{P}}$  and show that it is contained in  $[\underline{c}(w, \eta), \bar{c}(w, \eta)]$  almost surely. To achieve this goal, we divide the analysis into two cases.

**Case 1:** Consider the case where  $p_{1|w} = \underline{c}(w, \eta)$ . By direct calculation,

$$\underline{F}_{Y_1|W}(y | w) = \bar{F}_{Y_1|W}(y | w) = F_{Y|X,W}(y | 1, w)$$

and

$$\underline{F}_{Y_0|W}(y | w) = \bar{F}_{Y_0|W}(y | w) = F_{Y|X,W}(y | 0, w).$$

Based on condition 1 we verified above, we have

$$\tilde{\mathbb{P}}(Y_1 \leq y_1 | W = w) = F_{Y|X,W}(y_1 | 1, w) \quad \text{and} \quad \tilde{\mathbb{P}}(Y_0 \leq y_0 | W = w) = F_{Y|X,W}(y_0 | 0, w).$$

Since  $p_{1|w} = \underline{c}(w, \eta)$ , by Assumption 12, it is straightforwardly to see that  $\mathbb{P}(X = 1 | Y_1, W = w) = \mathbb{P}(X = 1 | Y_0, W = w) = \underline{c}(w, \eta)$  almost surely, which further implies

$$\begin{aligned}
\tilde{\mathbb{E}} \left[ \frac{\mathbb{1}[Y \leq y_1]X}{\underline{c}(w, \eta)} | W = w \right] &= \mathbb{E} \left[ \frac{\mathbb{1}[Y \leq y_1]X}{\underline{c}(w, \eta)} | W = w \right] \\
&= F_{Y|X,W}(y_1 | 1, w) \\
&= \tilde{\mathbb{P}}(Y_1 \leq y_1 | W = w)
\end{aligned}$$

and

$$\begin{aligned}
\tilde{\mathbb{E}} \left[ \frac{\mathbb{1}[Y \leq y_0](1-X)}{1-\underline{c}(w, \eta)} \middle| W = w \right] &= \mathbb{E} \left[ \frac{\mathbb{1}[Y \leq y_0](1-X)}{1-\underline{c}(w, \eta)} \middle| W = w \right] \\
&= F_{Y|X,W}(y_0|0, w) \\
&= \tilde{\mathbb{P}}(Y_0 \leq y_0 \mid W = w).
\end{aligned}$$

Following Lemma C.3.2, this implies  $\tilde{\mathbb{E}}(X \mid Y_1, W = w) = \tilde{\mathbb{E}}(X \mid Y_0, W = w) = \underline{c}(w, \eta)$  almost surely under  $\tilde{\mathbb{P}}$ , which is naturally bounded within  $[\underline{c}(w, \eta), \bar{c}(w, \eta)]$ , as desired. The proof for the case where  $p_{1|w} = \bar{c}(w, \eta)$  follows the same argument by interchanging  $\underline{c}(w, \eta)$  with  $\bar{c}(w, \eta)$  and thus omitted.

**Case 2:** Consider the case where  $\underline{c}(w, \eta) < p_{1|w} < \bar{c}(w, \eta)$ . Define

$$p_1(y, w; \varepsilon) = \frac{1}{\varepsilon \underline{p}_1(y, w)^{-1} + (1 - \varepsilon) \bar{p}_1(y, w)^{-1}},$$

where  $\bar{p}_1(y, w)$  and  $\underline{p}_1(y, w)$  are defined in Appendix C.1.

By Lemma C.3.4.4,  $\underline{p}_1(Y_1, w), \bar{p}_1(Y_1, w) \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$  almost surely. Therefore,

$$p_1(Y_1, w; \varepsilon) = \frac{1}{\varepsilon \underline{p}_1(Y_1, w)^{-1} + (1 - \varepsilon) \bar{p}_1(Y_1, w)^{-1}} \leq \frac{1}{\varepsilon \bar{c}(w, \eta)^{-1} + (1 - \varepsilon) \bar{c}(w, \eta)^{-1}} = \bar{c}(w, \eta)$$

and

$$p_1(Y_1, w; \varepsilon) = \frac{1}{\varepsilon \underline{p}_1(Y_1, w)^{-1} + (1 - \varepsilon) \bar{p}_1(Y_1, w)^{-1}} \geq \frac{1}{\varepsilon \underline{c}(w, \eta)^{-1} + (1 - \varepsilon) \underline{c}(w, \eta)^{-1}} = \underline{c}(w, \eta).$$

Therefore  $p_1(Y_1, w; \varepsilon) \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$  almost surely.

Next we will show that  $\tilde{\mathbb{E}}[X \mid Y_1, W = w] = p_1(Y_1, w; \varepsilon)$  via Lemma C.3.2 by verifying that for all  $y \in \mathbb{R}$ :

$$\tilde{\mathbb{E}} \left[ \frac{\mathbb{1}(Y_1 \leq y)X}{p_1(Y_1, w; \varepsilon)} \middle| W = w \right] = \tilde{\mathbb{P}}(Y_1 \leq y \mid W = w) = \varepsilon \underline{F}_{Y_1|W}(y \mid w) + (1 - \varepsilon) \bar{F}_{Y_1|W}(y \mid w).$$

To show this, we have the following derivations:

$$\begin{aligned}
\tilde{\mathbb{E}} \left[ \frac{\mathbb{1}(Y_1 \leq y)X}{p_1(Y_1, w; \varepsilon)} \mid W = w \right] &= \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)X}{p_1(Y, w; \varepsilon)} \mid W = w \right] \\
&= \mathbb{E} \left[ \mathbb{1}(Y \leq y)X \left( \frac{\varepsilon}{\underline{p}_1(Y, w)} + \frac{1 - \varepsilon}{\bar{p}_1(Y, w)} \right) \mid W = w \right] \\
&= \varepsilon \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)X}{\underline{p}_1(Y, w)} \mid W = w \right] + (1 - \varepsilon) \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)X}{\bar{p}_1(Y, w)} \mid W = w \right] \\
&= \varepsilon F_{Y_1|W}(y \mid w) + (1 - \varepsilon) \bar{F}_{Y_1|W}(y \mid w).
\end{aligned}$$

The first equality holds by noting that the distribution of  $Y_1$  conditional  $X = 1$  and  $W = w$  under  $\tilde{\mathbb{P}}$  is the same as the one under the population  $\mathbb{P}$  as verified by condition 2. The last equality follows by applying Lemma C.3.4.5.

For the cdf of  $Y_0$ , define

$$p_0(Y_0, w; \gamma) = 1 - \frac{1}{\gamma(1 - \underline{p}_0(Y_0, w))^{-1} + (1 - \gamma)(1 - \bar{p}_0(Y_0, w))^{-1}}.$$

Since  $1 - \underline{p}_0(Y_0, w), 1 - \bar{p}_0(Y_0, w) \in [1 - \bar{c}(w, \eta), 1 - \underline{c}(w, \eta)]$  almost surely, we have that

$$\begin{aligned}
p_0(Y_0, w; \gamma) &= 1 - \frac{1}{\gamma(1 - \underline{p}_0(Y_0, w))^{-1} + (1 - \gamma)(1 - \bar{p}_0(Y_0, w))^{-1}} \\
&\leq 1 - \frac{1}{\gamma(1 - \bar{c}(w, \eta))^{-1} + (1 - \gamma)(1 - \bar{c}(w, \eta))^{-1}} \\
&= \bar{c}(w, \eta),
\end{aligned}$$

and

$$\begin{aligned}
p_0(Y_0, w; \gamma) &= 1 - \frac{1}{\gamma(1 - \underline{p}_0(Y_0, w))^{-1} + (1 - \gamma)(1 - \bar{p}_0(Y_0, w))^{-1}} \\
&\geq 1 - \frac{1}{\gamma(1 - \underline{c}(w, \eta))^{-1} + (1 - \gamma)(1 - \underline{c}(w, \eta))^{-1}} \\
&= \underline{c}(w, \eta).
\end{aligned}$$

Therefore,  $p_0(Y_0, w; \eta) \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$  almost surely. We can also see that

$$\begin{aligned}
& \tilde{\mathbb{E}} \left[ \frac{\mathbb{1}(Y_0 \leq y)(1 - X)}{1 - p_0(Y_0, w; \gamma)} \mid W = w \right] \\
&= \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)(1 - X)}{1 - p_0(Y, w; \eta)} \mid W = w \right] \\
&= \mathbb{E} \left[ \mathbb{1}(Y \leq y)(1 - X) \left( \frac{\gamma}{1 - \underline{p}_0(Y_0, w)} + \frac{1 - \gamma}{1 - \bar{p}_0(Y_0, w)} \right) \mid W = w \right] \\
&= \gamma \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)(1 - X)}{1 - \underline{p}_0(Y_0, w)} \mid W = w \right] \\
&\quad + (1 - \gamma) \mathbb{E} \left[ \frac{\mathbb{1}(Y \leq y)(1 - X)}{1 - \bar{p}_0(Y_0, w)} \mid W = w \right] \\
&= \gamma \underline{F}_{Y_0|W}(y \mid w) + (1 - \gamma) \bar{F}_{Y_0|W}(y \mid w),
\end{aligned}$$

where the last equality follows by Lemma C.3.4.5. Therefore, by Lemma C.3.2,  $\tilde{\mathbb{P}}(X = 1 \mid Y_0, W = w) = p_0(Y_0, w; \eta) \in [\underline{c}(w, \eta), \bar{c}(w, \eta)]$  almost surely, which concludes the proof.  $\square$

### C.3.2 Proof of Theorem 4.4.2

This appendix provides a proof of Theorem 4.4.2 and all of its auxiliary lemmas. First, we define four latent propensity score functions. For  $w \in \text{supp}(W)$ , let

$$p^{ul}(y_1, y_0, w; B) = \begin{cases} \underline{c} & \text{if } y_1 \leq \bar{Q}_1, y_0 \leq \underline{Q}_0, (y_1, y_0) \neq (\bar{Q}_1, \underline{Q}_0) \\ B & \text{if } (y_1, y_0) = (\bar{Q}_1, \underline{Q}_0) \\ \bar{c} & \text{if } y_1 \geq \bar{Q}_1, y_0 \geq \underline{Q}_0, (y_1, y_0) \neq (\bar{Q}_1, \underline{Q}_0) \\ p_{1|w} & \text{otherwise,} \end{cases} \quad (\text{C.16})$$

$$p^{uu}(y_1, y_0, w; B) = \begin{cases} \underline{c} & \text{if } y_1 \leq \bar{Q}_1, y_0 \geq \bar{Q}_0, (y_1, y_0) \neq (\bar{Q}_1, \bar{Q}_0) \\ B & \text{if } (y_1, y_0) = (\bar{Q}_1, \bar{Q}_0) \\ \bar{c} & \text{if } y_1 \geq \bar{Q}_1, y_0 \leq \bar{Q}_0, (y_1, y_0) \neq (\bar{Q}_1, \bar{Q}_0) \\ p_{1|w} & \text{otherwise,} \end{cases} \quad (\text{C.17})$$

$$p^{lu}(y_1, y_0, w; B) = \begin{cases} \bar{c} & \text{if } y_1 \leq \bar{Q}_1, y_0 \leq \underline{Q}_0, (y_1, y_0) \neq (\bar{Q}_1, \underline{Q}_0) \\ B & \text{if } (y_1, y_0) = (\bar{Q}_1, \underline{Q}_0) \\ \underline{c} & \text{if } y_1 \geq \bar{Q}_1, y_0 \geq \underline{Q}_0, (y_1, y_0) \neq (\bar{Q}_1, \underline{Q}_0) \\ p_{1|w} & \text{otherwise,} \end{cases} \quad (\text{C.18})$$

$$p^u(y_1, y_0, w; B) = \begin{cases} \bar{c} & \text{if } y_1 \leq \bar{Q}_1, y_0 \geq \bar{Q}_0, (y_1, y_0) \neq (\bar{Q}_1, \bar{Q}_0) \\ B & \text{if } (y_1, y_0) = (\bar{Q}_1, \bar{Q}_0) \\ \underline{c} & \text{if } y_1 \geq \bar{Q}_1, y_0 \leq \bar{Q}_0, (y_1, y_0) \neq (\bar{Q}_1, \bar{Q}_0) \\ p_{1|w} & \text{otherwise.} \end{cases} \quad (\text{C.19})$$

By appropriately specifying the constant  $B$  in these propensity scores, we can show that they correspond to the propensity scores  $\mathbb{P}(X = 1 \mid Y_1, Y_0, W = w)$  under joint  $c$ -dependence for all four pairs of cdf bounds. Before showing this, we state and prove three auxiliary lemmas.

**Lemma C.3.5.** *Let  $w \in \text{supp}(W)$ . Suppose  $m(\cdot)$  is a Borel measurable function and  $\mathbb{P}(m(Y_1, Y_0) > \delta \mid W = w) = 1$  for some  $\delta > 0$ . The following statements are equivalent:*

1. *Conditional on  $W = w$ , the following statement holds almost surely:*

$$m(Y_1, Y_0) = \mathbb{E}[X \mid Y_1, Y_0, W = w]. \quad (\text{C.20})$$

2. *For all  $(y_1, y_0) \in \mathbb{R}^2$ , the following equality holds:*

$$\mathbb{E}[\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0)m(Y_1, Y_0) \mid W = w] = \mathbb{P}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 \mid W = w). \quad (\text{C.21})$$

*Proof of Lemma C.3.5.* We first show (C.20) implies (C.21), note that

$$\begin{aligned} \mathbb{P}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 \mid W = w) &= \mathbb{E}[\mathbb{1}[Y_1 \leq y_1, Y_0 \leq y_0]X \mid W = w] \\ &= \mathbb{E}(\mathbb{E}[\mathbb{1}[Y_1 \leq y_1, Y_0 \leq y_0]X \mid Y_1, Y_0, W = w] \mid W = w) \\ &= \mathbb{E}(\mathbb{1}[Y_1 \leq y_1, Y_0 \leq y_0]\mathbb{E}[X \mid Y_1, Y_0, W = w] \mid W = w) \\ &= \mathbb{E}(\mathbb{1}[Y_1 \leq y_1, Y_0 \leq y_0]m(Y_1, Y_0) \mid W = w). \end{aligned} \quad (\text{C.22})$$

where we use the law of iterated expectation in the second line and use (C.20) in the last line of derivation.

Next, we show that (C.21) implies (C.20). To establish this result, we first note a few key facts:

1. Following from the last two lines of (C.22), the law of iterated expectations implies

$$\begin{aligned} & \mathbb{E}[\mathbb{1}[Y_1 \leq y_1, Y_0 \leq y_0] \mathbb{E}(X \mid Y_1, Y_0, W = w) \mid W = w] \\ &= \mathbb{E}[\mathbb{1}[Y_1 \leq y_1, Y_0 \leq y_0] m(Y_1, Y_0) \mid W = w] \end{aligned}$$

for each  $(y_1, y_0) \in \mathbb{R}^2$ .

2. For  $(y_1, y_0) \in \mathbb{R}^2$ , define the preimage from a half-space on  $\mathbb{R}^2$ :

$$I_{y_1, y_0} = \{\omega \in \Omega : Y_1(\omega) \leq y_1, Y_0(\omega) \leq y_0\}$$

and let  $\mathcal{A}_2 := \{I_{y_1, y_0} : (y_1, y_0) \in \mathbb{R}^2\}$ . Similar to the proof of lemma C.3.2, the class of sets  $\mathcal{A}_2$  is a  $\pi$ -system.

3. The sample space can be written as a countable union of sets in  $\mathcal{A}_2$ :

$$\Omega = \{\omega \in \Omega : Y_1(\omega) < \infty, Y_0(\omega) < \infty\} = \bigcup_{n=1}^{\infty} I_{n, n}.$$

4. The random variable  $m(Y_1, Y_0)$  is measurable with respect to the  $\sigma$ -algebra generated by  $(Y_1, Y_0)$  due to the Borel measurability of  $m(\cdot)$ , and it is integrable since  $\mathbb{E}(m(Y_1, Y_0) \mid W = w) = \mathbb{P}(X = 1 \mid W = w) < \infty$  by sending  $y_1$  and  $y_0$  to infinity in (C.22).
5. The  $\sigma$ -algebra generated by  $\mathcal{A}_2$  equals the  $\sigma$ -algebra generated by  $(Y_1, Y_0)$ , i.e.,

$$\sigma(\mathcal{A}_2) = \sigma(Y_1, Y_0).$$

To show this, define the mapping  $f : \Omega \rightarrow \mathbb{R}^2$  as  $f(\omega) \mapsto (Y_1(\omega), Y_0(\omega))$  and  $\mathcal{F} = \{(-\infty, y_1] \times (-\infty, y_0] : (y_1, y_0) \in \mathbb{R}^2\}$ . Note that

$$\sigma(\mathcal{A}_2) = \sigma(f^{-1}(\mathcal{F})) = f^{-1}(\sigma(\mathcal{F})).$$

Since the Borel  $\sigma$ -algebra on  $\mathbb{R}^2$  can be generated by elements in  $\mathcal{F}$ , we have  $\sigma(\mathcal{F}) = \mathcal{B}(\mathbb{R}^2)$ . This implies

$$f^{-1}(\sigma(\mathcal{F})) = f^{-1}(\mathcal{B}(\mathbb{R}^2)) := \sigma(Y_1, Y_0).$$

Therefore the desired conclusion holds.

Given the above results, it follows by Billingsley (1995, Theorem 34.1) that

$$\mathbb{E}[X \mid Y_1, Y_0, W = w] = m(Y_1, Y_0),$$

almost surely conditional on  $W = w$ , as desired.  $\square$

**Lemma C.3.6.** *Let  $w \in \text{supp}(W)$ . Consider a probability distribution defined by*

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x \mid W = w) \\ &= \min\{F_{Y|X,W}(y_1 \mid 1, w), \underline{F}_{Y_0|X,W}(y_0 \mid 1, w)\}p_{1|w}x \\ &+ \min\{\overline{F}_{Y_1|X,W}(y_1 \mid 0, w), F_{Y|X,W}(y_0|0, w)\}p_{0|w}(1 - x), \end{aligned}$$

then

$$\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) = \min\{\overline{F}_{Y_1|W}(y_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w)\}.$$

Also for the following distribution,

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x \mid W = w) \\ &= \min\{F_{Y|X,W}(y_1 \mid 1, w), \overline{F}_{Y_0|X,W}(y_0 \mid 1, w)\}p_{1|w}x \\ &+ \min\{\underline{F}_{Y_1|X,W}(y_1 \mid 0, w), F_{Y|X,W}(y_0|0, w)\}p_{0|w}(1 - x) \end{aligned}$$

implies

$$\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) = \min\{\underline{F}_{Y_1}(y_1 \mid w), \overline{F}_{Y_0}(y_0 \mid w)\}.$$

*Proof of Lemma C.3.6.* Consider the first statement with  $p_{1|w} = c$ . Then it follows that  $\overline{F}_{Y_1|X,W}(y_1 \mid 0, w) = \overline{F}_{Y_1|W}(y_1 \mid w) = F_{Y|X,W}(y_1 \mid 1, w)$  and  $\underline{F}_{Y_0|X,W}(y_0 \mid 1, w) = \underline{F}_{Y_0|W}(y_0 \mid w) = F_{Y|X,W}(y_0|0, w)$ . Therefore,

$$\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x \mid W = w) = \min\{F_{Y|X,W}(y_1 \mid 1, w), F_{Y|X,W}(y_0|0, w)\}p_{x|w}.$$

This implies

$$\begin{aligned} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) &= \min\{F_{Y|X,W}(y_1 \mid 1, w), F_{Y|X,W}(y_0|0, w)\} \\ &= \min\{\overline{F}_{Y_1|W}(y_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w)\} \end{aligned}$$

as desired. The proof for the case where  $p_{1|w} = \bar{c}$  follows the same arguments and thus omitted.

Next consider  $\underline{c} < p_{1|w} < \bar{c}$ . We have that

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) \\
&= \min\{F_{Y|X,W}(y_1 \mid 1, w), \underline{F}_{Y_0|X,W}(y_0 \mid 1, w)\}p_{1|w} \\
&\quad + \min\{\bar{F}_{Y_1|X,W}(y_1 \mid 0, w), F_{Y|X,W}(y_0|0, w)\}p_{0|w} \\
&= F_{Y|X,W}(y_1 \mid 1, w)p_{1|w}\mathbb{1}(\underline{F}_{Y_0|X,W}(y_0 \mid 1, w) \geq F_{Y|X,W}(y_1 \mid 1, w)) \\
&\quad + \underline{F}_{Y_0|X,W}(y_0 \mid 1, w)p_{1|w}\mathbb{1}(\underline{F}_{Y_0|X,W}(y_0 \mid 1, w) < F_{Y|X,W}(y_1 \mid 1, w)) \\
&\quad + \bar{F}_{Y_1|X,W}(y_1 \mid 0, w)p_{0|w}\mathbb{1}(F_{Y|X,W}(y_0|0, w) \geq \bar{F}_{Y_1|X,W}(y_1 \mid 0, w)) \\
&\quad + F_{Y|X,W}(y_0|0, w)p_{0|w}\mathbb{1}(F_{Y|X,W}(y_0|0, w) < \bar{F}_{Y_1|X,W}(y_1 \mid 0, w)) \\
&= (F_{Y|X,W}(y_1 \mid 1, w)p_{1|w} + \bar{F}_{Y_1|X,W}(y_1 \mid 0, w)p_{0|w})\mathbb{1}[\underline{F}_{Y_0|W}(y_0 \mid w) \geq \bar{F}_{Y_1|W}(y_1 \mid w)] \\
&\quad + (\underline{F}_{Y_0|X,W}(y_0 \mid 1, w)p_{1|w} + F_{Y|X,W}(y_0|0, w)p_{0|w})\mathbb{1}[\underline{F}_{Y_0|W}(y_0 \mid w) < \bar{F}_{Y_1|W}(y_1 \mid w)] \\
&= \min\{\bar{F}_{Y_1|W}(y_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w)\}.
\end{aligned}$$

The third equality follows by the first set of equivalences in Lemma C.3.4.6 after setting  $x = 1$ . The last equality follows by Lemma C.3.1.2. The second statement follows similar arguments but instead uses Lemma C.3.4.6 by setting  $x = 0$ . Therefore, the proof is complete.  $\square$

**Lemma C.3.7.** *Let  $w \in \text{supp}(W)$ . Consider a probability distribution defined by*

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x \mid W = w) \\
&= \max\{F_{Y|X,W}(y_1 \mid 1, w) + \bar{F}_{Y_0|X,W}(y_0 \mid 1, w) - 1, 0\}p_{1|w}x \\
&\quad + \max\{\bar{F}_{Y_1|X,W}(y_1 \mid 0, w) + F_{Y|X,W}(y_0|0, w) - 1, 0\}p_{0|w}(1 - x)
\end{aligned}$$

then

$$\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) = \max\{\bar{F}_{Y_1|W}(y_1 \mid w) + \bar{F}_{Y_0|W}(y_0 \mid w) - 1, 0\}.$$

Also for the following distribution,

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x \mid W = w) \\ &= \max\{F_{Y|X,W}(y_1 \mid 1, w) + \underline{E}_{Y_0|X,W}(y_0 \mid 1, w) - 1, 0\}p_{1|w}x \\ & \quad + \max\{\underline{E}_{Y_1|X,W}(y_1 \mid 0, w) + F_{Y|X,W}(y_0|0, w) - 1, 0\}p_{0|w}(1 - x) \end{aligned}$$

implies

$$\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) = \max\{\underline{E}_{Y_1|W}(y_1 \mid w) + \underline{E}_{Y_0|W}(y_0 \mid w) - 1, 0\}.$$

*Proof of Lemma C.3.7.* Consider the first statement. Similar arguments from the proof of Lemma C.3.6 can be used to establish the desired result for  $p_{1|w} = \underline{c}$  or  $p_{1|w} = \bar{c}$ . Thus we consider the case where  $\underline{c} < p_{1|w} < \bar{c}$ . Then we have that

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) \\ &= \max\{F_{Y|X,W}(y_1 \mid 1, w) + \bar{F}_{Y_0|X,W}(y_0 \mid 1, w) - 1, 0\}p_{1|w} \\ & \quad + \max\{\bar{F}_{Y_1|X,W}(y_1 \mid 0, w) + F_{Y|X,W}(y_0|0, w) - 1, 0\}p_{0|w} \\ &= \max\{p_{1|w}(F_{Y|X,W}(y_1 \mid 1, w) + \bar{F}_{Y_0|X,W}(y_0 \mid 1, w) - 1), 0\} \\ & \quad + \max\{p_{0|w}(\bar{F}_{Y_1|X,W}(y_1 \mid 0, w) + F_{Y|X,W}(y_0|0, w) - 1), 0\} \\ &= \max\left\{\sum_{x=0,1} p_{x|w}F_{Y|X,W}(y_x|x, w) + p_{0|w}\bar{F}_{Y_1|X,W}(y_1 \mid 0, w) + p_{1|w}\bar{F}_{Y_0|X,W}(y_0 \mid 1, w) - (p_{1|w} + p_{0|w}), 0\right\} \\ &= \max\{\bar{F}_{Y_1|W}(y_1 \mid w), \bar{F}_{Y_0|W}(y_0 \mid w) - 1, 0\}. \end{aligned}$$

The second equality follows by the first set of equivalences in Lemma C.3.4.10 by setting  $x = 1$ . The last equality holds by Lemma C.3.1.2. The second statement follows similar arguments but instead uses Lemma C.3.4.10 on the second set of equivalences regarding lower bounds of cdfs. Therefore, the proof is complete.  $\square$

**Lemma C.3.8.** *Let assumptions 11 and 13 hold. Let  $\bar{\mathcal{C}}_{1,0|X,W}$  and  $\underline{\mathcal{C}}_{1,0|X,W}$  denote classes of comonotonic (and counter-monotonic) copulas where  $\bar{\mathcal{C}}_{1,0|x,w}(u, v) = \min\{u, v\}$  and  $\underline{\mathcal{C}}_{1,0|x,w}(u, v) = \max\{u + v - 1, 0\}$  for all  $(x, w) \in \{0, 1\} \times \text{supp}(W)$ . Then each of the following terms is contained in the identified set  $\mathcal{I}_0^j(F_{Y,X,W})$ :*

1.  $(\bar{F}_{Y_1|W}, \underline{E}_{Y_0|W}, \bar{\mathcal{C}}_{1,0|X,W})$ ;

2.  $(\overline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, \underline{C}_{1,0|X,W});$
3.  $(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, \overline{C}_{1,0|X,W});$
4.  $(\underline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, \underline{C}_{1,0|X,W}).$

*Proof of Lemma C.3.8. Proof of Part 1:* Fix a  $w \in \text{supp}(W)$ . We prove the first statement by constructing a probability distribution  $\tilde{\mathbb{P}}$  for  $(Y_1, Y_0, X)$  conditional on  $W = w$  such that for all  $y \in \mathbb{R}$  and  $x \in \{0, 1\}$ , we have

1.  $\tilde{\mathbb{P}}(Y_1 \leq y | W = w) = \overline{F}_{Y_1|W}(y | w)$  and  $\tilde{\mathbb{P}}(Y_0 \leq y | W = w) = \underline{F}_{Y_0|W}(y | w);$
2.  $\tilde{\mathbb{P}}(X = x | W = w) = p_{x|w};$
3.  $\tilde{\mathbb{P}}(Y_x \leq y | X = x, W = w) = F_{Y|X,W}(Y | X, w);$
4. The following equality holds:

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 | X = x, W = w) \\ &= \min \left\{ \tilde{\mathbb{P}}(Y_1 \leq y_1 | X = x, W = w), \tilde{\mathbb{P}}(Y_0 \leq y_0 | X = x, W = w) \right\}; \end{aligned}$$

5.  $\tilde{\mathbb{P}}(X = 1 | Y_1, Y_0, W = w) \in [\underline{c}, \bar{c}]$  for  $\tilde{\mathbb{P}}$ -almost surely.

Similar to the arguments in the proof of Theorem 4.4.1, Conditions 1–5 ensures that the constructed distribution  $\tilde{\mathbb{P}}$  generates the desired marginal cdfs and copulas in Lemma C.3.8.1 and satisfies all the requirements in the definition of identified set  $\mathcal{I}_0^j(F_{Y,X,W})$ .

Let

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x | W = w) \\ &= x \min\{F_{Y|X,W}(y_1 | 1, w), \underline{F}_{Y_0|X,W}(y_0 | 1, w)\} p_{1|w} \\ &+ (1 - x) \min\{\overline{F}_{Y_1|X,W}(y_1 | 0, w), F_{Y|X,W}(y_0 | 0, w)\} p_{0|w}. \end{aligned} \tag{C.23}$$

By Lemma C.3.1.1,  $\underline{F}_{Y_0|X,W}(y_0 | 1, w)$  and  $\overline{F}_{Y_1|X,W}(y_1 | 0, w)$  are cdfs. Also note that  $(u, v) \mapsto \min\{u, v\}$  is the comonotonic copula. By Sklar's Theorem,  $\tilde{\mathbb{P}}$  is a joint distribution function for  $(Y_1, Y_0, X)$  conditional on  $W = w$ .

Following the same steps as in the proof of Theorem 4.4.1, it can be shown that conditions 1–4 are satisfied because the distribution in (C.23) is the same as in (C.14) but

for a specific rather than an arbitrary choice of copulas. By Lemma C.3.6,  $\tilde{\mathbb{P}}$  implies the following co-monotonic joint distribution of  $(Y_1, Y_0)$ :

$$\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) = \min \left\{ \overline{F}_{Y_1|W}(y_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w) \right\}. \quad (\text{C.24})$$

To show condition 5 holds, and thus complete the proof, we construct a function  $p^{ul}$  such that  $p^{ul}(Y_1, Y_0) = \tilde{\mathbb{E}}[X \mid Y_1, Y_0, W = w]$  and  $p^{ul}(Y_1, Y_0, w) \in [\underline{c}, \bar{c}]$  almost surely under  $\tilde{\mathbb{P}}$ .

First consider  $p_{1|w} = \underline{c}$ , then we have

$$\underline{F}_{Y_0|W}(y_0 \mid w) = \underline{F}_{Y_0|X,W}(y_0 \mid 1, w) = F_{Y|X,W}(y_0|0, w)$$

and

$$\overline{F}_{Y_1|W}(y_1 \mid w) = \overline{F}_{Y_1|X,W}(y_1 \mid 0, w) = F_{Y|X,W}(y_1 \mid 1, w).$$

This implies the following derivation

$$\begin{aligned} \tilde{\mathbb{E}}[\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) \underline{c} \mid W = w] &= p_{1|w} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) \\ &= p_{1|w} \min \left\{ F_{Y|X,W}(y_1 \mid 1, w), F_{Y|X,W}(y_0|0, w) \right\} \\ &= \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 \mid W = w). \end{aligned}$$

The first line holds by  $\underline{c} = p_{1|w}$ , the second line holds by (C.24), and the last line holds by (C.23). Following Lemma C.3.5, we conclude that  $\tilde{\mathbb{P}}(X = 1 \mid Y_1, Y_0, W = w) = \underline{c}$  almost surely, which is naturally bounded between  $\underline{c}$  and  $\bar{c}$ . The proof of the case where  $p_{1|w} = \bar{c}$  is similar and thus omitted.

Next consider  $\underline{c} < p_{1|w} < \bar{c}$ . Let  $p^{ul}(Y_1, Y_0, w) = p^{ul}(Y_1, Y_0, w; B^{ul})$  defined in (C.16), where

$$B^{ul} = \frac{\tilde{\mathbb{P}}(Y_1 = \overline{Q}_1, Y_0 = \underline{Q}_0, X = 1 \mid W = w)}{\tilde{\mathbb{P}}(Y_1 = \overline{Q}_1, Y_0 = \underline{Q}_0 \mid W = w)}$$

whenever  $\tilde{\mathbb{P}}(Y_1 = \overline{Q}_1, Y_0 = \underline{Q}_0 \mid W = w) > 0$ . Let  $B^{ul} = p_{1|w}$  otherwise.

We verify that  $B^{ul} \in [\underline{c}, \bar{c}]$  if the denominator is nonzero.

First, note that the denominator of  $B^{ul}$  can be expanded below

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0 \mid W = w) \\
&= \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq \underline{Q}_0 \mid W = w) - \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 < \underline{Q}_0 \mid W = w) \\
&\quad - \tilde{\mathbb{P}}(Y_1 < \bar{Q}_1, Y_0 \leq \underline{Q}_0 \mid W = w) + \tilde{\mathbb{P}}(Y_1 < \bar{Q}_1, Y_0 < \underline{Q}_0 \mid W = w) \\
&= \min\{\bar{F}_{Y_1|W}(\bar{Q}_1 \mid w), \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w)\} - \min\{\bar{F}_{Y_1|W}(\bar{Q}_1 \mid w), \underline{F}_{Y_0|W}(\underline{Q}_0 - |w)\} \\
&\quad - \min\{\bar{F}_{Y_1|W}(\bar{Q}_1 - |w), \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w)\} + \min\{\bar{F}_{Y_1|W}(\bar{Q}_1 - |w), \underline{F}_{Y_0|W}(\underline{Q}_0 - |w)\},
\end{aligned}$$

where the second equality holds via (C.24). By Lemma C.3.4.8, this expression simplifies to

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0 \mid W = w) \\
&= \min\{\bar{F}_{Y_1|W}(\bar{Q}_1 \mid w), \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w)\} - \underline{F}_{Y_0|W}(\underline{Q}_0 - |w) \\
&\quad - \bar{F}_{Y_1|W}(\bar{Q}_1 - |w) + \min\{\bar{F}_{Y_1|W}(\bar{Q}_1 - |w), \underline{F}_{Y_0|W}(\underline{Q}_0 - |w)\} \\
&= \min\{\bar{F}_{Y_1|W}(\bar{Q}_1 \mid w), \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w)\} - \max\{\bar{F}_{Y_1|W}(\bar{Q}_1 - |w), \underline{F}_{Y_0|W}(\underline{Q}_0 - |w)\}.
\end{aligned}$$

Second, we expand the numerator of  $B^{ul}$ . We have that

$$\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0, X = 1 \mid W = w) = \tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0 \mid X = 1, W = w)p_{1|w}$$

and that

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0 \mid X = 1, W = w) \\
&= \min\{F_{Y|X,W}(\bar{Q}_1 \mid 1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 \mid 1, w)\} - \min\{F_{Y|X,W}(\bar{Q}_1 \mid 1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 - |1, w)\} \\
&\quad - \min\{F_{Y|X,W}(\bar{Q}_1 - |1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 \mid 1, w)\} + \min\{F_{Y|X,W}(\bar{Q}_1 - |1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 - |1, w)\} \\
&= \min\{F_{Y|X,W}(\bar{Q}_1 \mid 1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 \mid 1, w)\} - \underline{F}_{Y_0|X,W}(\underline{Q}_0 - |1, w) \\
&\quad - F_{Y|X,W}(\bar{Q}_1 - |1, w) + \min\{F_{Y|X,W}(\bar{Q}_1 - |1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 - |1, w)\} \\
&= \min\{F_{Y|X,W}(\bar{Q}_1 \mid 1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 \mid 1, w)\} - \max\{F_{Y|X,W}(\bar{Q}_1 - |1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 - |1, w)\},
\end{aligned}$$

where the second to last equality follows from Lemma C.3.4.9.

From Part 6 and 7 of Lemma C.3.4, we observe that  $B^{ul}$  can take four possible values

as follows:

$$B^{ul} = \frac{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0 | X = 1, W = w)p_{1|w}}{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0 | W = w)}$$

$$= \frac{(F_{Y|X,W}(\bar{Q}_1 | 1, w) - F_{Y|X,W}(\bar{Q}_1 - | 1, w))p_{1|w}}{\bar{F}_{Y_1|W}(\bar{Q}_1 | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \mathbb{1} \left( \begin{array}{l} \bar{F}_{Y_1|W}(\bar{Q}_1 | w) \leq \underline{F}_{Y_0|W}(\underline{Q}_0 | w), \\ \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) > \underline{F}_{Y_0|W}(\underline{Q}_0 - | w) \end{array} \right) \quad (\text{C.25})$$

$$+ \frac{(F_{Y|X,W}(\bar{Q}_1 | 1, w) - \underline{F}_{Y_0|X,W}(\underline{Q}_0 - | 1, w))p_{1|w}}{\bar{F}_{Y_1|W}(\bar{Q}_1 | w) - \underline{F}_{Y_0|W}(\underline{Q}_0 - | w)} \mathbb{1} \left( \begin{array}{l} \bar{F}_{Y_1|W}(\bar{Q}_1 | w) \leq \underline{F}_{Y_0|W}(\underline{Q}_0 | w), \\ \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) \leq \underline{F}_{Y_0|W}(\underline{Q}_0 - | w) \end{array} \right) \quad (\text{C.26})$$

$$+ \frac{(\underline{F}_{Y_0|X,W}(\underline{Q}_0 | 1, w) - F_{Y|X,W}(\bar{Q}_1 - | 1, w))p_{1|w}}{\underline{F}_{Y_0|W}(\underline{Q}_0 | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \mathbb{1} \left( \begin{array}{l} \bar{F}_{Y_1|W}(\bar{Q}_1 | w) > \underline{F}_{Y_0|W}(\underline{Q}_0 | w), \\ \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) > \underline{F}_{Y_0|W}(\underline{Q}_0 - | w) \end{array} \right) \quad (\text{C.27})$$

$$+ \frac{(\underline{F}_{Y_0|X,W}(\underline{Q}_0 | 1, w) - \underline{F}_{Y_0|X,W}(\underline{Q}_0 - | 1, w))p_{1|w}}{\underline{F}_{Y_0|W}(\underline{Q}_0 | w) - \underline{F}_{Y_0|W}(\underline{Q}_0 - | w)} \mathbb{1} \left( \begin{array}{l} \bar{F}_{Y_1|W}(\bar{Q}_1 | w) > \underline{F}_{Y_0|W}(\underline{Q}_0 | w), \\ \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) \leq \underline{F}_{Y_0|W}(\underline{Q}_0 - | w) \end{array} \right). \quad (\text{C.28})$$

All the terms have positive denominators since we focus on the case where  $\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0 | W = w) > 0$ . As shown in Lemma C.3.4.4, terms (C.25) and (C.28) lie in  $[\underline{c}, \bar{c}]$ .

Next we examine the term (C.26), which can be written as follows

$$\begin{aligned} & \frac{(F_{Y|X,W}(\bar{Q}_1 | 1, w) - \underline{F}_{Y_0|X,W}(\underline{Q}_0 - | 1, w))p_{1|w}}{\bar{F}_{Y_1|W}(\bar{Q}_1 | w) - \underline{F}_{Y_0|W}(\underline{Q}_0 - | w)} \\ &= \frac{F_{Y|X,W}(\bar{Q}_1 | 1, w)p_{1|w} - F_{Y|X,W}(\underline{Q}_0 - | 0, w)\frac{p_{0|w}\underline{c}}{1-\underline{c}}}{\frac{\bar{c}-p_{1|w}}{\bar{c}} + F_{Y|X,W}(\bar{Q}_1 | 1, w)\frac{p_{1|w}}{\bar{c}} - F_{Y|X,W}(\underline{Q}_0 - | 0, w)\frac{p_{0|w}}{1-\underline{c}}} \\ &= \underline{c} + \frac{(\bar{c} - \underline{c})p_{1|w}}{\bar{c}} \frac{F_{Y|X,W}(\bar{Q}_1 | 1, w) - \bar{\tau}_1}{\frac{\bar{c}-p_{1|w}}{\bar{c}} + F_{Y|X,W}(\bar{Q}_1 | 1, w)\frac{p_{1|w}}{\bar{c}} - F_{Y|X,W}(\underline{Q}_0 - | 0, w)\frac{p_{0|w}}{1-\underline{c}}} \\ &= \underline{c} + \frac{(\bar{c} - \underline{c})p_{1|w}}{\bar{c}} \frac{F_{Y|X,W}(\bar{Q}_1 | 1, w) - \bar{\tau}_1}{\bar{F}_{Y_1|W}(\bar{Q}_1 | w) - \underline{F}_{Y_0|W}(\underline{Q}_0 - | w)} \\ &\geq \underline{c} \end{aligned}$$

where the last line follows by  $\bar{c} \geq \underline{c}$  and  $F_{Y|X,W}(\bar{Q}_1 | 1, w) \geq \bar{\tau}_1$  via Lemma C.3.3.2. Also

note that

$$\begin{aligned}
& \frac{(F_{Y|X,W}(\bar{Q}_1 | 1, w) - \underline{F}_{Y_0|X,W}(\underline{Q}_0 - | 1, w))p_{1|w}}{\bar{F}_{Y_1|W}(\bar{Q}_1 | w) - \underline{F}_{Y_0|W}(\underline{Q}_0 - | w)} \\
&= \bar{c} + \frac{p_{0|w}(\bar{c} - \underline{c})}{1 - \underline{c}} \frac{F_{Y|X,W}(\underline{Q}_0 - | 0, w) - \underline{\tau}_0}{\bar{F}_{Y_1|W}(\bar{Q}_1 | w) - \underline{F}_{Y_0|W}(\underline{Q}_0 - | w)} \\
&\leq \bar{c},
\end{aligned}$$

where the inequality follows by  $\bar{c} \geq \underline{c}$  and  $F_{Y|X,W}(\underline{Q}_0 - | 0, w) \leq \underline{\tau}_0$  via Lemma C.3.3.3.

Thus we have shown the term (C.26) is bounded within  $[\underline{c}, \bar{c}]$ .

Then consider the term (C.27). Following the same arguments, we have

$$\begin{aligned}
& \frac{(\underline{F}_{Y_0|X,W}(\underline{Q}_0 | 1, w) - F_{Y|X,W}(\bar{Q}_1 - | 1, w))p_{1|w}}{\underline{F}_{Y_0|W}(\underline{Q}_0 | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \\
&= \underline{c} + \frac{p_{0|w}(\bar{c} - \underline{c})}{1 - \bar{c}} \frac{F_{Y|X,W}(\underline{Q}_0 | 0, w) - \underline{\tau}_0}{\underline{F}_{Y_0|W}(\underline{Q}_0 | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \\
&\geq \underline{c},
\end{aligned}$$

and

$$\begin{aligned}
& \frac{(\underline{F}_{Y_0|X,W}(\underline{Q}_0 | 1, w) - F_{Y|X,W}(\bar{Q}_1 - | 1, w))p_{1|w}}{\underline{F}_{Y_0|W}(\underline{Q}_0 | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \\
&= \bar{c} + \frac{p_{1|w}(\bar{c} - \underline{c})}{\underline{c}} \frac{F_{Y|X,W}(\bar{Q}_1 - | 1, w) - \bar{\tau}_1}{\underline{F}_{Y_0|W}(\underline{Q}_0 | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \\
&\leq \bar{c}.
\end{aligned}$$

So we have shown that all the four terms (C.25)–(C.28) are bounded within  $[\underline{c}, \bar{c}]$ , thus concluding  $B^{ul} \in [\underline{c}, \bar{c}]$ , which then establishes  $p^{ul}(Y_1, Y_0, w) \in [\underline{c}, \bar{c}]$  almost surely.

To finish this proof, we demonstrate that  $\tilde{\mathbb{E}}[X | Y_1, Y_0, W = w] = p^{ul}(Y_1, Y_0, w)$  almost surely. To do so, we use Lemma C.3.5 and show that

$$\begin{aligned}
& \tilde{\mathbb{E}} \left[ \mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) p^{ul}(Y_1, Y_0, w) \mid W = w \right] \\
&= \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 \mid W = w) \\
&= p_{1|w} \min\{F_{Y|X,W}(y_1 | 1, w), \underline{F}_{Y_0|X,W}(y_0 | 1, w)\}
\end{aligned} \tag{C.29}$$

for all  $(y_1, y_0) \in \mathbb{R}^2$ . To complete the proof, we break this up into different cases.

**(Part 1) Case 1:**  $y_1 < \bar{Q}_1$  and  $y_0 < \underline{Q}_0$ .

In this case,  $p^{ul}(Y_1, Y_0, w) \mathbb{1}[Y_1 \leq y_1, Y_0 \leq y_0] = \underline{c} \mathbb{1}[Y_1 \leq y_1, Y_0 \leq y_0]$ . Thus we have

$$\begin{aligned}
& \tilde{\mathbb{E}} \left[ \mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) p^{ul}(Y_1, Y_0, w) \mid W = w \right] \\
&= \underline{c} \min \left\{ \bar{F}_{Y_1|W}(y_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w) \right\} \\
&= \underline{c} \min \left\{ F_{Y|X,W}(y_1 \mid 1, w) \frac{p_{1|w}}{\underline{c}}, F_{Y|X,W}(y_0 \mid 0, w) \frac{p_{0|w}}{1 - \underline{c}} \right\} \\
&= \min \left\{ F_{Y|X,W}(y_1 \mid 1, w) p_{1|w}, F_{Y|X,W}(y_0 \mid 0, w) \frac{p_{0|w} \underline{c}}{1 - \underline{c}} \right\} \\
&= p_{1|w} \min \left\{ F_{Y|X,W}(y_1 \mid 1, w), \underline{F}_{Y_0|X,W}(y_0 \mid 1, w) \right\}.
\end{aligned}$$

The second line holds by the assumption that  $y_1 < \bar{Q}_1$  and  $y_0 < \underline{Q}_0$ . Therefore, we have shown that (C.29) holds.

**(Part 1) Case 2:**  $y_1 \geq \bar{Q}_1$  and  $y_0 < \underline{Q}_0$ .

First, note that the joint cdf from (C.24) implies

$$\tilde{\mathbb{P}}(Y_1 > \bar{Q}_1, Y_0 < \underline{Q}_0 \mid W = w) = \tilde{\mathbb{P}}(Y_1 < \bar{Q}_1, Y_0 > \underline{Q}_0 \mid W = w) = 0. \quad (\text{C.30})$$

These equalities follow by

$$\begin{aligned}
\tilde{\mathbb{P}}(Y_1 > \bar{Q}_1, Y_0 < \underline{Q}_0 \mid W = w) &= \tilde{\mathbb{P}}(Y_0 < \underline{Q}_0 \mid W = w) - \tilde{\mathbb{P}}(Y_0 < \underline{Q}_0, Y_1 \leq \bar{Q}_1 \mid W = w) \\
&= \underline{F}_{Y_0|W}(\underline{Q}_0 - |w) - \min \left\{ \bar{F}_{Y_1|W}(\bar{Q}_1 \mid w), \underline{F}_{Y_0|W}(\underline{Q}_0 - |w) \right\} \\
&= \underline{F}_{Y_0|W}(\underline{Q}_0 - |w) - \underline{F}_{Y_0|W}(\underline{Q}_0 - |w) \\
&= 0,
\end{aligned}$$

where the third line holds by Lemma C.3.4.8. Similarly,

$$\begin{aligned}
\tilde{\mathbb{P}}(Y_1 < \bar{Q}_1, Y_0 > \underline{Q}_0 \mid W = w) &= \tilde{\mathbb{P}}(Y_1 < \bar{Q}_1 \mid W = w) - \tilde{\mathbb{P}}(Y_1 < \bar{Q}_1, Y_0 \leq \underline{Q}_0 \mid W = w) \\
&= \bar{F}_{Y_1|W}(\bar{Q}_1 - \mid w) - \min \left\{ \bar{F}_{Y_1|W}(\bar{Q}_1 - \mid w), \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w) \right\} \\
&= \bar{F}_{Y_1|W}(\bar{Q}_1 - \mid w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - \mid w) \\
&= 0.
\end{aligned}$$

Based on (C.30), we can decompose the left-hand-side term of (C.29) as below

$$\begin{aligned}
&\tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) p^{ul}(Y_1, Y_0, w) \mid W = w) \\
&= \underline{c} \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq y_0 \mid W = w) + p_{1|w} \cdot \tilde{\mathbb{P}}(\bar{Q}_1 < Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) \\
&= \underline{c} \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq y_0 \mid W = w) \\
&= \underline{c} \min \left\{ \bar{F}_{Y_1|W}(\bar{Q}_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w) \right\}.
\end{aligned}$$

Note that  $\bar{F}_{Y_1|W}(\bar{Q}_1 \mid w) \geq \underline{F}_{Y_0|W}(\underline{Q}_0 - \mid w) \geq \underline{F}_{Y_0|W}(y_0 \mid w)$  by Lemma C.3.4.8 and the condition that  $y_0 < \underline{Q}_0$ . We have

$$\underline{c} \min \left\{ \bar{F}_{Y_1|W}(\bar{Q}_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w) \right\} = \underline{c} \underline{F}_{Y_0|W}(y_0 \mid w) = p_{1|w} \underline{F}_{Y_0|X,W}(y_0 \mid 1, w).$$

By Lemma C.3.4.6 (the second set of equivalent results),  $\bar{F}_{Y_1|W}(\bar{Q}_1 \mid w) \geq \underline{F}_{Y_0|W}(y_0 \mid w)$  also implies

$$\underline{F}_{Y_0|X,W}(y_0 \mid 1, w) \leq F_{Y|X,W}(\bar{Q}_1 \mid 1, w) \leq F_{Y|X,W}(y_1 \mid 1, w),$$

hence we have

$$p_{1|w} \underline{F}_{Y_0|X,W}(y_0 \mid 1, w) = p_{1|w} \min \left\{ F_{Y|X,W}(y_1 \mid 1, w), \underline{F}_{Y_0|X,W}(y_0 \mid 1, w) \right\}.$$

Combining those results then yields (C.29), as desired.

**(Part 1) Case 3:**  $y_1 < \bar{Q}_1$  and  $y_0 \geq \underline{Q}_0$ .

Following similar arguments in Case 2, we have the following equality

$$\begin{aligned}
& \tilde{\mathbb{E}} \left( \mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) p^{ul}(Y_1, Y_0, w) \mid W = w \right) \\
&= \underline{c} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq \underline{Q}_0 \mid W = w) + p_{1|w} \cdot \tilde{\mathbb{P}}(Y_1 \leq y_1, \underline{Q}_0 < Y_0 \leq y_0 \mid W = w) \\
&= \underline{c} \min \left\{ \overline{F}_{Y_1|W}(y_1 \mid w), \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w) \right\} \\
&= \underline{c} \overline{F}_{Y_1|W}(y_1 \mid w) \\
&= p_{1|w} F_{Y|X,W}(y_1 \mid 1, w) \\
&= p_{1|w} \min \left\{ F_{Y|X,W}(y_1 \mid 1, w), \underline{F}_{Y_0|X,W}(y_0 \mid 1, w) \right\}
\end{aligned}$$

where we use (C.30) in the second equality, the third equality follows by Lemma C.3.4.8, and the condition that  $y_1 < \overline{Q}_1$ , and the last line holds by Lemma C.3.4.9, where we deduce that

$$F_{Y|X,W}(y_1 \mid 1, w) \leq F_{Y|X,W}(\overline{Q}_1 - \mid 1, w) \leq \underline{F}_{Y_0|X,W}(\underline{Q}_0 \mid 1, w) \leq \underline{F}_{Y_0|X,W}(y_0 \mid 1, w).$$

Therefore, we established (C.29).

**(Part 1) Case 4:**  $y_1 = \overline{Q}_1$  and  $y_0 = \underline{Q}_0$ .

We can decompose the LHS probability of (C.29) as follows:

$$\begin{aligned}
& \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0)p^{ul}(Y_1, Y_0, w) \mid W = w) \\
&= \tilde{\mathbb{E}}(\mathbb{1}(Y_1 = y_1, Y_0 = y_0)p^{ul}(Y_1, Y_0) \mid W = w) + \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 < y_0)p^{ul}(Y_1, Y_0, w) \mid W = w) \\
&\quad + \tilde{\mathbb{E}}(\mathbb{1}(Y_1 < y_1, Y_0 \leq y_0)p^{ul}(Y_1, Y_0, w) \mid W = w) - \tilde{\mathbb{E}}(\mathbb{1}(Y_1 < y_1, Y_0 < y_0)p^{ul}(Y_1, Y_0, w) \mid W = w) \\
&= B^{ul}\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0 \mid W = w) \\
&\quad + \lim_{u \nearrow \underline{Q}_0} \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq \bar{Q}_1, Y_0 \leq u \mid w = w)p^{ul}(Y_1, Y_0, w) \mid W = w) \\
&\quad + \lim_{v \nearrow \bar{Q}_1} \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq v, Y_0 \leq \underline{Q}_0)p^{ul}(Y_1, Y_0, w) \mid W = w) \\
&\quad - \lim_{v \nearrow \bar{Q}_1, u \nearrow \underline{Q}_0} \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq v, Y_0 \leq u)p^{ul}(Y_1, Y_0, w) \mid W = w) \\
&= \tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \underline{Q}_0, X = 1 \mid W = w) + \lim_{u \nearrow \underline{Q}_0} \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq u, X = 1 \mid W = w) \\
&\quad + \lim_{v \nearrow \bar{Q}_1} \tilde{\mathbb{P}}(Y_1 \leq v, Y_0 \leq \underline{Q}_0, X = 1 \mid W = w) - \lim_{v \nearrow \bar{Q}_1, u \nearrow \underline{Q}_0} \tilde{\mathbb{P}}(Y_1 \leq v, Y_0 \leq u, X = 1 \mid W = w) \\
&= \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq \underline{Q}_0, X = 1 \mid W = w).
\end{aligned}$$

The second equality holds by the monotone convergence theorem. The third equality holds by the conclusion proved in Case 1–3. The last equality holds by the continuity of probability measure. Thus (C.29) has been verified.

**(Part 1) Case 5:**  $(y_1, y_0) \geq (\bar{Q}_1, \underline{Q}_0)$ .

Given the above results, we have the following derivation:

$$\begin{aligned}
& \tilde{\mathbb{E}} \left( \mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) p^{ul}(Y_1, Y_0, w) \mid W = w \right) \\
&= \tilde{\mathbb{E}} \left( \mathbb{1}(Y_1 \leq \bar{Q}_1, Y_0 \leq \underline{Q}_0) p^{ul}(Y_1, Y_0, w) \mid W = w \right) \\
&\quad + \tilde{\mathbb{E}} \left( \bar{c} \left[ \mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) - \mathbb{1}(Y_1 \leq \bar{Q}_1, Y_0 \leq \underline{Q}_0) \right] \mid W = w \right) \\
&= \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq \underline{Q}_0, X = 1 \mid W = w) \\
&\quad + \bar{c} \left( \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) - \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq \underline{Q}_0 \mid W = w) \right) \\
&= \bar{c} \min \left\{ \bar{F}_{Y_1|W}(y_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w) \right\} + p_{1|w} \min \left\{ F_{Y|X,W}(\bar{Q}_1 \mid 1, w), \underline{F}_{Y_0|X,W}(\underline{Q}_0 \mid 1, w) \right\} \\
&\quad - \bar{c} \min \left\{ \bar{F}_{Y_1|W}(\bar{Q}_1 \mid w), \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w) \right\} \\
&= \bar{c} \min \left\{ \bar{F}_{Y_1|W}(y_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w) \right\} \\
&\quad + \left[ F_{Y|X,W}(\bar{Q}_1 \mid 1, w) p_{1|w} - \bar{F}_{Y_1|W}(\bar{Q}_1 \mid w) \bar{c} \right] \mathbb{1}(\bar{F}_{Y_1|W}(\bar{Q}_1 \mid w) \leq \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w)) \\
&\quad + \left[ \underline{F}_{Y_0|X,W}(\underline{Q}_0 \mid 1, w) p_{1|w} - \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w) \bar{c} \right] \mathbb{1}(\bar{F}_{Y_1|W}(\bar{Q}_1 \mid w) > \underline{F}_{Y_0|W}(\underline{Q}_0 \mid w)) \\
&= \bar{c} \min \left\{ \bar{F}_{Y_1|W}(y_1 \mid w), \underline{F}_{Y_0|W}(y_0 \mid w) \right\} - (\bar{c} - p_{1|w}) \\
&= \min \left\{ F_{Y|X,W}(y_1 \mid 1, w) p_{1|w}, \frac{p_{1|w} - \bar{c}}{1 - \bar{c}} + F_{Y|X,W}(y_0 \mid 0, w) \frac{p_{0|w} \bar{c}}{1 - \bar{c}} \right\} \\
&= p_{1|w} \min \left\{ F_{Y|X,W}(y_1 \mid 1, w), \underline{F}_{Y_0|X,W}(y_0 \mid 0, w) \right\},
\end{aligned}$$

where the first equality follows by (C.30) that  $(Y_1, Y_0)$  has no mass on the off-diagonal area, the second equality follows by the result established in Case 4 above, and the fourth equality follows by Lemma C.3.4.6. Thus we have verified (C.29).

Since  $\mathbb{R}^2$  is partitioned by these 5 cases, we have established that  $\tilde{\mathbb{E}}[X \mid Y_1, Y_0, W = w] = p^{ul}(Y_1, Y_0, w)$  almost surely, which concludes the proof of Part 1.

**Proof of Part 2:** We prove this by constructing a probability distribution  $\tilde{\mathbb{P}}$  for  $(Y_1, Y_0, X)$  conditional on  $W = w$  such that for all  $y \in \mathbb{R}$  and  $x \in \{0, 1\}$ , we have

$$1. \tilde{\mathbb{P}}(Y_1 \leq y \mid W = w) = \bar{F}_{Y_1|W}(y \mid w) \text{ and } \tilde{\mathbb{P}}(Y_0 \leq y \mid W = w) = \bar{F}_{Y_0|W}(y \mid w);$$

2.  $\tilde{\mathbb{P}}(X = x \mid W = w) = p_{x|w}$ ;
3.  $\tilde{\mathbb{P}}(Y_x \leq y \mid X = x, W = w) = F_{Y|X,W}(Y \mid X, w)$ ;
4. The following equality holds:

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid X = x, W = w) \\ &= \max \left\{ \tilde{\mathbb{P}}(Y_1 \leq y_1 \mid X = x, W = w) + \tilde{\mathbb{P}}(Y_0 \leq y_0 \mid X = x, W = w) - 1, 0 \right\}; \end{aligned}$$

5.  $\tilde{\mathbb{P}}(X = 1 \mid Y_1, Y_0, W = w) \in [\underline{c}, \bar{c}]$  for  $\tilde{\mathbb{P}}$ -almost surely.

Let

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x \mid W = w) \\ &= x \max\{F_{Y|X,W}(y_1 \mid 1, w) + \bar{F}_{Y_0|X,W}(y_0 \mid 1, w) - 1, 0\}p_{1|w} \\ &+ (1 - x) \max\{\bar{F}_{Y_1|X,W}(y_1 \mid 0, w) + F_{Y|X,W}(y_0|0, w) - 1, 0\}p_{0|w}. \end{aligned} \tag{C.31}$$

By Lemma C.3.1.1,  $\bar{F}_{Y_0|X,W}(y_0 \mid 1, w)$  and  $\bar{F}_{Y_1|X,W}(y_1 \mid 0, w)$  are cdfs. Also note that  $(u, v) \mapsto \max\{u + v - 1, 0\}$  is the counter-monotonic copula. Following Sklar's Theorem,  $\tilde{\mathbb{P}}$  is a joint distribution function for  $(Y_1, Y_0, X)$  conditional on  $W = w$ .

Following the same steps as in the proof of Theorem 4.4.1, we can show that conditions 1–4 are satisfied because the distribution in (C.31) is the same as in (C.14) but for a specific rather than an arbitrary choice of copulas. By Lemma C.3.7,  $\tilde{\mathbb{P}}$  leads to the following counter-monotonic joint distribution of  $(Y_1, Y_0)$ :

$$\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) = \max \left\{ \bar{F}_{Y_1|W}(y_1 \mid w) + \bar{F}_{Y_0|W}(y_0 \mid w) - 1, 0 \right\}. \tag{C.32}$$

To show condition 5 holds, and thus complete the proof, we must find a function  $p^{uu}$  such that  $p^{uu}(Y_1, Y_0, w) = \tilde{\mathbb{E}}[X \mid Y_1, Y_0, W = w]$  and  $p^{uu}(Y_1, Y_0, w) \in [\underline{c}, \bar{c}]$  almost surely under  $\tilde{\mathbb{P}}$ .

First consider  $p_{1|w} = \underline{c}$ , then we have

$$\bar{F}_{Y_1|X,W}(y_1 \mid 0, w) = \bar{F}_{Y_1|W}(y_1 \mid w) = F_{Y|X,W}(y_1 \mid 1, w)$$

and

$$\bar{F}_{Y_0|X,W}(y_0 \mid 1, w) = \bar{F}_{Y_0|W}(y_0 \mid w) = F_{Y|X,W}(y_0|0, w).$$

This implies the following derivations:

$$\begin{aligned}
\tilde{\mathbb{E}}[\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) \underline{c} | W = w] &= p_{1|w} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 | W = w) \\
&= p_{1|w} \max \{ F_{Y|X,W}(y_1 | 1, w) + F_{Y|X,W}(y_0 | 0, w) - 1, 0 \} \\
&= \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 | W = w).
\end{aligned}$$

The first line holds by  $\underline{c} = p_{1|w}$ , the second line holds by (C.32), and the last line holds by (C.31). Following Lemma C.3.5, we conclude that  $\tilde{\mathbb{P}}(X = 1 | Y_1, Y_0, W = w) = \underline{c} \in [\underline{c}, \bar{c}]$  almost surely. The proof of the case where  $p_{1|w} = \bar{c}$  is similar and thus omitted.

Next consider  $\underline{c} < p_{1|w} < \bar{c}$ . Let  $p^{uu}(Y_1, Y_0, w) = p^{uu}(Y_1, Y_0, w; B^{uu})$  defined in (C.17), where

$$B^{uu} = \frac{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0, X = 1 | W = w)}{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 | W = w)}$$

whenever  $\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 | W = w) > 0$ . Set  $B^{uu} = p_{1|w}$  otherwise.

We verify that  $B^{uu} \in [\underline{c}, \bar{c}]$  if the denominator is nonzero.

First, note that the denominator of  $B^{uu}$  can be expanded below

$$\begin{aligned}
&\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 | W = w) \\
&= \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq \bar{Q}_0 | W = w) - \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 < \bar{Q}_0 | W = w) \\
&\quad - \tilde{\mathbb{P}}(Y_1 < \bar{Q}_1, Y_0 \leq \bar{Q}_0 | W = w) + \tilde{\mathbb{P}}(Y_1 < \bar{Q}_1, Y_0 < \bar{Q}_0 | W = w) \\
&= \max\{\bar{F}_{Y_1|W}(\bar{Q}_1 | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 | w) - 1, 0\} - \max\{\bar{F}_{Y_1|W}(\bar{Q}_1 | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 - |w) - 1, 0\} \\
&\quad - \max\{\bar{F}_{Y_1|W}(\bar{Q}_1 - |w) + \bar{F}_{Y_0|W}(\bar{Q}_0 | w) - 1, 0\} + \max\{\bar{F}_{Y_1|W}(\bar{Q}_1 - |w) + \bar{F}_{Y_0|W}(\bar{Q}_0 - |w) - 1, 0\}.
\end{aligned}$$

where the second equality holds via (C.32). By Lemma C.3.4.8, we observe that

$$\begin{aligned}
\bar{F}_{Y_1|W}(\bar{Q}_1 | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 | w) - 1 &\geq \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} + \frac{p_{1|w} - \underline{c}}{\bar{c} - \underline{c}} - 1 = 0 \\
\bar{F}_{Y_1|W}(\bar{Q}_1 - |w) + \bar{F}_{Y_0|W}(\bar{Q}_0 - |w) - 1 &\leq \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} + \frac{p_{1|w} - \underline{c}}{\bar{c} - \underline{c}} - 1 = 0,
\end{aligned}$$

hence this expression simplifies to

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 \mid W = w) \\
&= \bar{F}_{Y_1|W}(\bar{Q}_1 \mid w) + \bar{F}_{Y_0|W}(\bar{Q}_0 \mid w) - 1 \\
&\quad - \max\{\bar{F}_{Y_1|W}(\bar{Q}_1 \mid w) + \bar{F}_{Y_0|W}(\bar{Q}_0 - |w) - 1, 0\} - \max\{\bar{F}_{Y_1|W}(\bar{Q}_1 - |w) + \bar{F}_{Y_0|W}(\bar{Q}_0 \mid w) - 1, 0\} \\
&= \min\{1 - \bar{F}_{Y_0|W}(\bar{Q}_0 - |w), \bar{F}_{Y_1|W}(\bar{Q}_1 \mid w)\} + \min\{1 - \bar{F}_{Y_1|W}(\bar{Q}_1 - |w), \bar{F}_{Y_0|W}(\bar{Q}_0 \mid w)\} - 1.
\end{aligned}$$

Second, we expand the numerator of  $B^{uu}$ . We have that

$$\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0, X = 1 \mid W = w) = \tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 \mid X = 1, W = w)p_{1|w}$$

and that

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 \mid X = 1, W = w) \\
&= \max\{F_{Y|X,W}(\bar{Q}_1 \mid 1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 \mid 1, w) - 1, 0\} \\
&\quad - \max\{F_{Y|X,W}(\bar{Q}_1 \mid 1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 - |1, w) - 1, 0\} \\
&\quad - \max\{F_{Y|X,W}(\bar{Q}_1 - |1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 \mid 1, w) - 1, 0\} \\
&\quad + \max\{F_{Y|X,W}(\bar{Q}_1 - |1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 - |1, w) - 1, 0\} \\
&= F_{Y|X,W}(\bar{Q}_1 \mid 1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 \mid 1, w) - 1 \\
&\quad - \max\{F_{Y|X,W}(\bar{Q}_1 \mid 1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 - |1, w) - 1, 0\} \\
&\quad - \max\{F_{Y|X,W}(\bar{Q}_1 - |1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 \mid 1, w) - 1, 0\} + 0 \\
&= \min\{1 - \bar{F}_{Y_0|X,W}(\bar{Q}_0 - |1, w), F_{Y|X,W}(\bar{Q}_1 \mid 1, w)\} \\
&\quad + \min\{1 - F_{Y|X,W}(\bar{Q}_1 - |1, w), \bar{F}_{Y_0|X,W}(\bar{Q}_0 \mid 1, w)\} - 1,
\end{aligned}$$

where the second to last equality follows from Lemma C.3.4.9, where we note that

$$\bar{\tau}_1 + \underline{\tau}_1 = 1.$$

From Lemma C.3.4.11,  $B^{uu}$  can take four possible values as follows:

$$\begin{aligned}
B^{uu} &= \frac{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 \mid X = 1, W = w)p_{1|w}}{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 \mid W = w)} \\
&= \frac{(1 - \bar{F}_{Y_0|X,W}(\bar{Q}_0 - | 1, w) - F_{Y|X,W}(\bar{Q}_1 - | 1, w)) p_{1|w}}{1 - \bar{F}_{Y_0|W}(\bar{Q}_0 - | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \\
&\quad \times \mathbb{1} \left( \begin{array}{l} \bar{F}_{Y_1|W}(\bar{Q}_1 | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 - | w) \geq 1, \\ \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 | w) \geq 1 \end{array} \right) \tag{C.33}
\end{aligned}$$

$$\begin{aligned}
&+ \frac{(F_{Y|X,W}(\bar{Q}_1 | 1, w) - F_{Y|X,W}(\bar{Q}_1 - | 1, w)) p_{1|w}}{\bar{F}_{Y_1|W}(\bar{Q}_1 | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \\
&\quad \times \mathbb{1} \left( \begin{array}{l} \bar{F}_{Y_1|W}(\bar{Q}_1 | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 - | w) < 1, \\ \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 | w) \geq 1 \end{array} \right) \tag{C.34}
\end{aligned}$$

$$\begin{aligned}
&+ \frac{(\bar{F}_{Y_0|X,W}(\bar{Q}_0 | 1, w) - \bar{F}_{Y_0|X,W}(\bar{Q}_0 - | 1, w)) p_{1|w}}{\bar{F}_{Y_0|W}(\bar{Q}_0 | w) - \bar{F}_{Y_0|W}(\bar{Q}_0 - | w)} \\
&\quad \times \mathbb{1} \left( \begin{array}{l} \bar{F}_{Y_1|W}(\bar{Q}_1 | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 - | w) \geq 1, \\ \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 | w) < 1 \end{array} \right) \tag{C.35}
\end{aligned}$$

$$\begin{aligned}
&+ \frac{(\bar{F}_{Y_0|X,W}(\bar{Q}_0 | 1, w) + F_{Y|X,W}(\bar{Q}_1 | 1, w) - 1) p_{1|w}}{\bar{F}_{Y_0|W}(\bar{Q}_0 | w) + \bar{F}_{Y_1|W}(\bar{Q}_1 | w) - 1} \\
&\quad \times \mathbb{1} \left( \begin{array}{l} \bar{F}_{Y_1|W}(\bar{Q}_1 | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 - | w) < 1, \\ \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 | w) < 1 \end{array} \right) \tag{C.36}
\end{aligned}$$

As shown in Lemma C.3.4.3, terms (C.34) and (C.35) lie in  $[\underline{c}, \bar{c}]$ .

Next we examine the term (C.33), which can be written as below

$$\begin{aligned}
& \frac{(1 - \bar{F}_{Y_0|X,W}(\bar{Q}_0 - | 1, w) - F_{Y|X,W}(\bar{Q}_1 - | 1, w)) p_{1|w}}{1 - \bar{F}_{Y_0|W}(\bar{Q}_0 - | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \\
&= \frac{p_{1|w} \left( 1 - F_{Y|X,W}(\bar{Q}_0 - | 0, w) \frac{p_{0|w} \bar{c}}{p_{1|w} (1 - \bar{c})} - F_{Y|X,W}(\bar{Q}_1 - | 1, w) \right)}{1 - F_{Y|X,W}(\bar{Q}_0 - | 0, w) \frac{p_{0|w}}{1 - \bar{c}} - F_{Y|X,W}(\bar{Q}_1 - | 1, w) \frac{p_{1|w}}{\underline{c}}} \\
&= \underline{c} + \frac{p_{1|w} - \underline{c} - F_{Y|X,W}(\bar{Q}_0 - | 0, w) \frac{p_{0|w} (\bar{c} - \underline{c})}{1 - \bar{c}}}{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 | W = w)} \\
&\geq \underline{c} + \frac{p_{1|w} - \underline{c} - \bar{\tau}_0 \frac{p_{0|w} (\bar{c} - \underline{c})}{1 - \bar{c}}}{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 | W = w)} \\
&= \underline{c},
\end{aligned}$$

where the inequality follows by Lemma C.3.3.3 that

$$F_{Y|X,W}(\bar{Q}_0 - | 0, w) = F_{Y|X,W}(Q_{Y|X,W}(\bar{\tau}_0 | 0, w) - | 0, w) \leq \bar{\tau}_0.$$

Also note that

$$\begin{aligned}
& \frac{(1 - \bar{F}_{Y_0|X,W}(\bar{Q}_0 - | 1, w) - F_{Y|X,W}(\bar{Q}_1 - | 1, w)) p_{1|w}}{1 - \bar{F}_{Y_0|W}(\bar{Q}_0 - | w) - \bar{F}_{Y_1|W}(\bar{Q}_1 - | w)} \\
&= \bar{c} + \frac{(p_{1|w} - \bar{c}) + F_{Y|X,W}(\bar{Q}_1 - | 1, w) \frac{p_{1|w} (\bar{c} - \underline{c})}{\underline{c}}}{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 | W = w)} \\
&\leq \bar{c} + \frac{(p_{1|w} - \bar{c}) + \bar{\tau}_1 \frac{p_{1|w} (\bar{c} - \underline{c})}{\underline{c}}}{\tilde{\mathbb{P}}(Y_1 = \bar{Q}_1, Y_0 = \bar{Q}_0 | W = w)} \\
&= \bar{c},
\end{aligned}$$

where the inequality follows by Lemma C.3.3.3 that

$$F_{Y|X,W}(\bar{Q}_1 - | 1, w) = F_{Y|X,W}(Q_{Y|X,W}(\bar{\tau}_1 | 1, w) - | 1, w) \leq \bar{\tau}_1.$$

Then we have shown that the term (C.33) is bounded between  $\underline{c}$  and  $\bar{c}$ .

Then consider the term (C.36). Following the same arguments, we have

$$\begin{aligned}
& \frac{(\overline{F}_{Y_0|X,W}(\overline{Q}_0 | 1, w) + F_{Y|X,W}(\overline{Q}_1 | 1, w) - 1) p_{1|w}}{\overline{F}_{Y_0|W}(\overline{Q}_0|w) + \overline{F}_{Y_1|W}(\overline{Q}_1 | w) - 1} \\
&= \underline{c} + \frac{\frac{p_{1|w}(\overline{c}-\underline{c})}{\underline{c}} F_{Y|X,W}(\overline{Q}_1 | 1, w) - \frac{(\overline{c}-p_{1|w})\underline{c}}{\underline{c}}}{\tilde{\mathbb{P}}(Y_1 = \overline{Q}_1, Y_0 = \overline{Q}_0 | W = w)} \\
&\geq \underline{c}
\end{aligned}$$

and

$$\begin{aligned}
& \frac{(\overline{F}_{Y_0|X,W}(\overline{Q}_0 | 1, w) + F_{Y|X,W}(\overline{Q}_1 | 1, w) - 1) p_{1|w}}{\overline{F}_{Y_0|W}(\overline{Q}_0|w) + \overline{F}_{Y_1|W}(\overline{Q}_1 | w) - 1} \\
&= \overline{c} + \frac{\frac{(p_{1|w}-\underline{c})(1-\overline{c})}{1-\underline{c}} - F_{Y|X,W}(\overline{Q}_0|0, w) \frac{p_{0|w}(\overline{c}-\underline{c})}{1-\underline{c}}}{\tilde{\mathbb{P}}(Y_1 = \overline{Q}_1, Y_0 = \overline{Q}_0 | W = w)} \\
&\leq \overline{c},
\end{aligned}$$

where inequalities follow by Lemma C.3.3.2 that  $F_{Y|X,W}(\overline{Q}_x|x, w) \geq \overline{\tau}_x$  for  $x = 0, 1$ . So we have shown that all four terms (C.33)–(C.36) are bounded within  $[\underline{c}, \overline{c}]$ , thus concluding  $B^{uu} \in [\underline{c}, \overline{c}]$ , which then establishes  $p^{uu}(Y_1, Y_0, w) \in [\underline{c}, \overline{c}]$  almost surely.

To finish this proof, we demonstrate that  $\tilde{\mathbb{E}}[X | Y_1, Y_0, W = w] = p^{uu}(Y_1, Y_0, w)$  almost surely. To do so, we use Lemma C.3.5 and show that

$$\begin{aligned}
& \tilde{\mathbb{E}}[\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) p^{uu}(Y_1, Y_0, w) | W = w] \\
&= \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 | W = w) \\
&= p_{1|w} \max\{F_{Y|X,W}(y_1 | 1, w) + \overline{F}_{Y_0|X,W}(y_0 | 1, w) - 1, 0\}
\end{aligned} \tag{C.37}$$

for all  $(y_1, y_0) \in \mathbb{R}^2$ . To complete the proof, we break this up into different cases.

**(Part 2) Case 1:**  $y_1 < \overline{Q}_1$  and  $y_0 < \overline{Q}_0$ .

First, note that the joint cdf from (C.32) implies

$$\tilde{\mathbb{P}}(Y_1 < \overline{Q}_1, Y_0 < \overline{Q}_0 | W = w) = \tilde{\mathbb{P}}(\overline{Q}_1 < Y_1, \overline{Q}_0 < Y_0 | W = w) = 0. \tag{C.38}$$

These equalities can be verified by the arguments below:

$$\begin{aligned}
\tilde{\mathbb{P}}(Y_1 < \bar{Q}_1, Y_0 < \bar{Q}_0 \mid W = w) &= \max \{ \bar{F}_{Y_1|W}(\bar{Q}_1 - | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 - | w) - 1, 0 \} \\
&\leq \max \left\{ \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} + \frac{p_{1|w} - \underline{c}}{\bar{c} - \underline{c}} - 1, 0 \right\} \\
&= 0,
\end{aligned}$$

where the inequality follows by Lemma C.3.4.8, and similarly,

$$\begin{aligned}
\tilde{\mathbb{P}}(\bar{Q}_1 < Y_1, \bar{Q}_0 < Y_0 \mid W = w) &= 1 - \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1 \mid W = w) - \tilde{\mathbb{P}}(Y_0 \leq \bar{Q}_0 \mid W = w) \\
&\quad + \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq \bar{Q}_0 \mid W = w) \\
&= 1 - \bar{F}_{Y_1|W}(\bar{Q}_1 | w) - \bar{F}_{Y_0|W}(\bar{Q}_0 | w) \\
&\quad + \max \{ \bar{F}_{Y_1|W}(\bar{Q}_1 | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 | w) - 1, 0 \} \\
&= 1 - \min \{ 1, \bar{F}_{Y_1|W}(\bar{Q}_1 | w) + \bar{F}_{Y_0|W}(\bar{Q}_0 | w) \} \\
&\leq 1 - \min \left\{ 1, \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} + \frac{p_{1|w} - \underline{c}}{\bar{c} - \underline{c}} \right\} \\
&= 0,
\end{aligned}$$

where the inequality follows by Lemma C.3.4.8.

On the one hand, (C.38) implies

$$\begin{aligned}
\tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) p^{uu}(Y_1, Y_0, w) \mid W = w) &= p_{1|w} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid W = w) \\
&\leq p_{1|w} \tilde{\mathbb{P}}(Y_1 < \bar{Q}_1, Y_0 < \bar{Q}_0 \mid W = w) \\
&= 0.
\end{aligned}$$

This shows  $\tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) p^{uu}(Y_1, Y_0, w) \mid W = w) = 0$  due to the construction that  $p^{uu}(Y_1, Y_0, w)$  is non-negative. On the other hand,

$$\begin{aligned}
&\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 \mid W = w) \\
&= p_{1|w} \max \{ F_{Y|X,W}(y_1 | 1, w) + \bar{F}_{Y_0|X,W}(y_0 | 1, w) - 1, 0 \} \\
&\leq p_{1|w} \max \{ F_{Y|X,W}(\bar{Q}_1 - | 1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 - | 1, w) - 1, 0 \} \\
&\leq p_{1|w} \max \{ \bar{\tau}_1 + \underline{\tau}_1 - 1, 0 \} \\
&= 0,
\end{aligned}$$

where the second inequality follows by Lemma C.3.4.9. This implies  $\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 | W = w) = 0$ , thus establishing (C.37), as desired.

**(Part 2) Case 2:**  $y_1 \geq \bar{Q}_1, y_0 < \bar{Q}_0$ .

First, note that (C.38) implies that

$$\begin{aligned} & \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0)p^{uu}(Y_1, Y_0, w) | W = w) \\ &= \tilde{\mathbb{E}}(\mathbb{1}(\bar{Q}_1 \leq Y_1 \leq y_1, Y_0 \leq y_0)\bar{c}|W = w) + \tilde{\mathbb{E}}(\mathbb{1}(Y_1 < \bar{Q}_1, Y_0 \leq y_0)p_{1|w}|W = w) \\ &= \tilde{\mathbb{E}}(\mathbb{1}(\bar{Q}_1 \leq Y_1 \leq y_1, Y_0 \leq y_0)\bar{c}|W = w) + \tilde{\mathbb{E}}(\mathbb{1}(Y_1 < \bar{Q}_1, Y_0 \leq y_0)\bar{c}|W = w) \\ &= \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0)\bar{c}|W = w), \end{aligned}$$

where the second equality follows by the fact that  $\tilde{\mathbb{P}}$  takes no mass on  $\{Y_1 < \bar{Q}_1, Y_0 \leq y_0\} \subseteq \{Y_1 < \bar{Q}_1, Y_0 < \bar{Q}_0\}$  by (C.38). Next we expand the last expression

$$\begin{aligned} & \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0)\bar{c}|W = w) \\ &= \bar{c}\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 | W = w) \\ &= \bar{c} \max \{ \bar{F}_{Y_1|W}(y_1 | w) + \bar{F}_{Y_0|W}(y_0 | w) - 1, 0 \} \\ &= \bar{c} \max \left\{ \frac{\bar{c} - p_{1|w}}{\bar{c}} + F_{Y|X,W}(y_1 | 1, w) \frac{p_{1|w}}{\bar{c}} + F_{Y|X,W}(y_0|0, w) \frac{p_{0|w}}{1 - \bar{c}} - 1, 0 \right\} \\ &= p_{1|w} \max \left\{ F_{Y|X,W}(y_1 | 1, w) + \frac{\bar{c} - p_{1|w}}{p_{1|w}} + F_{Y|X,W}(y_0|0, w) \frac{p_{0|w}\bar{c}}{p_{1|w}(1 - \bar{c})} - \frac{\bar{c}}{p_{1|w}}, 0 \right\} \\ &= p_{1|w} \max \left\{ F_{Y|X,W}(y_1 | 1, w) + F_{Y|X,W}(y_0|0, w) \frac{p_{0|w}\bar{c}}{p_{1|w}(1 - \bar{c})} - 1, 0 \right\} \\ &= p_{1|w} \max \{ F_{Y|X,W}(y_1 | 1, w) + \bar{F}_{Y_0|X,W}(y_0 | 1, w) - 1, 0 \}, \end{aligned}$$

thus establishing (C.37), as desired.

**(Part 2) Case 3:**  $y_1 < \bar{Q}_1, y_0 \geq \bar{Q}_0$ .

Similar to the proof of case 2 above, we have

$$\tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0)p^{uu}(Y_1, Y_0, w) | W = w) = \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0)\underline{c}|W = w)$$

due to (C.38). Next we expand the expression on the right hand side.

$$\begin{aligned}
& \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) \underline{c} | W = w) \\
&= \underline{c} \max \left\{ \overline{F}_{Y_1|W}(y_1 | w) + \overline{F}_{Y_0|W}(y_0 | w) - 1, 0 \right\} \\
&= \underline{c} \max \left\{ F_{Y|X,W}(y_1 | 1, w) \frac{p_{1|w}}{\underline{c}} + \frac{p_{1|w} - \underline{c}}{1 - \underline{c}} + F_{Y|X,W}(y_0|0, w) \frac{p_{0|w}}{1 - \underline{c}} - 1, 0 \right\} \\
&= p_{1|w} \max \left\{ F_{Y|X,W}(y_1 | 1, w) + \frac{(p_{1|w} - \underline{c})\underline{c}}{p_{1|w}(1 - \underline{c})} + F_{Y|X,W}(y_0|0, w) \frac{p_{0|w}\underline{c}}{p_{1|w}(1 - \underline{c})} - \frac{\underline{c}}{p_{1|w}}, 0 \right\} \\
&= p_{1|w} \max \left\{ F_{Y|X,W}(y_1 | 1, w) + \frac{p_{1|w} - \underline{c}}{p_{1|w}(1 - \underline{c})} + F_{Y|X,W}(y_0|0, w) \frac{p_{0|w}\underline{c}}{p_{1|w}(1 - \underline{c})} - 1, 0 \right\} \\
&= p_{1|w} \max \left\{ F_{Y|X,W}(y_1 | 1, w) + \overline{F}_{Y_0|X,W}(y_0 | 1, w) - 1, 0 \right\},
\end{aligned}$$

thus establishing (C.37), as desired.

**(Part 2) Case 4:**  $y_1 = \overline{Q}_1$ ,  $y_0 = \overline{Q}_0$ .

Note that the equality (C.37) can be established following the same arguments from the proof of Part 1, case 4. Once the results are established for cases 1–3, the equality (C.37) holds for  $y_1 = \overline{Q}_1$ ,  $y_0 = \overline{Q}_0$  by applying monotone convergence theorem and continuity of measure. To this end, the proof is omitted.

**(Part 2) Case 5:**  $(y_1, y_0) \geq (\overline{Q}_1, \overline{Q}_0)$ .

We start by noting that

$$\begin{aligned}
& \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq y_1, Y_0 \leq y_0) p^{uu}(Y_1, Y_0, w) | W = w) \\
&= p_{1|w} \tilde{\mathbb{P}}(Y_1 \in (\overline{Q}_1, y_1], Y_0 \in (\overline{Q}_0, y_0] | W = w) + \underline{c} \tilde{\mathbb{P}}(Y_1 \leq \overline{Q}_1, Y_0 \in (\overline{Q}_0, y_0] | W = w) \\
&\quad + \overline{c} \tilde{\mathbb{P}}(Y_1 \in (\overline{Q}_1, y_1], Y_0 \leq \overline{Q}_0 | W = w) + \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq \overline{Q}_1, Y_0 \leq \overline{Q}_0) p^{uu}(Y_1, Y_0, w) | W = w) \\
&= \underline{c} \left( \tilde{\mathbb{P}}(Y_1 \leq \overline{Q}_1, Y_0 \in (\overline{Q}_0, y_0] | W = w) + \tilde{\mathbb{P}}(Y_1 > \overline{Q}_1, Y_0 \in (\overline{Q}_0, y_0] | W = w) \right) \\
&\quad + \overline{c} \left( \tilde{\mathbb{P}}(Y_1 \in (\overline{Q}_1, y_1], Y_0 \leq \overline{Q}_0 | W = w) + \tilde{\mathbb{P}}(Y_1 \in (\overline{Q}_1, y_1], Y_0 > \overline{Q}_0) | W = w) \right) \\
&\quad + \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq \overline{Q}_1, Y_0 \leq \overline{Q}_0) p^{uu}(Y_1, Y_0, w) | W = w) \\
&= \underline{c} \tilde{\mathbb{P}}(Y_0 \leq (\overline{Q}_0, y_0] | W = w) + \overline{c} \tilde{\mathbb{P}}(Y_1 \in (\overline{Q}_1, y_1] | W = w) + \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq \overline{Q}_1, Y_0 \leq \overline{Q}_0) p^{uu}(Y_1, Y_0, w) | W = w),
\end{aligned}$$

where the second equality follows by (C.38) that  $\tilde{\mathbb{P}}$  takes no mass on diagonal area. Next we expand the last line.

$$\begin{aligned}
& \underline{c}\tilde{\mathbb{P}}(Y_0 \leq (\bar{Q}_0, y_0) | W = w) + \bar{c}\tilde{\mathbb{P}}(Y_1 \in (\bar{Q}_1, y_1] | W = w) + \tilde{\mathbb{E}}(\mathbb{1}(Y_1 \leq \bar{Q}_1, Y_0 \leq \bar{Q}_0)p^{uu}(Y_1, Y_0) | W = w) \\
&= \underline{c}[\bar{F}_{Y_0|W}(y_0 | w) - \bar{F}_{Y_0|W}(\bar{Q}_0 | w)] + \bar{c}[\bar{F}_{Y_1|W}(y_1 | w) - \bar{F}_{Y_1|W}(\bar{Q}_0 | w)] \\
&\quad + \tilde{\mathbb{P}}(Y_1 \leq \bar{Q}_1, Y_0 \leq \bar{Q}_0, X = 1 | W = w) \\
&= \frac{p_{0|w}\underline{c}}{1-\underline{c}}(F_{Y|X,W}(y_0|0, w) - F_{Y|X,W}(\bar{Q}_0|0, w)) + p_{1|w}(F_{Y|X,W}(y_1 | 1, w) - F_{Y|X,W}(\bar{Q}_1 | 1, w)) \\
&\quad + p_{1|w} \max\{F_{Y|X,W}(\bar{Q}_1 | 1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 | 1, w) - 1, 0\} \\
&= \frac{p_{0|w}\underline{c}}{1-\underline{c}}(F_{Y|X,W}(y_0|0, w) - F_{Y|X,W}(\bar{Q}_0|0, w)) + p_{1|w}(F_{Y|X,W}(y_1 | 1, w) - F_{Y|X,W}(\bar{Q}_1 | 1, w)) \\
&\quad + p_{1|w}(F_{Y|X,W}(\bar{Q}_1 | 1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 | 1, w) - 1) \\
&= p_{1|w} \left[ F_{Y|X,W}(y_1 | 1, w) + \frac{p_{1|w} - \underline{c}}{p_{1|w}(1-\underline{c})} + \frac{p_{0|w}\underline{c}}{p_{1|w}(1-\underline{c})} F_{Y|X,W}(y_0|0, w) - 1 \right] \\
&= p_{1|w} [F_{Y|X,W}(y_1 | 1, w) + \bar{F}_{Y_0|X,W}(y_0 | 1, w) - 1] \\
&= p_{1|w} \max\{F_{Y|X,W}(y_1 | 1, w) + \bar{F}_{Y_0|X,W}(y_0 | 1, w) - 1, 0\}.
\end{aligned}$$

The third and the last equality hold by the following derivation

$$\begin{aligned}
F_{Y|X,W}(y_1 | 1, w) + \bar{F}_{Y_0|X,W}(y_0 | 1, w) - 1 &\geq F_{Y|X,W}(\bar{Q}_1 | 1, w) + \bar{F}_{Y_0|X,W}(\bar{Q}_0 | 1, w) - 1 \\
&\geq \bar{\tau}_1 + \underline{\tau}_1 - 1 \\
&= 0,
\end{aligned}$$

where the second inequality follows by Lemma C.3.4.9. Hence we have established (C.37), as desired.

Since  $\mathbb{R}^2$  is partitioned by these 5 cases, we haven shown that  $\tilde{\mathbb{E}}[X | Y_1, Y_0, W = w] = p^{uu}(Y_1, Y_0, w)$  almost surely, which concludes the proof of Part 2.

**Proof of Part 3:** One can show that  $(\underline{F}_{Y_1|W}, \bar{F}_{Y_0|W}, \bar{C}_{1,0|X,W})$  can be achieved by the joint distribution of  $(Y_1, Y_0, X)$  conditional on  $W = w$  constructed as below:

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 | W = w) \\
&= x \min\{F_{Y|X,W}(y_1 | 1, w), \bar{F}_{Y_0|X,W}(y_0 | 1, w)\} p_{1|w} \\
&\quad + (1-x) \min\{\underline{F}_{Y_1|X,W}(y_1 | 0, w), F_{Y|X,W}(y_0|0, w)\} p_{0|w}.
\end{aligned}$$

It can be verified that this joint distribution satisfies the following 5 conditions: for all  $y \in \mathbb{R}$  and  $x \in \{0, 1\}$ ,

1.  $\tilde{\mathbb{P}}(Y_1 \leq y \mid W = w) = \underline{F}_{Y_1|W}(y \mid w)$  and  $\tilde{\mathbb{P}}(Y_0 \leq y \mid W = w) = \overline{F}_{Y_0}(y \mid w)$ ;
2.  $\tilde{\mathbb{P}}(X = x \mid W = w) = p_{x|w}$ ;
3.  $\tilde{\mathbb{P}}(Y_x \leq y \mid X = x, W = w) = F_{Y|X,W}(Y \mid X, w)$ ;
4. The following equality holds:

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid X = x, W = w) \\ &= \min \left\{ \tilde{\mathbb{P}}(Y_1 \leq y_1 \mid X = x, W = w), \tilde{\mathbb{P}}(Y_0 \leq y_0 \mid X = x, W = w) \right\}; \end{aligned}$$

5.  $\tilde{\mathbb{E}}(X \mid Y_1, Y_0, W = w) = p^{lu}(Y_1, Y_0, w; B^{lu}) \in [\underline{c}, \overline{c}]$ , for  $\tilde{\mathbb{P}}$ -almost surely with

$$B^{lu} = \frac{\tilde{\mathbb{P}}(Y_1 = y_1, Y_0 = y_0, X = 1 \mid W = w)}{\tilde{\mathbb{P}}(Y_1 = y_1, Y_0 = y_0 \mid W = w)}.$$

The arguments are similar to the proof of Part 1 and thus omitted.

**Proof of Part 4:** One can show that  $(\underline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, \underline{C}_{1,0|X,W})$  can be achieved by the joint distribution of  $(Y_1, Y_0, X)$  conditional on  $W = w$  constructed as below:

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1 \mid W = w) \\ &= x \max \left\{ F_{Y|X,W}(y_1 \mid 1, w) + \underline{F}_{Y_0|X,W}(y_0 \mid 1, w) - 1, 0 \right\} p_{1|w} \\ &+ (1 - x) \max \left\{ \underline{F}_{Y_1|X,W}(y_1 \mid 0, w) + F_{Y|X,W}(y_0 \mid 0, w) - 1, 0 \right\} p_{0|w}. \end{aligned}$$

It can be verified that this joint distribution satisfies the following four conditions: for all  $y \in \mathbb{R}$  and  $x \in \{0, 1\}$ ,

1.  $\tilde{\mathbb{P}}(Y_1 \leq y \mid W = w) = \underline{F}_{Y_1|W}(y \mid w)$  and  $\tilde{\mathbb{P}}(Y_0 \leq y \mid W = w) = \underline{F}_{Y_0|W}(y \mid w)$ ;
2.  $\tilde{\mathbb{P}}(X = x \mid W = w) = p_{x|w}$ ;
3.  $\tilde{\mathbb{P}}(Y_x \leq y \mid X = x, W = w) = F_{Y|X,W}(Y \mid X, w)$ ;
4. The following equality holds:

$$\begin{aligned} & \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0 \mid X = x, W = w) \\ &= \max \left\{ \tilde{\mathbb{P}}(Y_1 \leq y_1 \mid X = x, W = w) + \tilde{\mathbb{P}}(Y_0 \leq y_0 \mid X = x, W = w) - 1, 0 \right\}; \end{aligned}$$

5.  $\tilde{\mathbb{E}}(X | Y_1, Y_0, W = w) = p^l(Y_1, Y_0, w; B^l) \in [\underline{c}, \bar{c}]$ , for  $\tilde{\mathbb{P}}$ -almost surely with

$$B^l = \frac{\tilde{\mathbb{P}}(Y_1 = y_1, Y_0 = y_0, X = 1 | W = w)}{\tilde{\mathbb{P}}(Y_1 = y_1, Y_0 = y_0 | W = w)}.$$

The arguments are similar to the proof of Part 2 and thus omitted.  $\square$

Suppose  $\{F_{Y_1, Y_0, X}^k\}_{k=1}^K$  is a set of valid cdfs of  $(Y_1, Y_0, X)$  whose support of  $X$  is  $\{0, 1\}$ . Consider a mixture of cdfs defined as

$$F_{Y_1, Y_0, X}^{\text{mix}}(y_1, y_0, x) = \sum_{k=1}^K a_k F_{Y_1, Y_0, X}^k(y_1, y_0, x) \quad (\text{C.39})$$

where  $a_k \in [0, 1]$  for all  $k \in \{1, \dots, K\}$ ,  $\sum_{k=1}^K a_k = 1$ , and  $(y_1, y_0, x) \in \mathbb{R}^2 \times \{0, 1\}$ .

Let  $p^k(Y_1, Y_0) := \mathbb{E}^k[X | Y_1, Y_0]$ , where  $\mathbb{E}^k$  denotes the expectation under cdf  $F_{Y_1, Y_0, X}^k$  for  $k \in \{1, \dots, K, \text{mix}\}$ . The next lemma will be used to show that mixtures of distributions satisfying joint  $c$ -dependence also satisfy joint  $c$ -dependence.

**Lemma C.3.9.** *There exists a sequence of Borel measurable function  $\{f_k(\cdot, \cdot)\}_{k=1}^K : \mathbb{R}^2 \rightarrow \mathbb{R}$  such that*

1.  $f_k(Y_1, Y_0) \in [0, 1]$  for each  $k \in \{1, \dots, K\}$ , and  $\sum_{k=1}^K f_k(Y_1, Y_0) = 1$ ,
2.  $p^{\text{mix}}(Y_1, Y_0) = \sum_{k=1}^K f_k(Y_1, Y_0) p^k(Y_1, Y_0)$ ,

almost surely under the distribution  $F_{Y_1, Y_0, X}^{\text{mix}}$ .

*Proof of Lemma C.3.9.* Let  $\mathbb{P}^k$  denote the probability taken under cdf  $F_{Y_1, Y_0, X}^k$  for  $k \in \{1, \dots, K, \text{mix}\}$ . Then it follows by the definition of conditional probability that

$$\begin{aligned} \mathbb{P}^{\text{mix}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1) &= \mathbb{E}^{\text{mix}}[\mathbb{1}[Y_1 \leq y_1, Y_0 \leq y_0] p^{\text{mix}}(Y_1, Y_0)] \\ &= \int_{u \leq y_1, v \leq y_0} p^{\text{mix}}(u, v) dF_{Y_1, Y_0}^{\text{mix}}, \end{aligned}$$

where the last line denotes the Lebesgue-Stieltjes integral with respect to the cdf  $F_{Y_1, Y_0}^{\text{mix}}$ .

Likewise, for each  $k \in \{1, \dots, K\}$ , we have

$$\mathbb{P}^k(Y_1 \leq y_1, Y_0 \leq y_0, X = 1) = \int_{u \leq y_1, v \leq y_0} p^k(u, v) dF_{Y_1, Y_0}^k.$$

Since (C.39) implies that

$$\mathbb{P}^{\text{mix}}(Y_1 \leq y_1, Y_0 \leq y_0, X = 1) = \sum_{k=1}^K a_k \mathbb{P}^k(Y_1 \leq y_1, Y_0 \leq y_0, X = 1),$$

we have

$$\int_{w \leq y_1, v \leq y_0} p^{\text{mix}}(w, v) dF_{Y_1, Y_0}^{\text{mix}} = \sum_{k=1}^K a_k \int_{u \leq y_1, v \leq y_0} p^k(u, v) dF_{Y_1, Y_0}^k. \quad (\text{C.40})$$

Following the Carathéodory extension theorem (e.g., Ash and Doléans-Dade (2000, Theorem 1.4.9)), there exists unique Lebesgue-Stieltjes measures  $\nu^{\text{mix}}$  and  $\{\nu^k\}_{k=1}^K$  defined on  $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$  that are consistent with cdfs  $F_{Y_1, Y_0}^{\text{mix}}$  and  $\{F_{Y_1, Y_0}^k\}_{k=1}^K$ , respectively. Combined with (C.39), this implies

$$\nu^{\text{mix}}(A) = \sum_{k=1}^K a_k \nu^k(A), \quad \text{for all } A \in \mathcal{A} := \{(-\infty, y_0] \times (-\infty, y_1] : (y_0, y_1) \in \mathbb{R}^2\}. \quad (\text{C.41})$$

It can be seen that  $\mathcal{A}$  is a  $\pi$ -system,  $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R}^2)$ , and the class of sets satisfying (C.41) constitutes a  $\lambda$ -system. Following from  $\pi - \lambda$  Theorem,

$$\nu^{\text{mix}}(A) = \sum_{k=1}^K a_k \nu^k(A), \quad \text{for all } A \in \mathcal{B}(\mathbb{R}^2). \quad (\text{C.42})$$

From the above identity (C.42), we note that  $a_k \nu^k \ll \nu^{\text{mix}}$  for all  $k \in \{1, \dots, K\}$ . By Radon-Nikodym Theorem (e.g., Royden and Fitzpatrick (2010, p.386, Problem 54.1)), there exist nonnegative Borel measurable functions  $d(a_k \nu^k)/d\nu^{\text{mix}}$  such that the following equalities hold for all Borel sets  $A \in \mathcal{B}(\mathbb{R}^2)$  and for each  $k \in \{1, \dots, K\}$ :

$$\int_A p^k(u, v) \frac{d(a_k \nu^k)}{d\nu^{\text{mix}}} d\nu^{\text{mix}} = \int_A p^k(u, v) d(a_k \nu^k) = \int_A a_k p^k(u, v) d\nu^k.$$

Taking  $A = (-\infty, y_1] \times (-\infty, y_0]$  and combining these equalities across  $k \in \{1, \dots, K\}$  then gives

$$\begin{aligned} \int_{u \leq y_1, v \leq y_0} \sum_{k=1}^K p^k(u, v) \frac{d(a_k \nu^k)}{d\nu^{\text{mix}}} d\nu^{\text{mix}} &= \sum_{k=1}^K a_k \int_{u \leq y_1, v \leq y_0} p^k(u, v) d\nu^k \\ &= \int_{u \leq y_1, v \leq y_0} p^{\text{mix}}(u, v) d\nu^{\text{mix}} \end{aligned}$$

The second equality holds by (C.40). Since  $\mathcal{A}$  constitutes a  $\pi$ -system, and  $\mathbb{R}^2$  can be written as a countable union of elements in the class  $\mathcal{A}$ , applying Billingsley (1995, Theorem 16.10.(iii)) then leads to the following equality:

$$p^{\text{mix}}(u, v) = \sum_{k=1}^K p^k(u, v) \frac{d(a_k \nu^k)}{d\nu^{\text{mix}}}(u, v)$$

almost surely with respect to the measure  $\nu^{\text{mix}}$  on  $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ . Replacing  $(u, v)$  with random variables  $(Y_1, Y_0)$ , the above equality is equivalent to

$$p^{\text{mix}}(Y_1, Y_0) = \sum_{k=1}^K p^k(Y_1, Y_0) \frac{d(a_k \nu^k)}{d\nu^{\text{mix}}}(Y_1, Y_0)$$

almost surely under the distribution  $F_{Y_1, Y_0}^{\text{mix}}$ . Next we show that the weights add up to one and they are non-negative.

Following from (C.42), the conclusion in Royden and Fitzpatrick (2010, p.386, Problem 54.2) implies

$$\sum_{k=1}^K \frac{d(a_k \nu^k)}{d\nu^{\text{mix}}} = \frac{d\nu^{\text{mix}}}{d\nu^{\text{mix}}} = 1, \quad \text{almost surely-}\nu^{\text{mix}}.$$

By definition, Radon-Nikodym derivatives are nonnegative. So we have

$$\frac{d(a_k \nu^k)}{d\nu^{\text{mix}}} = 1 - \sum_{j \neq k} \frac{d(a_j \nu^j)}{d\nu^{\text{mix}}} \in [0, 1], \quad \text{almost surely-}\nu^{\text{mix}}.$$

Therefore, we have shown that  $p^{\text{mix}}(Y_1, Y_0)$  is a convex combination of  $\{p^k(Y_1, Y_0)\}_{k=1}^K$  almost surely, with weights  $d(a_k \nu^k)/d\nu^{\text{mix}}$  being a non-negative measurable function of  $(Y_1, Y_0)$ , as desired.  $\square$

*Proof of Theorem 4.4.2.* Fix a  $w \in \text{supp}(W)$  and  $(\varepsilon, \gamma) \in [0, 1]^2$ , we prove this by constructing a probability distribution  $\tilde{\mathbb{P}}$  for  $(Y_1, Y_0, X)$  conditional on  $W = w$  such that for all  $y \in \mathbb{R}$  and  $x \in \{0, 1\}$ , the following conditions hold

1.  $\tilde{\mathbb{P}}(Y_1 \leq y \mid W = w) = \varepsilon \underline{F}_{Y_1|W}(y \mid w) + (1 - \varepsilon) \overline{F}_{Y_1|W}(y \mid w)$  and  
 $\tilde{\mathbb{P}}(Y_0 \leq y \mid W = w) = \gamma \underline{F}_{Y_0|W}(y \mid w) + (1 - \gamma) \overline{F}_{Y_0|W}(y \mid w);$

2.  $\tilde{\mathbb{P}}(X = x | W = w) = p_{x|w}$ ;
3.  $\tilde{\mathbb{P}}(Y_x \leq y | X = x, W = w) = F_{Y|X,W}(Y | X, w)$ ;
4.  $\tilde{\mathbb{P}}(X = 1 | Y_1, Y_0, W = w) \in [\underline{c}, \bar{c}]$  for  $\tilde{\mathbb{P}}$ -almost surely.

Compared to the proof of Theorem 4.4.1, we remove the requirement that the copulas between  $Y_1$  and  $Y_0$  conditional on  $(X, W)$  can be arbitrary.

As in Lemma C.3.8, we have constructed the following four joint (conditional) cdfs of  $(Y_1, Y_0, X) | W = w$ :

$$\begin{aligned}
F^{ul}(y_1, y_0, x|w) &= x \min \left\{ F_{Y|X,W}(y_1 | 1, w), \underline{F}_{Y_0|X}(y_0 | 1) \right\} p_{1|w} \\
&\quad + (1-x) \min \left\{ \overline{F}_{Y_1|X,W}(y_1 | 0, w), F_{Y|X,W}(y_0|0, w) \right\} p_{0|w} \\
F^{lu}(y_1, y_0, x|w) &= x \min \left\{ F_{Y|X,W}(y_1 | 1), \overline{F}_{Y_0|X,W}(y_0 | 1, w) \right\} p_{1|w} \\
&\quad + (1-x) \min \left\{ \underline{F}_{Y_1|X,W}(y_1 | 0, w), F_{Y|X,W}(y_0|0, w) \right\} p_{0|w}
\end{aligned}$$

and

$$\begin{aligned}
F^{uu}(y_1, y_0, x|w) &= x \max \left\{ F_{Y|X,W}(y_1 | 1, w) + \overline{F}_{Y_0|X,W}(y_0 | 1, w) - 1, 0 \right\} p_{1|w} \\
&\quad + (1-x) \min \left\{ \overline{F}_{Y_1|X,W}(y_1 | 0, w) + F_{Y|X,W}(y_0|0, w) - 1, 0 \right\} p_{0|w} \\
F^{ll}(y_1, y_0, x|w) &= x \max \left\{ \underline{F}_{Y_1|X,W}(y_1 | 1, w) + \underline{F}_{Y_0|X,W}(y_0 | 1, w) - 1, 0 \right\} p_{1|w} \\
&\quad + (1-x) \min \left\{ \underline{F}_{Y_1|X,W}(y_1 | 0, w) + F_{Y|X,W}(y_0|0, w) - 1, 0 \right\} p_{0|w}.
\end{aligned}$$

Let the joint distribution of  $(Y_1, Y_0, X) | W = w$  be defined as below:

$$\begin{aligned}
&\tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, X = x | W = w) \\
&= \gamma \left[ \varepsilon F^{ll}(y_1, y_0, x|w) + (1-\varepsilon) F^{ul}(y_1, y_0, x|w) \right] \\
&\quad + (1-\gamma) \left[ \varepsilon F^{lu}(y_1, y_0, x|w) + (1-\varepsilon) F^{uu}(y_1, y_0, x|w) \right].
\end{aligned} \tag{C.43}$$

As shown in Lemma C.3.8, the functions  $F^{ll}$ ,  $F^{ul}$ ,  $F^{lu}$ , and  $F^{uu}$  are valid cdfs, hence their convex combination (C.43) also yields a valid cdf. Next we verify that this cdf satisfies

conditions 1-4 listed above, thus concluding that  $(\varepsilon \underline{F}_{Y_1|W} + (1 - \varepsilon) \overline{F}_{Y_1|W}, \gamma \underline{F}_{Y_0|W} + (1 - \gamma) \overline{F}_{Y_0|W})$  is in the identified set.

**Verifying Condition 1:** For  $y \in \mathbb{R}$ , we have

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_1 \leq y \mid W = w) \\
&= \lim_{y_0 \rightarrow +\infty} \tilde{\mathbb{P}}(Y_1 \leq y, Y_0 \leq y_0, X = 1 \mid W = w) + \lim_{y_0 \rightarrow +\infty} \tilde{\mathbb{P}}(Y_1 \leq y, Y_0 \leq y_0, X = 0 \mid W = w) \\
&= \sum_{x=0,1} \gamma \left[ \varepsilon \lim_{y_0 \rightarrow +\infty} F^{ll}(y, y_0, x|w) + (1 - \varepsilon) \lim_{y_0 \rightarrow +\infty} F^{ul}(y, y_0, x|w) \right] \\
&\quad + \sum_{x=0,1} (1 - \gamma) \left[ \varepsilon \lim_{y_0 \rightarrow +\infty} F^{lu}(y, y_0, x|w) + (1 - \varepsilon) \lim_{y_0 \rightarrow +\infty} F^{uu}(y, y_0, x|w) \right] \\
&= \gamma \left[ \varepsilon \underline{F}_{Y_1|W}(y \mid w) + (1 - \varepsilon) \overline{F}_{Y_1|W}(y \mid w) \right] \\
&\quad + (1 - \gamma) \left[ \varepsilon \underline{F}_{Y_1|W}(y \mid w) + (1 - \varepsilon) \overline{F}_{Y_1|W}(y \mid w) \right] \\
&= \varepsilon \underline{F}_{Y_1|W}(y \mid w) + (1 - \varepsilon) \overline{F}_{Y_1|W}(y \mid w),
\end{aligned}$$

where the third equality uses the conclusion from condition 1 in the proof of Lemma C.3.8.

Likewise,

$$\begin{aligned}
& \tilde{\mathbb{P}}(Y_0 \leq y \mid W = w) \\
&= \lim_{y_1 \rightarrow +\infty} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y, X = 1 \mid W = w) + \lim_{y_1 \rightarrow +\infty} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y, X = 0 \mid W = w) \\
&= \sum_{x=0,1} \gamma \left[ \varepsilon \lim_{y_1 \rightarrow +\infty} F^{ll}(y_1, y, x|w) + (1 - \varepsilon) \lim_{y_1 \rightarrow +\infty} F^{ul}(y_1, y, x|w) \right] \\
&\quad + \sum_{x=0,1} (1 - \gamma) \left[ \varepsilon \lim_{y_1 \rightarrow +\infty} F^{lu}(y_1, y, x|w) + (1 - \varepsilon) \lim_{y_1 \rightarrow +\infty} F^{uu}(y_1, y, x|w) \right] \\
&= \gamma \left[ \varepsilon \underline{F}_{Y_0|W}(y \mid w) + (1 - \varepsilon) \overline{F}_{Y_0|W}(y \mid w) \right] \\
&\quad + (1 - \gamma) \left[ \varepsilon \overline{F}_{Y_0|W}(y \mid w) + (1 - \varepsilon) \overline{F}_{Y_0|W}(y \mid w) \right] \\
&= \gamma \underline{F}_{Y_0|W}(y \mid w) + (1 - \gamma) \overline{F}_{Y_0|W}(y \mid w).
\end{aligned}$$

**Verifying Condition 2:** For  $x \in \{0, 1\}$ , we have that

$$\begin{aligned}
& \tilde{\mathbb{P}}(X = x \mid W = w) \\
&= \lim_{y_0, y_1 \rightarrow +\infty} \tilde{\mathbb{P}}(Y_1 \leq y_1, Y_0 \leq y_0, x \mid W = w) \\
&= \gamma \left[ \varepsilon \lim_{y_0, y_1 \rightarrow +\infty} F^{ll}(y_1, y_0, x|w) + (1 - \varepsilon) \lim_{y_0, y_1 \rightarrow +\infty} F^{ul}(y_1, y_0, x|w) \right] \\
&\quad + (1 - \gamma) \left[ \varepsilon \lim_{y_0, y_1 \rightarrow +\infty} F^{lu}(y_1, y_0, x|w) + (1 - \varepsilon) \lim_{y_0, y_1 \rightarrow +\infty} F^{uu}(y_1, y_0, x|w) \right] \\
&= \gamma [\varepsilon p_{x|w} + (1 - \varepsilon) p_{x|w}] + (1 - \gamma) [\varepsilon p_{x|w} + (1 - \varepsilon) p_{x|w}] \\
&= p_{x|w},
\end{aligned}$$

where the third equality follows by the condition 2 in the proof of Lemma C.3.8.

**Verifying Condition 3:** Similar to the proof of condition 2, condition 3 follows by the fact that all the cdfs  $F^{ul}$ ,  $F^{lu}$ ,  $F^{uu}$ , and  $F^{ll}$  satisfy condition 3 as argued in Lemma C.3.8. Hence it follows that their convex combination  $\tilde{\mathbb{P}}$  also satisfies this condition.

**Verifying Condition 4:** As in the proof of Lemma C.3.8, we established propensity score functions  $p^{ul}$ ,  $p^{lu}$ ,  $p^{uu}$ , and  $p^{ll}$  under the cdfs  $F^{ul}$ ,  $F^{lu}$ ,  $F^{uu}$ , and  $F^{ll}$ , respectively. Applying Lemma C.3.9 to the conditional mixture distribution  $\tilde{\mathbb{P}}$  gives the propensity score as below

$$\begin{aligned}
\tilde{\mathbb{E}}(X \mid Y_1, Y_0, W = w) &= \omega^{ul}(Y_1, Y_0) p^{ul}(Y_1, Y_0, w) + \omega^{lu}(Y_1, Y_0) p^{lu}(Y_1, Y_0, w) \\
&\quad + \omega^{uu}(Y_1, Y_0) p^{uu}(Y_1, Y_0, w) + \omega^{ll}(Y_1, Y_0) p^{ll}(Y_1, Y_0, w)
\end{aligned}$$

almost surely under  $\tilde{\mathbb{P}}$ , where  $\omega^k(Y_1, Y_0) \in [0, 1]$ , and  $\sum_k \omega^k(Y_1, Y_0) = 1$  almost surely under  $\tilde{\mathbb{P}}$  for  $k \in \{ul, lu, uu, ll\}$ . Since we have argued that  $p^k(u, v, w) \in [\underline{c}, \bar{c}]$  for  $(u, v) \in \mathbb{R}^2$ , therefore,  $\tilde{\mathbb{E}}(X \mid Y_1, Y_0, W = w) \in [\underline{c}, \bar{c}]$  almost surely under  $\tilde{\mathbb{P}}$ , which concludes the proof.  $\square$

### C.3.3 Proofs for Section 4.4.2

*Proof of Theorem 4.4.3.* First, we prove this proposition when Assumption 12 holds. Fix an arbitrary  $(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y,X,W}; c)$ . By the definition of  $\mathcal{I}_0^{\text{marg}}(F_{Y,X,W}; c)$ , there

exists a joint cdf  $F_{Y_1, Y_0|X, W}$  that generates  $(F_1, F_0, C)$  and satisfies Assumption 12. By Lemma 4.4.1 and the monotonicity assumption, we have

$$\theta(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_{Y, X, W}) \leq \theta(F_1, F_0, F_{Y, X, W}) \leq \theta(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, F_{Y, X, W}).$$

Since  $(F_1, F_0, C)$  is arbitrary, we have

$$\theta(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_{Y, X, W}) \leq \inf_{(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y, X, W}; c)} \theta(F_1, F_0, F_{Y, X, W})$$

and

$$\sup_{(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y, X, W}; c)} \theta(F_1, F_0, F_{Y, X, W}) \leq \theta(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, F_{Y, X, W}).$$

Furthermore, as demonstrated by Theorem 4.4.1,  $(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, C)$  and  $(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, C)$  are contained in the identified set for any copula  $C \in \mathcal{C}_{1,0|X, W}$ . This implies

$$\inf_{(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y, X, W}; c)} \theta(F_1, F_0, F_{Y, X, W}) \leq \theta(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_{Y, X, W})$$

and

$$\sup_{(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y, X, W}; c)} \theta(F_1, F_0, F_{Y, X, W}) \geq \theta(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, F_{Y, X, W}).$$

Thus we conclude that

$$\inf_{(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y, X, W}; c)} \theta(F_1, F_0, F_{Y, X, W}) = \theta(\overline{F}_{Y_1|W}, \underline{F}_{Y_0|W}, F_{Y, X, W})$$

$$\sup_{(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y, X, W}; c)} \theta(F_1, F_0, F_{Y, X, W}) = \theta(\underline{F}_{Y_1|W}, \overline{F}_{Y_0|W}, F_{Y, X, W}).$$

Note that Theorem 4.4.1 also implies that  $(\varepsilon \underline{F}_{Y_1|W} + (1 - \varepsilon) \overline{F}_{Y_1|W}, \gamma \underline{F}_{Y_0|W} + (1 - \gamma) \overline{F}_{Y_0|W}, C_{1,0|X, W})$  belongs to the identified set  $\mathcal{I}_0^{\text{marg}}(F_{Y, X, W}; c)$  for each  $(\varepsilon, \gamma) \in [0, 1]^2$ . By the continuity of the mapping  $(\varepsilon, \gamma) \mapsto \theta(\varepsilon \underline{F}_{Y_1|W} + (1 - \varepsilon) \overline{F}_{Y_1|W}, \gamma \underline{F}_{Y_0|W} + (1 - \gamma) \overline{F}_{Y_0|W})$  and the definition of  $\mathcal{I}_\theta^{\text{marg}}(F_{Y, X, W}; c)$ , the sharpness of the interior then follows by the intermediate value theorem. The proof follows the same arguments when imposing Assumption 13, where we use Theorem 4.4.2 instead of Theorem 4.4.1.  $\square$

*Proof of Lemma 4.4.2. Part 1:* Because  $Q_{Y_x}(U) \sim Y_x$  when  $U \sim \text{Unif}(0, 1)$ , we can write

$$\mathbb{E}[Y_x] = \mathbb{E}[Q_{Y_x}(U)] = \int_0^1 Q_{Y_x}(\tau) d\tau = \int_0^1 \theta_Q(F_{Y_x}; \tau) d\tau.$$

By the proof of Part 2 below,  $\theta_Q(F_{Y_x}; \tau)$  is increasing in  $F_{Y_x}$  for all  $\tau \in (0, 1)$ . This implies  $\theta_{\mathbb{E}}(F_{Y_x}) = \int_0^1 \theta_Q(F_{Y_x}; \tau) d\tau$  is also increasing. Continuity follows from

$$\begin{aligned} \theta_{\mathbb{E}}(\varepsilon \underline{F}_{Y_x} + (1 - \varepsilon) \overline{F}_{Y_x}) &= \int y d(\varepsilon \underline{F}_{Y_x}(y) + (1 - \varepsilon) \overline{F}_{Y_x}(y)) \\ &= \int y d(\varepsilon \underline{F}_{Y_x}(y)) + \int y d((1 - \varepsilon) \overline{F}_{Y_x}(y)) \\ &= \varepsilon \theta_{\mathbb{E}}(\underline{F}_{Y_x}) + (1 - \varepsilon) \theta_{\mathbb{E}}(\overline{F}_{Y_x}) \end{aligned}$$

being continuous in  $\varepsilon$  over  $\varepsilon \in [0, 1]$ .

**Part 2:** Suppose  $F_{Y_x}(y) \leq F'_{Y_x}(y)$  for all  $y \in \mathbb{R}$ . Therefore, for any  $\tau \in (0, 1)$ ,  $\{y \in \mathbb{R} : F_{Y_x}(y) \geq \tau\} \subseteq \{y \in \mathbb{R} : F'_{Y_x}(y) \geq \tau\}$ . Hence,

$$\theta_Q(F_{Y_x}; \tau) = \inf\{y \in \mathbb{R} : F_1(y) \geq \tau\} \geq \inf\{y \in \mathbb{R} : F'_{Y_x}(y) \geq \tau\} = \theta_Q(F'_{Y_x}; \tau).$$

Since  $F_{Y_x} \geq F'_{Y_x}$ , we have that  $\theta_Q(F_{Y_x}; \tau)$  is increasing in  $F_{Y_x}$ .

**Part 3:** Suppose  $F_{Y_x}(y) \leq F'_{Y_x}(y)$  for all  $y \in \mathbb{R}$ . Denote by

$$\begin{aligned} F_{Y_x|X}(y | 1 - x) &= \frac{F_{Y_x}(y) - F_{Y|X}(Y | X)p_x}{p_{1-x}}, \\ F'_{Y_x|X}(y | 1 - x) &= \frac{F'_{Y_x}(y) - F_{Y|X}(Y | X)p_x}{p_{1-x}}. \end{aligned}$$

Then  $F_{Y_x|X}(\cdot | 1 - x) \geq F'_{Y_x|X}(\cdot | 1 - x)$  and, by Part 2,

$$\theta_{CQ}(F_{Y_x}; \tau) = \theta_Q(F_{Y_x|X}(\cdot | 1 - x); \tau) \geq \theta_Q(F'_{Y_x|X}(\cdot | 1 - x); \tau) = \theta_{CQ}(F'_{Y_x}; \tau)$$

for any  $\tau \in (0, 1)$ . Therefore,  $\theta_{CQ}$  is increasing in  $F_{Y_x}$ .

**Part 4:** Suppose  $F_{Y_1|W} \geq F_{Y'_1|W}$ . This implies

$$\int y dF_{Y_1|W}(y | w) \geq \int y dF_{Y'_1|W}(y | w) \text{ for all } w \in \text{supp}(W)$$

which in turn implies

$$\begin{aligned} & \mathbb{1} \left( \int y dF_{Y_1|W}(y | w) - \mathbb{E}[Y_0 | W = w]z \right) \\ & \leq \mathbb{1} \left( \int y dF_{Y'_1|W}(y | w) - \mathbb{E}[Y_0 | W = w] \leq z \right) \end{aligned}$$

for all  $w \in \text{supp}(W)$  and hence

$$\mathbb{P} \left( \int y dF_{Y_1|W}(y | W) - \mathbb{E}[Y_0 | W] \leq z \right) \leq \mathbb{P} \left( \int y dF_{Y'_1|W}(y | W) - \mathbb{E}[Y_0 | W] \leq z \right).$$

The first statement holds by Part 1 of this lemma, the second holds directly, and the third by integrating over the distribution of  $W$ . Therefore, this last cdf is decreasing in  $F_{Y_1|W}$ . By Part 2 of this lemma, its corresponding quantile will be decreasing in  $F_{Y_1|W}$ . This parameter is decreasing in  $F_{Y_0|W}$  because of the minus sign inside the CATE.

**Part 5:** Suppose  $F_{Y_x} \geq F'_{Y_x}$  for  $x \in \{0, 1\}$ . Then,  $F_{Y_x}(y) \leq F'_{Y_x}(y)$  for all  $y \in \mathbb{R}$ . Therefore, for any  $(y_1, y_0) \in \mathbb{R}^2$  and copula  $C$

$$\theta(F_{Y_1}, F_{Y_0}, C; y_1, y_0) = C(F_{Y_1}(y_0), F_{Y_0}(y_0)) \leq C(F'_{Y_1}(y_0), F'_{Y_0}(y_0)) = \theta(F'_{Y_1}, F'_{Y_0}, C; y_1, y_0)$$

because  $C$ , as a copula, is nondecreasing in its arguments. We conclude that this parameter is decreasing in both  $F_{Y_1}$  and  $F_{Y_0}$ .

**Part 6:** We begin by showing that  $(Y_1, Y_0) \sim (Q_{Y_1}(U_1), Q_{Y_0}(U_0))$  where  $(U_1, U_0)$  have joint cdf  $C$ . To see this, note that  $F_{Y_1, Y_0}(y_1, y_0) = C(F_{Y_1}(y_1), F_{Y_0}(y_0))$  by Sklar's Theorem. Also,

$$\begin{aligned} F_{Y_1, Y_0}(y_1, y_0) &= C(F_{Y_1}(y_1), F_{Y_0}(y_0)) \\ &= \mathbb{P}(U_1 \leq F_{Y_1}(y_1), U_0 \leq F_{Y_0}(y_0)) \\ &= \mathbb{P}(Q_{Y_1}(U_1) \leq y_1, Q_{Y_0}(U_0) \leq y_0), \end{aligned}$$

where the third equality follows from Lemma C.3.3.1.

Based on this, we can write the functional as

$$\begin{aligned}
\theta_{\text{DTE}}(F_{Y_1}, F_{Y_0}, C; z) &= \mathbb{E}[\mathbb{1}(Y_1 - Y_0 \leq z)] \\
&= \mathbb{E}[\mathbb{1}(Q_{Y_1}(U_1) - Q_{Y_0}(U_0) \leq z)] \\
&= \int \mathbb{1}(Q_{Y_1}(u_1) - Q_{Y_0}(u_0) \leq z) dC(u_1, u_0).
\end{aligned}$$

Now suppose that  $F_{Y_1} \geq F'_{Y_1}$ . By Part 2 above, this implies that  $Q_{Y_1}(u_1) \geq Q'_{Y_1}(u_1)$  for all  $u_1 \in (0, 1)$  and thus

$$\begin{aligned}
\theta_{\text{DTE}}(F_{Y_1}, F_{Y_0}, C; z) &= \int \mathbb{1}(Q_{Y_1}(u_1) - Q_{Y_0}(u_0) \leq z) dC(u_1, u_0) \\
&\leq \int \mathbb{1}(Q'_{Y_1}(u_1) - Q_{Y_0}(u_0) \leq z) dC(u_1, u_0) \\
&= \theta_{\text{DTE}}(F'_{Y_1}, F_{Y_0}, C; z).
\end{aligned}$$

Therefore,  $\theta_{\text{DTE}}(F_{Y_1}, F_{Y_0}, C; z)$  is decreasing in  $F_{Y_1}$ . An analogous argument shows that it is increasing in  $F_{Y_0}$ .  $\square$

## C.4 Proofs for Section 4.5

*Proof of Proposition 4.5.1.* By Lemma 4.4.1 and the monotonicity of copulas in their arguments, we have that

$$\begin{aligned}
&\sup_{(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y, X, W}; c)} \theta_{\text{CDF}}(F_1, F_0, C, F_{Y, X, W}; y_1, y_0) \\
&\leq \sup_{C \in \mathcal{C}_{1, 0|X, W}} \theta_{\text{CDF}}(\bar{F}_{Y_1|X, W}, \bar{F}_{Y_0|X, W}, C, F_{Y, X, W}; y_1, y_0) \\
&= \theta_{\text{CDF}}(\bar{F}_{Y_1|X, W}, \bar{F}_{Y_0|X, W}, \bar{C}, F_{Y, X, W}; y_1, y_0).
\end{aligned}$$

The equality follows from the Fréchet-Hoeffding bounds. Similarly,

$$\begin{aligned}
&\inf_{(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y, X, W}; c)} \theta_{\text{CDF}}(F_1, F_0, C, F_{Y, X, W}; y_1, y_0) \\
&\geq \inf_{C \in \mathcal{C}_{1, 0|X, W}} \theta_{\text{CDF}}(\underline{F}_{Y_1|X, W}, \underline{F}_{Y_0|X, W}, C, F_{Y, X, W}; y_1, y_0) \\
&= \theta_{\text{CDF}}(\underline{F}_{Y_1|X, W}, \underline{F}_{Y_0|X, W}, \underline{C}, F_{Y, X, W}; y_1, y_0).
\end{aligned}$$

To show sharpness of the interval (4.7), consider the following choices for conditional cdfs and copulas:

$$(\varepsilon_1 \underline{F}_{Y_1|X,W} + (1 - \varepsilon_1) \overline{F}_{Y_1|X,W}, \varepsilon_2 \underline{F}_{Y_0|X,W} + (1 - \varepsilon_2) \overline{F}_{Y_0|X,W}, \varepsilon_3 \underline{C} + (1 - \varepsilon_3) \overline{C})$$

for  $\varepsilon := (\varepsilon_1, \varepsilon_2, \varepsilon_3) \in [0, 1]^3$ . For any  $\varepsilon \in [0, 1]^3$ , this triple belongs to  $\mathcal{I}_0^{\text{marg}}(F_{Y,X,W}; c)$ . Setting  $\varepsilon = (0, 0, 0)$  and  $\varepsilon = (1, 1, 1)$  yields the upper and lower bounds of the interval, so the bounds are sharp. To show the interior is sharp, consider the function

$$\begin{aligned} \varepsilon &\mapsto \theta_{\text{CDF}} \left( \begin{array}{c} \varepsilon_1 \underline{F}_{Y_1|X,W} + (1 - \varepsilon_1) \overline{F}_{Y_1|X,W}, \\ \varepsilon_2 \underline{F}_{Y_0|X,W} + (1 - \varepsilon_2) \overline{F}_{Y_0|X,W}, \\ \varepsilon_3 \underline{C} + (1 - \varepsilon_3) \overline{C}, \\ F_{Y,X,W} \end{array} ; y_1, y_0 \right) \\ &= \varepsilon_3 \mathbb{E} \left[ \max \left\{ \begin{array}{l} \varepsilon_1 \underline{F}_{Y_1|X,W}(y_1 | X, W) + (1 - \varepsilon_1) \overline{F}_{Y_1|X,W}(y_1 | X, W) \\ + \varepsilon_2 \underline{F}_{Y_0|X,W}(y_0 | X, W) + (1 - \varepsilon_2) \overline{F}_{Y_0|X,W}(y_0 | X, W) - 1, 0 \end{array} \right\} \right] \\ &\quad + (1 - \varepsilon_3) \mathbb{E} \left[ \min \left\{ \begin{array}{l} \varepsilon_1 \underline{F}_{Y_1|X,W}(y_1 | X, W) + (1 - \varepsilon_1) \overline{F}_{Y_1|X,W}(y_1 | X, W), \\ \varepsilon_2 \underline{F}_{Y_0|X,W}(y_0 | X, W) + (1 - \varepsilon_2) \overline{F}_{Y_0|X,W}(y_0 | X, W) \end{array} \right\} \right]. \end{aligned}$$

This mapping is continuous in  $\varepsilon_3$ . It is also continuous in  $\varepsilon_1$  and  $\varepsilon_2$  since the functions  $(u, v) \mapsto \underline{C}(u, v)$  and  $(u, v) \mapsto \overline{C}(u, v)$  are both continuous, and by the dominated convergence theorem. Therefore, by the intermediate value theorem, all values in the interval (4.7) are attained and thus the identified set is this interval.  $\square$

*Proof of Proposition 4.5.2.* By lemmas 4.4.1 and 4.4.2.5

$$\begin{aligned} &\sup_{(F_1, F_0, C) \in \mathcal{I}_0^{\text{marg}}(F_{Y,X,W}; c)} \theta_{\text{DTE}}(F_{Y_1|X,W}, F_{Y_0|X,W}, C_{1,0|X,W}, F_{Y,X,W}; z) \\ &\leq \sup_{C \in \mathcal{C}_{1,0|X,W}} \theta_{\text{DTE}}(\overline{F}_{Y_1|X,W}, \underline{F}_{Y_0|X,W}, C, F_{Y,X,W}; z). \end{aligned}$$

By Lemma 2.1 in Fan and Park, 2010,

$$\begin{aligned} &\sup_{C \in \mathcal{C}_{1,0|X,W}} \theta_{\text{DTE}}(\overline{F}_{Y_1|X,W}, \underline{F}_{Y_0|X,W}, C, F_{Y,X,W}; z) \\ &\leq 1 + \mathbb{E} \left[ \min \left\{ \inf_{y \in \mathbb{R}} \left( \overline{F}_{Y_1|X,W}(y | X, W) - \underline{F}_{Y_0|X,W}(y - z | X, W) \right), 0 \right\} \right]. \end{aligned} \tag{C.44}$$

This bound can be attained because the cdf pair  $(\bar{F}_{Y_1|X,W}, \underline{F}_{Y_0|X,W})$  is attainable by Theorem 4.4.1, and the bound in (C.44) is attained by Lemma 2.1 in Fan and Park, 2010 since the set of conditional copulas under marginal  $c$ -dependence is unrestricted.

Similar proof can be used to show that the lower bound

$$\mathbb{E} \left[ \max \left\{ \sup_{y \in \mathbb{R}} \left( \underline{F}_{Y_1|X,W}(y | X, W) - \bar{F}_{Y_0|X,W}(y - z | X, W) \right), 0 \right\} \right]$$

is sharp as well.  $\square$

## C.5 Appendix: Explicit bounds on expected potential outcomes

**Lemma C.5.1.** *Let  $Y$  be random variable with cdf  $F$  and quantile function  $Q$ . Suppose  $\mathbb{E}(|Y|) < \infty$ . Then, for  $a \in (0, 1)$ :*

$$\begin{aligned} \int_0^a Q(u) du &= \mathbb{E}[Y | Y \leq Q(a)]F(Q(a)) - Q(a)(F(Q(a)) - a) \\ \int_a^1 Q(u) du &= \mathbb{E}[Y | Y > Q(a)](1 - F(Q(a))) + Q(a)(F(Q(a)) - a). \end{aligned}$$

If  $\mathbb{P}(Y = Q(a)) = 0$ , then

$$\begin{aligned} \int_0^a Q(u) du &= \mathbb{E}[Y | Y \leq Q(a)]a \\ \int_a^1 Q(u) du &= \mathbb{E}[Y | Y > Q(a)](1 - a). \end{aligned}$$

**Lemma C.5.2.** *Let Assumption 11 hold. Then,*

$$\begin{aligned} & \int y d\bar{F}_{Y_1|W}(y | w) \\ &= \frac{p_{1|w}}{c} \left[ \mathbb{E}[Y | Y \leq \bar{Q}_1, X = 1, W = w] F_{Y|X,W}(\bar{Q}_1 | 1, w) - \bar{Q}_1 (F_{Y|X,W}(\bar{Q}_1 | 1, w) - \bar{\tau}_1) \right] \\ &+ \frac{p_{1|w}}{c} \left[ \mathbb{E}[Y | Y > \bar{Q}_1, X = 1, W = w] (1 - F_{Y|X,W}(\bar{Q}_1 | 1, w)) + \bar{Q}_1 (\bar{F}_{Y|X,W}(\bar{Q}_1 | 1, w) - \bar{\tau}_1) \right], \end{aligned} \quad (\text{C.45})$$

$$\begin{aligned} & \int y d\underline{F}_{Y_1|W}(y | w) \\ &= \frac{p_{1|w}}{c} \left[ \mathbb{E}[Y | Y \leq \underline{Q}_1, X = 1, W = w] F_{Y|X,W}(\underline{Q}_1 | 1, w) - \underline{Q}_1 (F_{Y|X,W}(\underline{Q}_1 | 1, w) - \underline{\tau}_1) \right] \\ &+ \frac{p_{1|w}}{c} \left[ \mathbb{E}[Y | Y > \underline{Q}_1, X = 1, W = w] (1 - F_{Y|X,W}(\underline{Q}_1 | 1, w)) + \underline{Q}_1 (F_{Y|X,W}(\underline{Q}_1 | 1, w) - \underline{\tau}_1) \right] \end{aligned} \quad (\text{C.46})$$

and

$$\begin{aligned}
& \int y d\bar{F}_{Y_0|W}(y | w) \\
&= \frac{p_{0|w}}{1-\bar{c}} [\mathbb{E}[Y | Y \leq \bar{Q}_0, X = 0, W = w] F_{Y|X,W}(\bar{Q}_0|0, w) - \bar{Q}_0 (F_{Y|X,W}(\bar{Q}_0|0, w) - \bar{\tau}_0)] \\
&+ \frac{p_{0|w}}{1-\underline{c}} [\mathbb{E}[Y | Y > \bar{Q}_0, X = 0, W = w] (1 - F_{Y|X,W}(\bar{Q}_0|0, w)) + \bar{Q}_0 (F_{Y|X,W}(\bar{Q}_0|0, w) - \bar{\tau}_0)],
\end{aligned} \tag{C.47}$$

$$\begin{aligned}
& \int y d\underline{F}_{Y_0|W}(y | w) \\
&= \frac{p_{0|w}}{1-\underline{c}} [\mathbb{E}[Y | Y \leq \underline{Q}_0, X = 0, W = w] F_{Y|X,W}(\underline{Q}_0|0, w) - \underline{Q}_0 (F_{Y|X,W}(\underline{Q}_0|0, w) - \underline{\tau}_0)] \\
&+ \frac{p_{0|w}}{1-\bar{c}} [\mathbb{E}[Y | Y > \underline{Q}_0, X = 0, W = w] (1 - F_{Y|X,W}(\underline{Q}_0|0, w)) + \underline{Q}_0 (F_{Y|X,W}(\underline{Q}_0|0, w) - \underline{\tau}_0)].
\end{aligned} \tag{C.48}$$

If  $Y$  is continuously distributed conditionally on  $(X, W)$ , then these expressions simplify to

$$\begin{aligned}
\int y d\bar{F}_{Y_1|W}(y | w) &= \mathbb{E}[Y | Y \leq \bar{Q}_1, X = 1, W = w] \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} \\
&+ \mathbb{E}[Y | Y > \bar{Q}_1, X = 1, W = w] \frac{p_{1|w} - \underline{c}}{\bar{c} - \underline{c}} \\
\int y d\underline{F}_{Y_1|W}(y | w) &= \mathbb{E}[Y | Y \leq \underline{Q}_1, X = 1, W = w] \frac{p_{1|w} - \underline{c}}{\bar{c} - \underline{c}} \\
&+ \mathbb{E}[Y | Y > \underline{Q}_1, X = 1, W = w] \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}}
\end{aligned}$$

and

$$\begin{aligned}
\int y d\bar{F}_{Y_0|W}(y | w) &= \mathbb{E}[Y | Y \leq \bar{Q}_0, X = 0, W = w] \frac{p_{1|w} - \underline{c}}{\bar{c} - \underline{c}} \\
&+ \mathbb{E}[Y | Y > \bar{Q}_0, X = 0, W = w] \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} \\
\int y d\underline{F}_{Y_0|W}(y | w) &= \mathbb{E}[Y | Y \leq \underline{Q}_0, X = 0, W = w] \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} \\
&+ \mathbb{E}[Y | Y > \underline{Q}_0, X = 0, W = w] \frac{p_{1|w} - \underline{c}}{\bar{c} - \underline{c}}
\end{aligned}$$

where  $\bar{Q}_x$ ,  $\underline{Q}_x$ ,  $\bar{\tau}_x$ , and  $\underline{\tau}_x$  for  $x = 0, 1$  are defined in Appendix C.1.

### C.5.1 Proofs of analytical bounds on expectations

*Proof of Lemma C.5.1.* First we consider the equality involving  $\int_0^a Q(u) du$ . Note that

$$\begin{aligned}
 \int_0^a Q(u) du &= \int_0^1 Q(u) \mathbb{1}[u \leq a] du \\
 &= \int_0^1 Q(u) \mathbb{1}[Q(u) \leq Q(a), u \leq a] du \\
 &= \int_0^1 Q(u) \mathbb{1}[Q(u) \leq Q(a)] du - \int_0^1 Q(u) \mathbb{1}[Q(u) \leq Q(a), u > a] du \\
 &= \int_0^1 Q(u) \mathbb{1}[Q(u) \leq Q(a)] du - \int_0^1 Q(u) \mathbb{1}[Q(u) = Q(a), u > a] du \\
 &= \int_0^1 Q(u) \mathbb{1}[Q(u) \leq Q(a)] du - Q(a) \int_0^1 \mathbb{1}[Q(u) \leq Q(a), u > a] du,
 \end{aligned}$$

where the second, the fourth, and the last line follow by the monotonicity of quantile function  $Q(\cdot)$ .

The first term in the last line can be written as below:

$$\int_0^1 Q(u) \mathbb{1}[Q(u) \leq Q(a)] du = \mathbb{E}(Y \mathbb{1}[Y \leq Q(a)]) = \mathbb{E}(Y \mid Y \leq Q(a)) F(Q(a)),$$

where the first equality follows by that  $Q(U)$  has the same distribution as  $Y$  if  $U$  is uniformly distributed over  $[0, 1]$ . To expand the second term, note that  $\{u : Q(u) \leq Q(a), u > a\}$  is a half-open interval with the left endpoint  $a$  excluded, and right endpoint  $\sup\{u : Q(u) \leq Q(a)\}$  included in the interval due to the left-continuity of quantile function  $Q(\cdot)$ . So we have

$$\begin{aligned}
 Q(a) \int_0^1 \mathbb{1}[Q(u) \leq Q(a), u > a] du &= Q(a)(\sup\{u : Q(u) \leq Q(a)\} - a) \\
 &= Q(a)(\sup\{u : u \leq F(Q(a))\} - a) \\
 &= Q(a)(F(Q(a)) - a),
 \end{aligned}$$

where the second line holds by Lemma C.3.3.1. Given the above derivations, we conclude that

$$\int_0^a Q(u) du = \mathbb{E}[Y \mid Y \leq Q(a)] F(Q(a)) - Q(a)(F(Q(a)) - a),$$

as desired.

Regarding the second equality involving  $\int_a^1 Q(u)du$ , note that  $\int_0^1 Q(u) du = \mathbb{E}[Q(U)] = \mathbb{E}[Y]$ . This implies

$$\begin{aligned} \int_a^1 Q(u) du &= \int_0^1 Q(u)du - \int_0^a Q(u) du \\ &= \mathbb{E}[Y] - \mathbb{E}[Y | Y \leq Q(a)]F(Q(a)) + Q(a)(F(Q(a)) - a) \\ &= \mathbb{E}[Y | Y > Q(a)](1 - F(Q(a))) + Q(a)(F(Q(a)) - a), \end{aligned}$$

where the last line follows by the law of iterated expectation. So the second equality is established.

When  $\mathbb{P}(Y = Q(a)) = 0$ , the the CDF  $F$  is continuous at  $Q(a)$ , which implies that  $F(Q(a)) = a$  by Lemma C.3.3.2. Therefore,

$$\int_0^a Q(u) du = \mathbb{E}[Y | Y \leq Q(a)]F(Q(a)) - Q(a)(F(Q(a)) - a) = \mathbb{E}[Y | Y \leq Q(a)]a,$$

and similar arguments can be applied to  $\int_a^1 Q(u) du$  as well. Therefore we have established the desired result.  $\square$

*Proof of Lemma C.5.2.* We prove the claim for  $\int_0^1 y d\bar{F}_{Y_1|W}(y | w)$ , and note that the claims for the other terms can be derived analogously.

Let  $U \sim \text{Unif}(0, 1)$ , then  $\underline{Q}_{Y_1|W}(U | 1, w)$  has the distribution  $\bar{F}_{Y_1|W}(\cdot | 1, w)$ , which implies

$$\begin{aligned} \int y d\bar{F}_{Y_1|W}(y | w) &= \int_0^1 \underline{Q}_{Y_1|W}(\tau | w) d\tau \\ &= \int_0^1 Q_{Y|X,W} \left( \frac{\underline{c}\tau}{p_{1|w}} | 1, w \right) \mathbb{1} \left[ \tau \leq \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} \right] d\tau \end{aligned} \quad (\text{C.49})$$

$$+ \int_0^1 Q_{Y|X,W} \left( \frac{p_{1|w} - \bar{c} + \bar{c}\tau}{p_{1|w}} | 1, w \right) \mathbb{1} \left[ \tau > \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} \right] d\tau \quad (\text{C.50})$$

We expand the term (C.49) below:

$$\begin{aligned}
& \int_0^1 Q_{Y|X,W} \left( \frac{\underline{c}\tau}{p_{1|w}} \mid 1, w \right) \mathbb{1} \left[ \tau \leq \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} \right] d\tau \\
&= \frac{p_{1|w}}{\underline{c}} \int_0^{\bar{\tau}_1} Q_{Y|X,W}(u \mid 1, w) du \\
&= \frac{p_{1|w}}{\underline{c}} \left[ \mathbb{E}[Y \mid Y \leq Q_{Y|X,W}(\bar{\tau}_1 \mid 1, w), X = 1, W = w] F_{Y|X,W}(Q_{Y|X,W}(\bar{\tau}_1 \mid 1, w) \mid 1, w) \right] \\
&\quad - \frac{p_{1|w}}{\underline{c}} Q_{Y|X,W}(\bar{\tau}_1 \mid 1, w) \left[ F_{Y|X,W}(Q_{Y|X,W}(\bar{\tau}_1 \mid 1, w) \mid 1, w) - \bar{\tau}_1 \right] \\
&= \frac{p_{1|w}}{\underline{c}} \left[ \mathbb{E}[Y \mid Y \leq \bar{Q}_1, X = 1, W = w] F_{Y|X,W}(\bar{Q}_1 \mid 1, w) - \bar{Q}_1 (F_{Y|X,W}(\bar{Q}_1 \mid 1, w) - \bar{\tau}_1) \right].
\end{aligned} \tag{C.51}$$

The first equality uses the changes of variable  $u = \underline{c}\tau/p_{1|w}$  and recall that

$$\bar{\tau}_1 = \frac{\underline{c}}{p_{1|w}} \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}}.$$

The second equality follows by Lemma C.5.1, and the last line holds by recalling that

$$\bar{Q}_1 = Q_{Y|X,W}(\bar{\tau}_1 \mid 1, w).$$

Similarly, we can expand the term (C.50) below:

$$\begin{aligned}
& \int_0^1 Q_{Y|X,W} \left( \frac{p_{1|w} - \bar{c} + \bar{c}\tau}{p_{1|w}} \mid 1, w \right) \mathbb{1} \left[ \tau > \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} \right] d\tau \\
&= \int_{\bar{\tau}_1}^1 Q_{Y|X,W}(u \mid 1, w) du \\
&= \frac{p_{1|w}}{\bar{c}} \left[ \mathbb{E}[Y \mid Y > \bar{Q}_1, X = 1, W = w] (1 - F_{Y|X,W}(\bar{Q}_1 \mid 1, w)) + \bar{Q}_1 (F_{Y|X,W}(\bar{Q}_1 \mid 1, w) - \bar{\tau}_1) \right],
\end{aligned} \tag{C.52}$$

where we use the change of variable  $u = 1 - \frac{(1-\tau)\bar{c}}{p_{1|w}}$  in the second line.

Combining the above results, we can combine (C.51) and (C.52) to obtain the analytical formula of  $\int y d\bar{F}_{Y_1|W}$ :

$$\begin{aligned}
& \int y d\bar{F}_{Y_1|W}(y \mid w) \\
&= \frac{p_{1|w}}{\underline{c}} \left[ \mathbb{E}[Y \mid Y \leq \bar{Q}_1, X = 1, W = w] F_{Y|X,W}(\bar{Q}_1 \mid 1, w) - \bar{Q}_1 (F_{Y|X,W}(\bar{Q}_1 \mid 1, w) - \bar{\tau}_1) \right] \\
&\quad + \frac{p_{1|w}}{\bar{c}} \left[ \mathbb{E}[Y \mid Y > \bar{Q}_1, X = 1, W = w] (1 - F_{Y|X,W}(\bar{Q}_1 \mid 1, w)) + \bar{Q}_1 (F_{Y|X,W}(\bar{Q}_1 \mid 1, w) - \bar{\tau}_1) \right].
\end{aligned}$$

Finally, we note that if  $Y$  is continuously distributed conditional on  $(X, W)$ , then

$$F_{Y|X,W}(\bar{Q}_1 | 1, w) = \bar{\tau}_1,$$

which implies

$$\begin{aligned} & \int y d\bar{F}_{Y_1|W}(y | w) \\ &= \frac{p_{1|w}}{\underline{c}} \mathbb{E}[Y | Y \leq \bar{Q}_1, X = 1, W = w] \bar{\tau}_1 + \frac{p_{1|w}}{\bar{c}} \mathbb{E}[Y | Y > \bar{Q}_1, X = 1, W = w] (1 - \bar{\tau}_1) \\ &= \mathbb{E}[Y | Y \leq \bar{Q}_1, X = 1, W = w] \frac{\bar{c} - p_{1|w}}{\bar{c} - \underline{c}} + \mathbb{E}[Y | Y > \bar{Q}_1, X = 1, W = w] \frac{p_{1|w} - \underline{c}}{\bar{c} - \underline{c}}, \end{aligned}$$

as desired. □

## Bibliography

- Agan, A., Doleac, J. L., & Harvey, A. (2023). Misdemeanor prosecution. *The Quarterly Journal of Economics*, 138(3), 1453–1505.
- Aizawa, N., Mommaerts, C., & Rennane, S. L. (2023). *Firm accommodation after disability: Labor market impacts and implications for social insurance* (tech. rep.). National Bureau of Economic Research.
- Andresen, M. E. (2018). Exploring marginal treatment effects: Flexible estimation using stata. *The Stata Journal*, 18(1), 118–158.
- Andrews, D. W. (2017). Identification-robust subvector inference. *Available at SSRN 3032675*.
- Andrews, D. W., & Cheng, X. (2012). Estimation and inference with weak, semi-strong, and strong identification. *Econometrica*, 80(5), 2153–2211.
- Andrews, D. W., & Cheng, X. (2013). Maximum likelihood estimation and uniform inference with sporadic identification failure. *Journal of Econometrics*, 173(1), 36–56.
- Andrews, D. W., & Cheng, X. (2014). Gmm estimation and uniform subvector inference with possible identification failure. *Econometric Theory*, 30(2), 287–333.
- Andrews, D. W., Cheng, X., & Guggenberger, P. (2020). Generic results for establishing the asymptotic size of confidence sets and tests. *Journal of Econometrics*, 218(2), 496–531.
- Andrews, D. W., & Guggenberger, P. (2010). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory*, 26(2), 426–468.
- Andrews, D. W., & Guggenberger, P. (2017). Asymptotic size of kleibergens lm and conditional lr tests for moment condition models. *Econometric Theory*, 33(5), 1046–1080.
- Andrews, D. W., & Guggenberger, P. (2019). Identification-and singularity-robust inference for moment condition models. *Quantitative Economics*, 10(4), 1703–1746.
- Andrews, I. (2016). Conditional linear combination tests for weakly identified models. *Econometrica*, 84(6), 2155–2182.
- Andrews, I. (2018). Valid two-step identification-robust confidence sets for gmm. *Review of Economics and Statistics*, 100(2), 337–348.

- Andrews, I., & Mikusheva, A. (2016b). Conditional inference with a functional nuisance parameter. *Econometrica*, *84*(4), 1571–1612.
- Andrews, I., & Mikusheva, A. (2016a). A geometric approach to nonlinear econometric models. *Econometrica*, *84*(3), 1249–1264.
- Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, *11*(1), 727–753.
- Ash, R. B., & Doléans-Dade, C. A. (2000). *Probability and measure theory*. Academic Press.
- Baron, E. J., & Gross, M. (2023). Is there a foster care-to-prison pipeline? evidence from quasi-randomly assigned investigators. *Working paper*.
- Bhuller, M., Dahl, G. B., Løken, K. V., & Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of Political Economy*, *128*(4), 1269–1324.
- Billingsley, P. (1995). *Probability and measure* (3rd). John Wiley & Sons.
- Björklund, A., & Moffitt, R. (1987). The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics*, 42–49.
- Bonvini, M., & Kennedy, E. H. (2022). Sensitivity analysis via the proportion of unmeasured confounding. *Journal of the American Statistical Association*, *117*(539), 1540–1550.
- Brave, S., & Walstrum, T. (2014). Estimating marginal treatment effects using parametric and semiparametric methods. *The Stata Journal*, *14*(1), 191–217.
- Brinch, C. N., Mogstad, M., & Wiswall, M. (2017). Beyond late with a discrete instrument. *Journal of Political Economy*, *125*(4), 985–1039.
- Bugni, F. A., & Horowitz, J. L. (2021). Permutation tests for equality of distributions of functional data. *Journal of Applied Economics*, *36*(7), 861–877.
- Bugni, F. A., Bunting, J., & Ura, T. (2024). Testing homogeneity in dynamic discrete games in finite samples. *arXiv preprint arXiv:2010.02297*.
- Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2010). Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica*, *78*(1), 377–394.
- Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2011). Estimating marginal returns to education. *American Economic Review*, *101*(6), 2754–81.

- Carneiro, P., & Lee, S. (2009). Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality. *Journal of Econometrics*, *149*(2), 191–208.
- Chaudhuri, S., & Zivot, E. (2011). A new method of projection-based inference in gmm with weakly identified nuisance parameters. *Journal of Econometrics*, *164*(2), 239–251.
- Chen, Q., & Fang, Z. (2019). Improved inference on the rank of a matrix. *Quantitative Economics*, *10*(4), 1787–1824.
- Cheng, X. (2015). Robust inference in nonlinear models with mixed identification strength. *Journal of Econometrics*, *189*(1), 207–228.
- Chernozhukov, V., Fernández-Val, I., & Luo, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica*, *86*(6), 1911–1938.
- Chernozhukov, V., & Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, *73*(1), 245–261.
- Chung, E., & Romano, J. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, *41*(2), 484–507.
- Chung, E., & Romano, J. (2016). Multivariate and multiple permutation tests. *Journal of Econometrics*, *193*(1), 76–91.
- Chung, E., & Olivares, M. (2021). Permutation test for heterogeneous treatment effects with a nuisance parameter. *Journal of Econometrics*, *225*(2), 148–174.
- Chyn, E., Frandsen, B., & Leslie, E. C. (2024). *Examiner and judge designs in economics: A practitioner’s guide* (tech. rep.). National Bureau of Economic Research.
- Cox, G. (2022). Weak identification with bounds in a class of minimum distance models. *arXiv preprint arXiv:2012.11222*.
- Davidson, J. (1994). *Stochastic limit theory*. Oxford University Press.
- Devereux, P. J. (2022). *Fragility of the marginal treatment effect* (tech. rep.). Working Paper.
- DiCiccio, C. J., & Romano, J. P. (2017). Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*, *112*(519), 1211–1220.

- Ding, P., & VanderWeele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, *27*(3), 368.
- Ditzhaus, M., & Gaigall, D. (2022). Testing marginal homogeneity in hilbert spaces with applications to stock market returns. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, *31*(3), 749–770.
- Dorn, J., & Guo, K. (2023). Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, *118*(544), 2645–2657.
- Dorn, J., Guo, K., & Kallus, N. (2024). Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, 1–12.
- Dorn, J., & Yap, L. (2024). Sensitivity analysis for linear estimands. *Working Paper*.
- Duarte, G. (2024). A unified approach for assessing sensitivity to violations of causal assumptions. *Working paper*.
- Dubin, J. A., & McFadden, D. L. (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica: Journal of the Econometric Society*, 345–362.
- Dufour, J.-M., & Taamouti, M. (2005). Projection-based statistical inference in linear structural models with possibly weak instruments. *Econometrica*, *73*(4), 1351–1365.
- Fan, Y., & Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, *26*(03), 931–951.
- Frandsen, B., Lefgren, L., & Leslie, E. (2023). Judging judge fixed effects. *American Economic Review*, *113*(1), 253–277.
- Friedrich, S., Brunner, E., & Pauly, M. (2017). Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, *153*, 255–265.
- Gaigall, D. (2020). Testing marginal homogeneity of a continuous bivariate distribution with possibly incomplete paired data. *Metrika*, *83*(4), 437–465.
- Gu, J., Russell, T., & Stringham, T. (2024). Counterfactual identification and latent space enumeration in discrete outcome models. *Review of Economic Studies* (forthcoming).

- Guggenberger, P., Kleibergen, F., & Mavroeidis, S. (2019). A more powerful subvector anderson rubin test in linear instrumental variables regression. *Quantitative Economics*, *10*(2), 487–526.
- Guggenberger, P., Kleibergen, F., & Mavroeidis, S. (2023). A powerful subvector anderson rubin test in linear instrumental variables regression with conditional heteroskedasticity. *Econometric Theory*, forthcoming.
- Guggenberger, P., Kleibergen, F., Mavroeidis, S., & Chen, L. (2012). On the asymptotic sizes of subset anderson–rubin and lagrange multiplier tests in linear instrumental variables regression. *Econometrica*, *80*(6), 2649–2666.
- Han, S., & McCloskey, A. (2019). Estimation and inference with a (nearly) singular jacobian. *Quantitative Economics*, *10*(3), 1019–1068.
- Hansen, B. (2022). *Probability and statistics for economists*. Princeton University Press.
- Heckman, J., Urzua, S., & Vytlacil, E. (2006). Estimation of treatment effects under essential heterogeneity. *Working Paper*.
- Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, *64*(4), 487–535.
- Heckman, J. J., & Vytlacil, E. (2001). Policy-relevant treatment effects. *American Economic Review*, *91*(2), 107–111.
- Heckman, J. J., & Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, *73*(3), 669–738.
- Heckman, J. J., & Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences*, *96*(8), 4730–4734.
- Huang, M., & Pimentel, S. D. (2025). Variance-based sensitivity analysis for weighting estimators results in more informative bounds. *Biometrika*, *112*(1).
- Igami, M., & Yang, N. (2016). Unobserved heterogeneity in dynamic games: Cannibalization and preemptive entry of hamburger chains in canada. *Quantitative Economics*, *7*(2), 483–521.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge University Press.

- Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics & probability letters*, 36(1), 9–21.
- Jochmans, K. (2023). Many (weak) judges in judge-leniency designs. *TSE Working Paper*.
- Kallus, N., & Zhou, A. (2018). Confounding-robust policy improvement. *Advances in Neural Information Processing Systems*, 31.
- Khmaladze, E. (2016). Unitary transformations, empirical processes and distribution free testing. *Bernoulli*, 22(1), 563–588.
- Kleibergen, F. (2002). Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5), 1781–1803.
- Kleibergen, F. (2005). Testing parameters in gmm without assuming that they are identified. *Econometrica*, 73(4), 1103–1123.
- Kleibergen, F., & Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics*, 133(1), 97–126.
- Kline, P., & Walters, C. R. (2019). On heckits, late, and numerical equivalence. *Econometrica*, 87(2), 677–696.
- Kowalski, A. E. (2023). Reconciling seemingly contradictory results from the oregon health insurance experiment and the massachusetts health reform. *Review of Economics and Statistics*, 105(3), 646–664.
- Lee, D. S., McCrary, J., Moreira, M. J., Porter, J. R., & Yap, L. (2023). *What to do when you can't use '1.96' confidence intervals for iv* (tech. rep.). National Bureau of Economic Research.
- Lee, J. (2015). Asymptotic sizes of subset anderson-rubin tests with weakly identified nuisance parameters and general covariance structure.
- Lehmann, E. L., & Romano, J. P. (2022). *Testing statistical hypothesis: Fourth edition*. Springer.
- Lewis, D. J., & Mertens, K. (2022). A robust test for weak instruments with multiple endogenous regressors. *Working Paper*.
- Ma, Y. (2023). Identification-robust inference for the late with high-dimensional covariates. *arXiv preprint arXiv:2302.09756*.

- Maestas, N., Mullen, K. J., & Strand, A. (2013). Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *American Economic Review*, *103*(5), 1797–1829.
- Makarov, G. (1982). Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & its Applications*, *26*(4), 803–806.
- Manski, C. F. (1997). Monotone treatment response. *Econometrica*, 1311–1334.
- Manski, C. F. (2003). *Partial identification of probability distributions*. Springer.
- Masten, M. A., & Poirier, A. (2016). Partial independence in nonseparable models. *cemmap working paper*, CWP26/16.
- Masten, M. A., & Poirier, A. (2018a). Identification of treatment effects under conditional partial independence. *Econometrica*, *86*(1), 317–351.
- Masten, M. A., & Poirier, A. (2018b). Salvaging falsified instrumental variable models. *arXiv preprint arXiv:1812.11598v1*.
- Masten, M. A., & Poirier, A. (2020). Inference on breakdown frontiers. *Quantitative Economics*, *11*(1), 41–111.
- Masten, M. A., & Poirier, A. (2023). Choosing exogeneity assumptions in potential outcome models. *The Econometrics Journal*, *26*(3), 327–349.
- Masten, M. A., Poirier, A., & Zhang, L. (2024). Assessing sensitivity to unconfoundedness: Estimation and inference. *Journal of Business & Economic Statistics*, *42*(1), 1–13.
- Mikusheva, A., & Sun, L. (2022). Inference with many weak instruments. *Review of Economic Studies*, *89*(5), 2663–2686.
- Moffitt, R. (2008). Estimating marginal treatment effects in heterogeneous populations. *Annales d’Economie et de Statistique*, 239–261.
- Mogstad, M., Santos, A., & Torgovitsky, A. (2017). *Using instrumental variables for inference about policy relevant treatment parameters* (tech. rep.). National Bureau of Economic Research.
- Mogstad, M., Santos, A., & Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, *86*(5), 1589–1619.

- Mogstad, M., & Torgovitsky, A. (2018). Identification and extrapolation of causal effects with instrumental variables. *Annual Review of Economics*, 10, 577–613.
- Mogstad, M., & Torgovitsky, A. (2024). *Instrumental variables with unobserved heterogeneity in treatment effects* (tech. rep.). NBER.
- Montiel Olea, J. L., & Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3), 358–369.
- Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4), 1027–1048.
- Nelsen, R. B. (2006). *An introduction to copulas* (Second). Springer.
- Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4, 2111–2245.
- Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica: Journal of the Econometric Society*, 1815–1820.
- Otsu, T., Pesendorfer, M., & Takahashi, Y. (2016). Pooling data across markets in dynamic markov games. *Quantitative Economics*, 7(2), 523–559.
- Pan, Z., Wang, Z., Zhang, J., & Zhou, Y. (2024). Marginal treatment effects in the absence of instrumental variables. *arXiv preprint arXiv:2401.17595*.
- Pauly, M., Brunner, E., & Konietzschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(2), 461–473.
- Rambachan, A., Coston, A., & Kennedy, E. (2023). Robust design and evaluation of predictive algorithms under unobserved confounding. *arXiv preprint arXiv:2212.09844*.
- Rosenbaum, P. R. (1995). *Observational studies*. Springer.
- Rosenbaum, P. R. (2002). *Observational studies* (Second). Springer.
- Rosenbaum, P. R. (2017). *Observation and experiment: An introduction to causal inference*. Harvard University Press.
- Royden, H., & Fitzpatrick, P. M. (2010). *Real analysis* (4th). Pearson.
- Sasaki, Y., & Ura, T. (2023). Estimation and inference for policy relevant treatment effects. *Journal of Econometrics*, 234(2), 394–450.

- Scheffe, H. (1959). *The analysis of variance*. John Wiley & Sons.
- Sjölander, A. (2024). Sharp bounds for causal effects based on ding and vanderweele’s sensitivity parameters. *Journal of Causal Inference*, *12*(1), 20230019.
- Sklar, M. (1959). *Fonctions de répartition à  $n$  dimensions et leurs marges*. Université Paris 8.
- Staiger, D. O., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557–586.
- Stock, J. H., & Wright, J. H. (2000). Gmm with weak identification. *Econometrica*, *68*(5), 1055–1096.
- Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear iv regression. *Identification and Inference for Econometric Models: Essays in honor of Thomas Rothenberg*, 80–108.
- Stoye, J. (2010). Partial identification of spread parameters. *Quantitative Economics*, *1*(2), 323–357.
- Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, *101*(476), 1619–1637.
- Tan, Z. (2024). Model-assisted sensitivity analysis for treatment effects under unmeasured confounding via regularized calibrated estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *86*(5), 1339–1363.
- Torgovitsky, A. (2019). Partial identification by extending subdistributions. *Quantitative Economics*, *10*(1), 105–144.
- Van de Sijpe, N., & Windmeijer, F. (2023). On the power of the conditional likelihood ratio and related tests for weak-instrument robust inference. *Journal of Econometrics*, *235*(1), 82–104.
- van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge University Press.
- van der Vaart, A., & Wellner, J. (1996). *Weak convergence and empirical processes*. Springer.
- van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.
- VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the e-value. *Annals of Internal Medicine*, *167*(4), 268–274.
- White, H. (1999). *Asymptotic theory for econometricians*. Academic press.

- Williamson, R. C., & Downs, T. (1990). Probabilistic arithmetic. i. numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4(2), 89–158.
- Zhao, Q., Small, D. S., & Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4), 735–761.