UNIVERSITY OF CALIFORNIA

Los Angeles

Essays in Econometrics

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Economics

by

Daniel Steven Ober-Reynolds

2024

ABSTRACT OF THE DISSERTATION

Essays in Econometrics

by

Daniel Steven Ober-Reynolds
Doctor of Philosophy in Economics
University of California, Los Angeles, 2024
Professor Andres Santos, Chair

This dissertation contains two chapters. The first chapter studies causal parameters that depend on a moment of the joint distribution of potential outcomes. Such parameters are especially relevant in policy evaluation settings, where noncompliance is common and accommodated through the model of Imbens & Angrist (1994). The sharp identified set for these parameters is an interval with endpoints characterized by the value of optimal transport problems. Sample analogue estimators are proposed based on the dual problem of optimal transport. These estimators are $\sqrt{n}$-consistent and converge in distribution under mild assumptions. Inference procedures based on the bootstrap are straightforward and computationally convenient. The ideas and estimators are demonstrated in an application revisiting the National Supported Work Demonstration job training program. Estimates suggest that workers who would see below average earnings without treatment tend to see above average benefits from treatment.

The second chapter proposes a methodology for studying the robustness of results drawn from incomplete datasets. Selection is measured as the squared Hellinger divergence between the distributions of complete and incomplete observations, which has a natural interpreta-

tion. The *breakdown point* is defined as the minimal amount of selection needed to overturn a given result. Reporting point estimates and lower confidence intervals of the breakdown point is a simple, concise way to communicate a result's robustness. An estimator of the breakdown point of results drawn from GMM models is proposed and shown $\sqrt{n}$-consistent and asymptotically normal under mild assumptions. Lower confidence intervals of the breakdown point are simple to construct. The chapter concludes with a simulation study illustrating the good finite sample performance of the procedure.

The dissertation of Daniel Steven Ober-Reynolds is approved.

Denis Nikolaye Chetverikov

Jinyong Hahn

Rosa Liliana Matzkin

Rodrigo Ribeiro Antunes Pinto

Andres Santos, Committee Chair

University of California, Los Angeles

2024

To Autumn, my family, my friends, and my teachers,
who help me through all the hard work
and make the hard work worth it.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to thank my advisor, Andres Santos. I learned so much by being his student, not only from his conscientious approach to research and teaching, but also from the kindness he shows everyone around him. I will always appreciate the support he gave me throughout these years. I also want to thank Denis Chetverikov. His thoughts on research and openness in discussing life decisions helped me through a very challenging time. Jinyong Hahn, Rosa Matzkin, and Rodrigo Pinto all offered me many helpful comments as they served on my committee, and I'm very grateful. My research also benefited from suggestions by Zhipeng Liao and Shuyang Sheng, and participants in the UCLA econometrics proseminar.

I had the pleasure of doing this PhD with wonderful friends, including Leo Shi, Kirill Ponomarev, Lucas Zhang, Nano Palleja, Fatih Ozturk, Nicole Gorton Caratelli, Daniel Perez, and Calvin Kuo. I especially want to thank my friend Manu Navjeevan, who helped me understand econometrics and life during countless weekly runs.

I have to include a few sentences here about friends, peers, colleagues, teachers, and mentors from elementary, middle, and high school, as well as Arizona State University and the Federal Reserve Bank of Richmond. I would name you all individually but this would take many pages, as many wonderful people have supported me through this journey. Besides, you know who you are.

Last but not least, I would not have gone to graduate school without the encouragement of my parents, Sharman and Steven, and my brothers, Andrew and Ben. I would not have made it through graduate school without the love and support of Autumn Moroney. I will always be grateful.

VITA

2012–2016	B.S. Economics, B.A. Mathematics, B.A. Philosophy, *Summa Cum Laude*, Arizona State University
2016–2018	Research Associate, Federal Reserve Bank of Richmond
2020	M.A. in Economics, University of California, Los Angeles
2019–2023	Research Assistant and Teaching Assistant, University of California, Los Angeles

CHAPTER 1

Estimating Functionals of the Joint Distribution of Potential Outcomes with Optimal Transport

1.1 Introduction

Researchers studying the causal effects of a binary treatment see an observation's treated or untreated outcome, but never both. As a result, the data identify the marginal distributions of each potential outcome, but not their joint distribution. This "fundamental problem of causal inference" (Holland, 1986) leaves parameters depending on the joint distribution partially identified.

This paper studies a wide class of parameters that depend on a moment of the joint distribution of potential outcomes. The setting is the canonical potential outcomes framework with binary treatment, a binary instrument satisfying a monotonicity restriction, and finitely supported covariates (Imbens & Angrist, 1994; Abadie, 2003). In this setting, the sharp identified set for such parameters is an interval with endpoints characterized by the value of optimal transport problems. Sample analogue estimators based on the dual problem of optimal transport are tractable, both for computation and asymptotic analysis. These estimators are shown to converge in distribution through the functional delta method. This allows for straightforward inference procedures based on the bootstrap.

The proposed estimators are especially attractive due to their wide applicability and computational simplicity. The class of parameters under study is broad, including the correlation between potential outcomes, the probability of benefitting from treatment, and many

more examples discussed in section 1.2. As argued in Heckman et al. (1997), such parameters are of particular interest to policymakers and economists carrying out econometric policy evaluation. Noncompliance with the assigned treatment status is common in these settings. Most studies accomodate noncompliance with the same framework adopted in this paper, and could make use of these estimators with no additional identifying assumptions. Computing the estimator and constructing confidence sets entails nothing more challenging than solving linear programming problems, for which there are fast and efficient algorithms readily available.

This paper contributes to a large econometrics literature studying parameters of the joint distribution of potential outcomes. Many papers in this literature focus on a subset of the parameters considered here, especially the cumulative distribution function (cdf) or quantiles of treatment effects (Manski, 1997; Heckman et al., 1997; Firpo, 2007; Fan & Park, 2010, 2012; Firpo & Ridder, 2019; Callaway, 2021; Frandsen & Lefgren, 2021). This limited focus allows greater use of known analytical expressions when deriving sharp bounds, especially the famed Makarov bounds on the cdf and Fréchet-Hoeffding bounds on the joint distribution. Several recent works develop methods applicable to broad parameter classes by employing procedures that do not require analytical expressions for the identified set. Russell (2021) studies continuous functionals of the joint distribution of discrete potential outcomes, through a computationally intensive (sometimes infeasible) search over all permissible distributions of model primitives. Fan et al. (2023) study parameters identified through moment conditions in several incomplete data settings – including potential outcome models – by searching over an infinite dimensional space of smooth copulas. This paper occupies a middle ground: by focusing on parameters that depend on a scalar moment of the joint distribution and working with optimal transport, I obtain expressions for the bounds with tractable sample analogues. This approach allows consideration of a wide variety of parameters while maintaining computational tractability.

This paper also contributes to a growing literature on applications of optimal transport

to econometrics; see Galichon (2017) for a survey. Several recent working papers utilize optimal transport for issues related to casual inference, including inverse propensity weighting (Dunipace, 2021), matching on covariates (Gunsilius & Xu, 2021), and obtaining counterfactual distributions (Torous et al., 2021). In concurrent and highly complementary work, Ji et al. (2023) consider a very similar class of parameters to the present paper and also propose inference based on the dual problem of optimal transport. Their focus, accomodating non-discrete covariates without resorting to parametric models, leads to theory based on cross fitting and high-level assumptions on first stage estimators. The goal of the present paper is to provide simple, low-level conditions and computationally convenient estimators in the common case where covariates are discrete. This leads to theory based on Hadamard directional differentiability and the functional delta method quite distinct from that of Ji et al. (2023).

The remainder of this paper is organized as follows. Section 1.2 formalizes the setting and introduces the class of parameters under study. Optimal transport is introduced in section 1.3, and used in identification in section 1.4. Section 1.5 proposes the estimators and contains the asymptotic results. Section 1.6 explores the finite sample properties of the estimators in a brief simulation study. Section 1.7 contains the application, showing suggestive evidence that the the National Supported Work Demonstration job training program was especially beneficial for workers who would see below average incomes without training. All formal results are proven in the appendix.

1.2 Setting and parameter class

1.2.1 Setting

Consider a potential outcomes framework with binary treatment, a binary instrument, and finitely supported covariates (Imbens & Angrist, 1994; Abadie, 2003). Let Y denote the scalar, real-valued outcome of interest and $D \in \{0, 1\}$ indicate treatment status. Further let

Y_1 denote the potential outcome when treated and Y_0 the potential outcome when untreated. The observed outcome Y is given by

$$Y = DY_1 + (1 - D)Y_0. \tag{1.1}$$

The difference in potential outcomes, $Y_1 - Y_0$, is called the treatment effect.

The binary instrument is denoted $Z \in \{0, 1\}$. Let D_1 denote the treatment status when $Z = 1$, and D_0 the treatment status when $Z = 0$. The observed treatment status D is given by

$$D = ZD_1 + (1 - Z)D_0. \tag{1.2}$$

It is assumed that the instrument itself does not affect the outcome.[1] Units with $1 = D_1 > D_0 = 0$ are known as *compliers*.

Assumption 1 formalizes the setting.

Assumption 1 (Setting)**.** *$\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$ is an i.i.d. sample with $(Y, D, Z, X) \sim P$,*

$$Y \in \mathcal{Y} \subseteq \mathbb{R}, \qquad D \in \{0, 1\}, \qquad Z \in \{0, 1\}, \qquad X \in \mathcal{X} = \{x_1, \ldots, x_M\} \subseteq \mathbb{R}^{d_x} \tag{1.3}$$

where Y, D, and Z are related to (Y_1, Y_0, D_1, D_0) through equations (1.1) *and* (1.2)*, and the random vector $(Y_1, Y_0, D_1, D_0, Z, X)$ satisfies*

(i) Instrument independence: $(Y_1, Y_0, D_1, D_0) \perp Z \mid X$,

(ii) Monotonicity: $P(D_1 \geq D_0) = 1$,

(iii) Existence of compliers: $P(D_1 > D_0, X = x) > 0$ for each x, and

(iv) $P(X = x, Z = z) > 0$ for each (x, z).

[1] One could hypothesize potential outcomes varying with the value of the instrument, i.e. Y_{dz} for each (d, z). The exposition here implicitly assumes *instrument exclusion*, also known as the *Stable Unit Treatment Value Assumption*: that $P(Y_{d1} = Y_{d0}) = 1$ for each d.

Assumption 1 is essentially equivalent to assumption 2.1 in Abadie (2003), with the addition that covariates are finitely supported. Instrument independence is sometimes referred to as *ignorability*, and satisfied in most randomized controlled trials where Z indicates being assigned to treatment. Monotonicity is typically a weak assumption in such settings.

It is worth emphasizing that this setting nests the case where treatment is exogenous. Specifically, when $D_1 = 1$ and $D_0 = 0$ (degenerately), every unit is a complier. In this case equation (1.2) shows treatment status equals the instrument: $D = Z$. Instrument independence simplifies to $(Y_1, Y_0) \perp D \mid X$, and monotonicity is trivially satisfied.

Interest focuses on the distribution of compliers. Such focus is especially policy relevant when "the policy is the instrument" i.e., the proposed change in policy is to assign $Z = 1$ to all units. Abadie (2003) shows that assumption 1 suffices to identify the marginal distributions of Y_1 and Y_0 for the subpopulation of compliers.

Lemma 1.2.1 (Abadie (2003))**.** *Suppose assumption 1 holds. Then the marginal distributions of Y_d conditional on $D_1 > D_0$ and $X = x$, denoted $P_{d|x}$, are identified by*

$$\begin{aligned} E_{P_{d|x}}[f(Y_d)] &\equiv E[f(Y_d) \mid D_1 > D_0, X = x] \\ &= \frac{E[f(Y)\mathbb{1}\{D = d\} \mid Z = d, X = x] - E[f(Y)\mathbb{1}\{D = d\} \mid Z = 1 - d, X = x]}{P(D = d \mid Z = d, X = x) - P(D = d \mid Z = 1 - d, X = x)} \end{aligned} \tag{1.4}$$

for any integrable function f. Furthermore, the distribution of X conditional on $D_1 > D_0$ is identified by

$$\begin{aligned} s_x &\equiv P(X = x \mid D_1 > D_0) \\ &= \frac{[P(D = 1 \mid Z = 1, X = x) - P(D = 1 \mid Z = 0, X = x)]\, P(X = x)}{\sum_{x'} [P(D = 1 \mid Z = 1, X = x') - P(D = 1 \mid Z = 0, X = x')]\, P(X = x')} \end{aligned} \tag{1.5}$$

The joint distribution of potential outcomes is not identified. This is a result of the fundamental problem of causal inference: there is no unit where both Y_1 and Y_0 are observed,

and as a result the joint distribution of (Y_1, Y_0) is not identified for any subpopulation. Let $P_{1,0}$ denote the joint distribution of (Y_1, Y_0) conditional on compliance, and $P_{1,0|x}$ denote the joint distribution conditional on compliance and $X = x$. These are related through the law of iterated expectations; for any function $c(y_1, y_0)$ with values in $\mathbb{R}$,

$$E_{P_{1,0}}[c(Y_1, Y_0)] = E[E[c(Y_1, Y_0) \mid D_1 > D_0, X] \mid D_1 > D_0] = \sum_x s_x E_{P_{1,0|x}}[c(Y_1, Y_0)].$$

This relation can also be expressed as $P_{1,0} = \sum_x s_x P_{1,0|x}$.

A joint distribution with marginals $P_{1|x}$ and $P_{0|x}$ is called a *coupling* of $P_{1|x}$ and $P_{0|x}$. $P_{1,0|x}$ is such a coupling, and is otherwise unrestricted by assumption 1. It follows that the identified set for $P_{1,0|x}$ is the set of distributions $\pi_{1,0|x}$ for (Y_1, Y_0) with marginals $\pi_{1|x} = P_{1|x}$ and $\pi_{0|x} = P_{0|x}$, denoted

$$\Pi(P_{1|x}, P_{0|x}) \equiv \left\{ \pi_{1,0|x} \ : \ \pi_{1|x} = P_{1|x}, \ \pi_{0|x} = P_{0|x} \right\}. \tag{1.6}$$

Moreover, the identified set for $P_{1,0}$ is $\left\{ \pi_{1,0} = \sum_x s_x \pi_{1,0|x} \ : \ \pi_{1,0|x} \in \Pi(P_{1|x}, P_{0|x}) \right\}$.

1.2.2 Parameter class

The idea at the core of this paper is to bound a moment of the joint distribution of potential outcomes by optimization. Accordingly, the focus is on scalar parameters of the form

$$\gamma \equiv g(\theta, \eta) \tag{1.7}$$

where g is a known function and $\theta \equiv E_{P_{1,0}}[c(Y_1, Y_0)] \in \mathbb{R}$ is a scalar moment of the joint distribution of (Y_1, Y_0) conditional on compliance. The function c is known, and referred to as a *cost function* in connection with the optimal transport literature. This class of parameters is broad, as illustrated by the examples given below. In each of these examples η is a finite collection of moments of the marginal distributions conditional on compliance:

$\eta = (E_{P_1}[\eta_1(Y_1)], E_{P_0}[\eta_0(Y_0)]) \in \mathbb{R}^{K_1+K_0}$. The formal results focus on this case, but could be generalized to allow η to be other point identified nuisance parameters.

The following conditions are stronger than necessary for identification of the sharp identified set of γ, but will be used when constructing and studying estimators. Assumption 2 places restrictions on the cost function to ensure optimal transport can be used characterize and estimate the sharp identified set for θ.

Assumption 2 (Cost function)**.** *Either*

(i) $c(y_1, y_0)$ *is Lipschitz continuous and* $\mathcal{Y}$ *is compact, or*

(ii) $c(y_1, y_0) = \mathbb{1}\{y_1 - y_0 \leq \delta\}$ *for a known* $\delta \in \mathbb{R}$ *and the cumulative distribution functions* $F_{d|x}(y) = P(Y_d \leq y \mid D_1 > D_0, X = x)$ *are continuous.*

Assumption 2 covers every example listed below. Continuous cost functions c are given a unified analysis, but for reasons discussed in section 1.3 discontinuous cost functions must be handled on a case-by-case basis. I focus on the leading case of interest in applications, $c(y_1, y_0) = \mathbb{1}\{y_1 - y_0 \leq \delta\}$, corresponding to the cumulative distribution of treatment effects. The approach developed in this paper could likely be generalized to cover other discontinuous cost functions; for example, results in the appendix allow estimation of the sharp lower bound of $P((Y_1, Y_0) \in C)$ for any open, convex set $C \subseteq \mathbb{R}^2$.

Assumption 2 (ii) requires the cdfs $F_{d|x}$ be continuous. As discussed in section 1.4, this ensures the set being estimated is the sharp identified set for the parameter of interest. However, the estimation and inference results of section 1.5 hold *regardless* of whether the cdfs are continuous or not; when the cdfs are not continuous, the estimand is a valid outer identified set.

Under assumptions 1 and 2, the sharp identified set for $\theta = E[c(Y_1, Y_0) \mid D_1 > D_0]$ is an interval $[\theta^L, \theta^H]$. Assumption 3 contains conditions on g and η.

Assumption 3 (Function of moments)**.** *The parameter is* $\gamma = g(\theta, \eta) \in \mathbb{R}$*, where*

$$\theta = E[c(Y_1, Y_0) \mid D_1 > D_0] \in \mathbb{R}, \qquad \eta = E\left[\eta_1(Y_1), \eta_0(Y_0) \mid D_1 > D_0\right] \in \mathbb{R}^{K_1+K_0}$$

for known functions g*,* c*,* η_1 *and* η_0 *such that*

(i) $E[\|\eta_d(Y)\|^2] < \infty$ *for* $d = 1, 0$*,*

(ii) $g(\cdot, \eta)$ *is continuous, and*

(iii) the functions

$$g^L(t^L, t^H, e) = \min_{t \in [t^L, t^H]} g(t, e), \qquad g^H(t^L, t^H, e) = \max_{t \in [t^L, t^H]} g(t, e)$$

are continuously differentiable at $(t^L, t^H, e) = (\theta^L, \theta^H, \eta)$*.*

Note that when θ itself is of interest, assumption 3 is satisfied with $g(\theta, \eta) = \theta$. Assumption 3 (ii) ensures the identified set for γ is the interval $[\gamma^L, \gamma^H]$, and assumption 3 (iii) is used to apply the delta method. It is straightforward to show assumption 3 (iii) holds when g is continuously differentiable in both arguments and $g(\cdot, \eta)$ is strictly increasing, as the latter condition implies $g^L(\theta^L, \theta^H, \eta) = g(\theta^L, \eta)$ and $g^H(\theta^L, \theta^H, \eta) = g(\theta^H, \eta)$ and the former condition implies they are continuously differentiable. This argument applies to every parameter listed below. When g is differentiable but $g(\cdot, \eta)$ is not monotonic, it is often possible to use the implicit function theorem applied to first order conditions to derive sufficient conditions for the corresponding arg min and arg max to be differentiable, and thus for assumption 3 (iii) to hold.

1.2.2.1 Examples

The following examples are intended both to fix ideas and illustrate the broad scope of the parameter class described above.

Example 1.2.1 (Summary statistics)**.** *Many summary statistics can be rewritten in the form* $\gamma = g(\theta, \eta)$*. For example, suppose interest is in the variance of treatment effects for compliers:* $\gamma = Var(Y_1 - Y_0 \mid D_1 > D_0)$*. This parameter can be rewritten as*

$$\gamma = Var(Y_1 - Y_0 \mid D_1 > D_0) = E_{P_{1,0}}[(Y_1 - Y_0)^2] - (E_{P_1}[Y_1] - E_{P_0}[Y_0])^2,$$

This parameter fits the form $\gamma = g(\theta, \eta)$ *required of display* (1.7)*, with* $\theta = E_{P_{1,0}}[(Y_1 - Y_0)^2]$*,* $\eta = (\eta^{(1)}, \eta^{(2)}) = (E_{P_1}[Y_1], E_{P_0}[Y_0])$*, and* $g(\theta, \eta) = \theta - (\eta^{(1)} - \eta^{(2)})^2$*. The cost function* $c(y_1, y_0) = (y_1 - y_0)^2$ *satisfies assumption 2 (i) when* $\mathcal{Y}$*, the support of the outcome* Y*, is bounded.*

Similarly, suppose the researcher is interested in the correlation between Y_1 *and* Y_0 *for compliers. Set* $\gamma = Corr(Y_1, Y_0 \mid D_1 > D_0)$*, which can be rewritten as*

$$\gamma = Corr(Y_1, Y_0 \mid D_1 > D_0) = \frac{E_{P_{1,0}}[Y_1 Y_0] - E_{P_1}[Y_1]E_{P_0}[Y_0]}{\sqrt{E_{P_1}[Y_1^2] - (E_{P_1}[Y_1])^2}\sqrt{E_{P_0}[Y_0^2] - (E_{P_0}[Y_0])^2}}$$

This parameter also fits the form $\gamma = g(\theta, \eta)$ *in display* (1.7)*, with* $\theta = E_{P_{1,0}}[Y_1 Y_0]$*,* $\eta = (\eta^{(1)}, \eta^{(2)}, \eta^{(3)}, \eta^{(4)}) = (E_{P_1}[Y_1], E_{P_1}[Y_1^2], E_{P_0}[Y_0], E_{P_0}[Y_0^2])$*, and* $g(\theta, \eta) = \frac{\theta - \eta^{(1)} \times \eta^{(3)}}{\sqrt{\eta^{(2)} - (\eta^{(1)})^2}\sqrt{\eta^{(4)} - (\eta^{(3)})^2}}$*. The cost function* $c(y_1, y_0) = y_1 y_0$ *satisfies assumption 2 (i) when* $\mathcal{Y}$ *is bounded.*

Example 1.2.2 (Expected percent change)**.** *The expected percent change in the outcome can be written as* $100 \times E\left[\frac{Y_1 - Y_0}{Y_0} \mid D_1 > D_0\right]\%$*. This is a unit-invariant causal parameter that is a natural summary measure when* Y_0 *exhibits considerable variation. For example, a treatment effect of* $Y_1 - Y_0 = 5$ *is typically of greater economic significance when the untreated outcome is small, say* $Y_0 = 10$*, than when* $Y_0 = 100$*.*

The expected percent change is proportional to

$$\gamma = E\left[\frac{Y_1 - Y_0}{Y_0} \mid D_1 > D_0\right] = E_{P_{1,0}}\left[\frac{Y_1 - Y_0}{Y_0}\right],$$

which fits the form of display (1.7)*, with* $\gamma = \theta = E_{P_{1,0}}\left[\frac{Y_1 - Y_0}{Y_0}\right]$*. The cost function* $c(y_1, y_0) = \frac{y_1 - y_0}{y_0}$ *satisfies assumption 2 (i) when* $\mathcal{Y}$ *is bounded and bounded away from zero.*

Example 1.2.3 (Equitable policies)**.** *Policy makers are often interested in whether a policy is equitable – that is, whether the benefits are concentrated among those who would have undesirable outcomes without treatment.*

One parameter that speaks to these concerns is the covariance between treatment effects and untreated outcomes among compliers: $\gamma = Cov(Y_1 - Y_0, Y_0 \mid D_1 > D_0)$*. Notice that* $\gamma < 0$ *implies those with below average* Y_0 *tend to see above average treatment effects. This parameter can be rewritten as*

$$\gamma = Cov(Y_1 - Y_0, Y_0 \mid D_1 > D_0) = E_{P_{1,0}}[(Y_1 - Y_0)Y_0] - (E_{P_1}[Y_1] - E_{P_0}[Y_0])E_{P_0}[Y_0]$$

and fits the form $g(\theta, \eta)$ *with* $\theta = E_{P_{1,0}}[(Y_1 - Y_0)Y_0]$*,* $\eta = (E_{P_1}[Y_1], E_{P_0}[Y_0])$*, and* $g(\theta, \eta) = \theta - (\eta^{(1)} - \eta^{(2)})\eta^{(2)}$*. The cost function* $c(y_1, y_0) = (y_1 - y_0)y_0$ *satisfies assumpion 2 (i) when* $\mathcal{Y}$ *is bounded.*

Many related parameters share a sign with $Cov(Y_1 - Y_0, Y_0 \mid D_1 > D_0)$ *and are also suitable for such an analysis. For example, consider the OLS slope when regressing* $Y_1 - Y_0$ *on* Y_0 *and a constant:* $\gamma = \frac{Cov(Y_1 - Y_0, Y_0 | D_1 > D_0)}{Var(Y_0 | D_1 > D_0)}$*. This parameter can be rewritten as*

$$\gamma = \frac{Cov(Y_1 - Y_0, Y_0 \mid D_1 > D_0)}{Var(Y_0 \mid D_1 > D_0)} = \frac{E_{P_{1,0}}[(Y_1 - Y_0)Y_0] - (E_{P_1}[Y_1] - E_{P_0}[Y_0])E_{P_0}[Y_0]}{E_{P_0}[Y_0^2] - (E_{P_0}[Y_0])^2},$$

which fits the form of display (1.7) *with* $\theta = E_{P_{1,0}}[(Y_1 - Y_0)Y_0]$*,* $\eta = (E_{P_1}[Y_1], E_{P_0}[Y_0], E_{P_0}[Y_0^2])$*, and* $g(\theta, \eta) = \frac{\theta - (\eta^{(1)} - \eta^{(2)})\eta^{(2)}}{\eta^{(3)} - (\eta^{(2)})^2}$*.*

Example 1.2.4 (Proportion that benefit)**.** *The share of compliers benefiting from treatment, written*

$$\gamma = P(Y_1 > Y_0 \mid D_1 > D_0),$$

is naturally of interest in applications where theory gives little indication whether the treatment will have a positive or negative effect. For example, Allcott et al. (2020) study the effect of deactivating facebook on subjective well-being. The authors find significant positive average effects of deactivation, but find substantial heterogeneity in follow-up interviews.

This parameter fits the form of display (1.7)*, with* $\gamma = \theta = E_{P_{1,0}}[\mathbb{1}\{Y_1 - Y_0 \leq 0\}]$*. The cost function* $c(y_1, y_0) = \mathbb{1}\{y_1 - y_0 \leq 0\}$ *satisfies assumption 2 (ii) if the cdfs* $F_{d|x}(y)$ *are continuous.*

The share benefiting from treatment is also of particular interest when the intervention comes at a financial cost and the outcome of interest is a pecuniary return. Examples include job training programs intended to increase a worker's income (e.g. the National Supported Work Demonstration studied in Couch (1992)) or management practices intended to raise a firm's accounting profit (e.g. the employee referral program studied in Friebel et al. (2023)). To illustrate, suppose the researcher observes $\{R_i, C_i, D_i, Z_i\}_{i=1}^n$*, where* R *is observed revenue and* C *is the observed cost. These are related to treatment status* $D \in \{0, 1\}$*, potential revenues* (R_1, R_0)*, and potential costs* (C_1, C_0) *by*

$$R = DR_1 + (1 - D)R_0, \qquad C = DC_1 + (1 - D)C_0.$$

The observed profit, $Y = R - C$*, is related to treatment status by*

$$Y = D\underbrace{(R_1 - C_1)}_{\equiv Y_1} + (1 - D)\underbrace{(R_0 - C_0)}_{\equiv Y_0}.$$

The probability the change in revenue exceeds the change in cost is

$$P(R_1 - R_0 > C_1 - C_0 \mid D_1 > D_0) = P(Y_1 > Y_0 \mid D_1 > D_0).$$

Example 1.2.5 (Quantiles)**.** *Suppose the parameter of interest is any q_τ solving*

$$P(Y_1 - Y_0 \leq q_\tau \mid D_1 > D_0) = \tau \tag{1.8}$$

This parameter has a similar interpretation to the τ-th quantile.[2] *q_τ cannot be viewed as $\gamma = g(\theta, \eta)$. However, by viewing $\theta(\delta) = P(Y_1 - Y_0 \leq \delta \mid D_1 > D_0) = E_{P_{1,0}}[\mathbb{1}\{Y_1 - Y_0 \leq \delta\}]$ as a function of δ, the results below can be adapted to construct a confidence set for the identified set of this parameter as described in appendix 1.9.1.2.*

1.3 Optimal transport

This section defines and discusses optimal transport, which is used to characterize the identified set and construct estimators. Given any marginal distributions P_1 and P_0 and a cost function $c(y_1, y_0)$, the Monge-Kantorovich formulation of *optimal transport* is the problem of choosing a coupling $\pi \in \Pi(P_1, P_0)$ to minimize the expected cost:

$$OT_c(P_1, P_0) \equiv \inf_{\pi \in \Pi(P_1, P_0)} E_\pi[c(Y_1, Y_0)]. \tag{1.9}$$

This minimization problem in (1.9) is referred to as the *primal problem*, and will be used to characterize the identified set of θ.

The dual problem of optimal transport will be used to construct and analyze estimators. Let Φ_c denote the set of functions $\varphi(y_1)$ and $\psi(y_0)$ whose pointwise sum is less than $c(y_1, y_0)$:

$$\Phi_c \equiv \{(\varphi, \psi)\ ;\ \varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)\}. \tag{1.10}$$

The *dual problem* chooses a pair of functions in Φ_c to maximize the sum of the corresponding

[2]The τ-th quantile is usually defined as the unique value $\tilde{q}_\tau = \inf\{y\ ;\ P(Y_1 - Y_0 \leq y) \geq \tau\}$. When the τ level set of the cumulative distribution function $P(Y_1 - Y_0 \leq \cdot)$ is nonempty, the τ-th quantile has the interpretation that $100 \times \tau\%$ of the population has treatment effect less than or equal to $\tilde{q}_\tau$. Every q_τ solving (1.8) has the same interpretation.

expectations:

$$\sup_{(\varphi,\psi)\in\Phi_c} E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)]. \tag{1.11}$$

When the cost function is lower semicontinuous and bounded from below, the primal problem is attained and *strong duality* holds:

$$OT_c(P_1, P_0) = \min_{\pi\in\Pi(P_1,P_0)} E_\pi[c(Y_1, Y_0)] = \sup_{(\varphi,\psi)\in\Phi_c} E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)]. \tag{1.12}$$

The dual problem is used to construct and analyze estimators. Indeed, the identification of $P_{d|x}$ in lemma 1.2.1 suggests straightforward sample analogues of $E_{P_{d|x}}[f(Y_d)]$ for a given f, which makes it possible to form a sample analogue of the dual problem in a setting with instruments.

Although it is clear how to form a sample analogue of the dual problem, it is not immediately clear how to analyze the resulting estimator. Fortunately, the dual problem can be simplified by restricting the maximization problem to a smaller set of functions. Estimators based on this restricted dual problem can then be studied with empirical process techniques.

The feasible set of the dual problem is restricted with the concept of c-concavity. Notice the dual problem's objective is monotonic, in the sense that $\varphi(y_1) \leq \tilde{\varphi}(y_1)$ for all y_1 implies

$$E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)] \leq E_{P_1}[\tilde{\varphi}(Y_1)] + E_{P_0}[\psi(Y_0)].$$

Increasing ψ pointwise will also increase the dual objective. Speaking loosely, any function pair $(\varphi, \psi) \in \Phi_c$ for which the constraint $\varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)$ is "slack" cannot be a solution to the dual problem and can therefore be ignored. This motivates the definition of the *c-transforms* of a function φ:

$$\varphi^c(y_0) \equiv \inf_{y_1}\{c(y_1, y_0) - \varphi(y_1)\}, \qquad \varphi^{cc}(y_1) \equiv \inf_{y_0}\{c(y_1, y_0) - \varphi^c(y_0)\}.$$

For any pair of functions $(\varphi, \psi) \in \Phi_c$, these definitions imply $\psi(y_0) \leq \varphi^c(y_0)$, $\varphi(y_1) \leq \varphi^{cc}(y_1)$,

and $\varphi^{cc}(y_1)+\varphi^c(y_0) \leq c(y_1, y_0)$. Further c-transformations are irrelevant because $(\varphi^{cc})^c = \varphi^c$, so a function φ is called *c-concave* if $\varphi^{cc} = \varphi$. If the c-transforms are integrable, the dual problem can be restricted to c-concave conjugate pairs, $(\varphi^{cc}, \varphi^c)$. c-concave functions often "inherit" properties of the cost function c. For example, if c is Lipschitz continuous then φ^c and φ^{cc} are Lipschitz continuous as well. These properties can be used to define sets of functions $\mathcal{F}_c$ and $\mathcal{F}_c^c$ (depending on the cost function c but not on the distributions P_1, P_0) such that

$$\sup_{(\varphi,\psi)\in\Phi_c} E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)] = \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)]. \tag{1.13}$$

Two cases suffice for the parameters considered in this paper. When the cost function $c(y_1, y_0)$ is Lipschitz continuous and $\mathcal{Y}$ is compact, define

$$\mathcal{F}_c \equiv \{\varphi : \mathcal{Y} \to \mathbb{R} \,;\; -\|c\|_\infty \leq \varphi(y_1) \leq \|c\|_\infty,\; |\varphi(y_1) - \varphi(y_1')| \leq L|y_1 - y_1'|\} \tag{1.14}$$

$$\mathcal{F}_c^c \equiv \{\psi : \mathcal{Y} \to \mathbb{R} \,;\; -2\|c\|_\infty \leq \psi(y_0) \leq 0,\; |\psi(y_0) - \psi(y_0')| \leq L|y_0 - y_0'|\} \tag{1.15}$$

where $\|c\|_\infty = \sup_{(y_1,y_0)}|c(y_1, y_0)|$ and L is the Lipschitz constant of c. When $c(y_1, y_0) = \mathbb{1}\{(y_1, y_0) \in C\}$ for an open, convex set C, let

$$\mathcal{F}_c \equiv \{\varphi : \mathcal{Y} \to \mathbb{R} \,;\; \varphi(y_1) = \mathbb{1}\{y_1 \in I\} \text{ for some interval } I\} \tag{1.16}$$

$$\mathcal{F}_c^c \equiv \{\psi : \mathcal{Y} \to \mathbb{R} \,;\; \psi(y_0) = -\mathbb{1}\{y_0 \in I^c\} \text{ for some interval } I\} \tag{1.17}$$

Equation (1.13) shows the optimal transport functional $OT_c(P_1, P_0)$ depends only on the values of $E_{P_1}[\varphi(Y_1)]$ and $E_{P_0}[\psi(Y_0)]$ for $(\varphi, \psi) \in \mathcal{F}_c \times \mathcal{F}_c^c$. For any set A, let $\ell^\infty(A)$ denote the space of real-valued bounded functions defined on A, equipped with the supremum norm: $\ell^\infty(A) = \{f : A \to \mathbb{R} \,;\; \|f\|_\infty = \sup_{a\in A}|f(a)| < \infty\}$. Optimal transport can be viewed as the

map $OT_c : \ell^\infty(\mathcal{F}_c) \times \ell^\infty(\mathcal{F}_c^c) \to \mathbb{R}$ given by

$$OT_c(P_1, P_0) = \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} E_{P_1}[\varphi(Y_1)] + E_{P_0}[\psi(Y_0)]. \tag{1.18}$$

This problem will be referred to as the *restricted dual problem*. Estimators formed with this map can be studied with empirical process techniques.

In summary, $OT_c(P_1, P_0)$ will be viewed as the functional in (1.9) when considering identification, and as the functional given in (1.18) when considering estimation. By ensuring c is either Lipschitz continuous or the indicator of an open convex set, strong duality and c-concavity ensures these functionals agree on the space of probability distributions.

1.4 Identification

This section derives expressions for the sharp bounds on the parameter of interest that make use of optimal transport. Recall the parameter of interest is $\gamma = g(\theta, \eta)$, where η is a point identified parameter, $\theta = E[c(Y_1, Y_0) \mid D_1 > D_0]$ is a scalar, and g and c are known functions.

Begin by rewriting θ with the law of iterated expectations:

$$\theta = E[E[c(Y_1, Y_0) \mid D_1 > D_0, X] \mid D_1 > D_0] = E[\theta_X \mid D_1 > D_0]$$

where $\theta_x \equiv E[c(Y_1, Y_0) \mid D_1 > D_0, X = x] = E_{P_{1,0|x}}[c(Y_1, Y_0)]$. As noted at the end of section 1.2.1, the identified set for $P_{1,0|x}$ is the set of couplings of $P_{1|x}$ and $P_{0|x}$, denoted $\Pi(P_{1|x}, P_{0|x})$. It follows that the identified set for θ_x is the set of values that can be expressed as $E_\pi[c(Y_1, Y_0)]$ for some $\pi \in \Pi(P_{1|x}, P_{0|x})$. The set $\Pi(P_{1|x}, P_{0|x})$ is convex, implying that the identified set for θ_x is an interval. Let θ_x^L and θ_x^H denote its lower and upper endpoint respectively.

To ensure the restricted dual problem can be used for estimation, θ_x^L and θ_x^H are characterized through an optimal transport problem with a suitable cost function. When assumption

2 (i) holds ($c(y_1, y_0)$ is Lipschitz continuous and $\mathcal{Y}$ is compact), define

$$\begin{aligned} c_L(y_1, y_0) &\equiv c(y_1, y_0), & c_H(y_1, y_0) &\equiv -c(y_1, y_0) \\ \theta^L(P_{1|x}, P_{0|x}) &\equiv OT_{c_L}(P_{1|x}, P_{0|x}), & \theta^H(P_{1|x}, P_{0|x}) &\equiv -OT_{c_H}(P_{1|x}, P_{0|x}). \end{aligned} \tag{1.19}$$

Note that $\theta_x^L = \theta^L(P_{1|x}, P_{0|x})$ and $\theta_x^H = \theta^H(P_{1|x}, P_{0|x})$.

The cumulative distribution function of $Y_1 - Y_0$ corresponds to the cost function $c(y_1, y_0) = \mathbb{1}\{y_1 - y_0 \leq \delta\}$, which is not lower semicontinuous. This challenge is circumvented by a small change in the cost function. When assumption 2 (ii) holds (the cost function is $c(y_1, y_0) = \mathbb{1}\{y_1 - y_0 \leq \delta\}$) define

$$\begin{aligned} c_L(y_1, y_0) &\equiv \mathbb{1}\{y_1 - y_0 < \delta\}, & c_H(y_1, y_0) &\equiv \mathbb{1}\{y_1 - y_0 > \delta\} \\ \theta^L(P_{1|x}, P_{0|x}) &\equiv OT_{c_L}(P_{1|x}, P_{0|x}), & \theta^H(P_{1|x}, P_{0|x}) &\equiv 1 - OT_{c_H}(P_{1|x}, P_{0|x}) \end{aligned} \tag{1.20}$$

It follows from definitions that $\theta_x^H = \theta^H(P_{1|x}, P_{0|x})$. Moreover, $c_L(y_1, y_0) \leq c(y_1, y_0)$ implies $\theta^L(P_{1|x}, P_{0|x})$ is a valid lower bound for θ_x. It is sharp if $P_{1|x}$, $P_{0|x}$ have continuous cumulative distribution functions, in which case $\theta_x^L = \theta^L(P_{1|x}, P_{0|x})$. It is worth emphasizing again that the estimation and inference results of section 1.5 ahead hold *regardless* of whether the cdfs are continuous or not; when the cdfs are not continuous, the estimand is a valid outer identified set.

Under assumptions 1 and 2, the identified set for $\theta = E_{P_{1,0}}[c(Y_1, Y_0)] = E[c(Y_1, Y_0) \mid D_1 > D_0]$ is the compact interval $[\theta^L, \theta^H]$ with endpoints

$$\theta^L = E[\theta_X^L \mid D_1 > D_0], \qquad \theta^H = E[\theta_X^H \mid D_1 > D_0]$$

Under assumptions 1, 2, and 3, the identified set for γ is $[\gamma^L, \gamma^H]$, with endpoints

$$\gamma^L = g^L(\theta^L, \theta^H, \eta) = \inf_{t \in [\theta^L, \theta^H]} g(t, \eta), \qquad \gamma^H = g^H(\theta^L, \theta^H, \eta) = \sup_{t \in [\theta^L, \theta^H]} g(t, \eta) \tag{1.21}$$

The following theorem summarizes the discussion above. Let $\theta^L(\cdot, \cdot)$ and $\theta^H(\cdot, \cdot)$ be given by (1.19) or (1.20) depending on the cost function, and set

$$\theta^L_x = \theta^L(P_{1|x}, P_{0|x}), \qquad \theta^H_x = \theta^H(P_{1|x}, P_{0|x}), \tag{1.22}$$

$$\theta^L = E[\theta^L_X \mid D_1 > D_0], \qquad \theta^H = E[\theta^H_X \mid D_1 > D_0], \tag{1.23}$$

$$\gamma^L = g^L(\theta^L, \theta^H, \eta), \qquad \gamma^H = g^H(\theta^L, \theta^H, \eta) \tag{1.24}$$

Theorem 1.4.1 (Identification of functions of moments)**.** *Suppose assumptions 1, 2, and 3 are satisfied. Then the sharp identified set for γ is $[\gamma^L, \gamma^H]$.*

All results are proven in the appendix.

It is worth pausing to consider the role of covariates. When covariates are available, ignoring them leads to wider bounds that are not sharp. Specifically, the marginal distributions P_1 and P_0 could be used to form a lower bound on θ with $\theta^L(P_1, P_0) = \inf_{\pi \in \Pi(P_1, P_0)} E_\pi[c_L(Y_1, Y_0)]$. This bound minimizes over the whole set $\Pi(P_1, P_0) = \{\pi_{1,0} \ ; \ \pi_1 = P_1, \pi_0 = P_0\}$, but the identified set for $P_{1,0}$ is the subset given by $\left\{\pi_{1,0} = \sum_x s_x \pi_{1,0|x} \ ; \ \pi_{1,0|x} \in \Pi(P_{1|x}, P_{0|x})\right\}$. The bounds defined by equations (1.22) and (1.23) is found while enforcing the additional constraints that $\pi_{1,0|x} \in \Pi(P_{1|x}, P_{0|x})$ for each x. These additional constraints imply $\theta^L(P_1, P_0) \leq \theta^L$, and similarly $\theta^H \leq \theta^H(P_1, P_0)$.

Extreme cases illustrate when covariates are informative. If X is independent of (Y_1, Y_0) conditional on $D_1 > D_0$, then $P_{d|x} = P_d$ for each x, $\Pi(P_{1|x}, P_{0|x}) = \Pi(P_1, P_0)$, and the inequalities in the preceding paragraph hold as equalities. In the other extreme, suppose $P_{1|x}$ or $P_{0|x}$ (or both) are degenerate. This would follow from Y_d being a function of X. When one or more of the distributions is degenerate, there is only one possible coupling.

Since $\Pi(P_{1|x}, P_{0|x})$ is a singleton, $\theta_x^L = \theta_x^H$ and $\theta_x = E[c(Y_1, Y_0) \mid D_1 > D_0, X = x]$ is point identified. If this occurs for all $x \in \mathcal{X}$, θ and γ are point identified.

Remark 1.4.1 (Makarov bounds). The proof of theorem 1.4.1 given in the appendix uses properties of optimal transport to argue that under assumptions 1 and 2 (ii), $[\theta^L, \theta^H]$ is the sharp identified set for $P(Y_1 - Y_0 \leq \delta \mid D_1 > D_0)$. Nonetheless, it is interesting to note that the proof shows

$$\theta_x^L = OT_{c_L}(P_{1|x}, P_{0|x}) = \sup_y \{F_{1|x}(y) - F_{0|x}(y - \delta)\}$$

$$\theta_x^H = 1 - OT_{c_H}(P_{1|x}, P_{0|x}) = 1 - \sup_y \{F_{0|x}(y - \delta) - F_{1|x}(y)\} = 1 + \inf_y \{F_{1|x}(y) - F_{0|x}(y - \delta)\}$$

which are the Makarov bounds on $P(Y_1 - Y_0 \leq \delta \mid D_1 > D_0, X = x)$ studied in Fan & Park (2010).

Remark 1.4.2 (Pointwise vs. uniformly sharp CDF bounds). Under assumptions 1 and 2 (ii), $[\theta^L, \theta^H]$ is the sharp identified set for $P(Y_1 - Y_0 \leq \delta \mid D_1 > D_0)$ at the *point* δ. Viewing these bounds as functions of δ, $\theta^L(\delta)$ and $\theta^H(\delta)$ are not *uniformly* sharp bounds for the cumulative distribution function $P(Y_1 - Y_0 \leq \delta \mid D_1 > D_0)$, in the sense that not every CDF $F(\cdot)$ satisfying $\theta^L(\delta) \leq F(\delta) \leq \theta^H(\delta)$ for all δ could be the CDF of $Y_1 - Y_0$. See Firpo & Ridder (2019) for a detailed discussion of this point.

1.5 Estimators

Sample analogues of the expressions identifying $P_{1|x}$, $P_{0|x}$, and s_x in lemma 1.2.1 provide convenient plug-in estimators of γ^L and γ^H. This section formally defines the estimators and studies their asymptotic properties.

The following notation simplifies expressions for the sample analogues. Let P denote the distribution of an observation (Y, D, Z, X), and f be a real-valued function. Use $P(f)$ to refer to $E_P[f(Y, D, Z, X)]$. Similarly, let $P_{d|x}(f) \equiv E_{P_{d|x}}[f(Y_d)] = E[f(Y_d) \mid D_1 > D_0, X = x]$.

Let $\mathbb{P}_n$ denote the empirical distribution formed from the sample $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$, and $\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^n f(Y_i, D_i, Z_i, X_i)$. The following indicator function notation also simplifies expressions:

$$\mathbb{1}_{d,x,z}(D, X, Z) = \mathbb{1}\{D = d, X = x, Z = z\},$$
$$\mathbb{1}_{x,z}(X, Z) = \mathbb{1}\{X = x, Z = z\}, \qquad \mathbb{1}_x(X) = \mathbb{1}\{X = x\}$$

For example, $P(D = d, X = x, Z = z)$ shortens to $P(\mathbb{1}_{d,x,z})$, and $\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{D_i = 1, X_i = x, Z_i = 0\}$ to $\mathbb{P}_n(\mathbb{1}_{1,x,0})$.

The probabilities $p_{d,x,z} = P(\mathbb{1}_{d,x,z})$, $p_{x,z} = P(\mathbb{1}_{x,z})$, and $p_x = P(\mathbb{1}_x)$ are estimated with empirical analogues:

$$\hat{p}_{d,x,z} \equiv \mathbb{P}_n(\mathbb{1}_{d,x,z}), \qquad \hat{p}_{x,z} \equiv \mathbb{P}_n(\mathbb{1}_{x,z}), \qquad \hat{p}_x \equiv \mathbb{P}_n(\mathbb{1}_x)$$

In this notation, $s_x = P(X = x \mid D_1 > D_0)$ and its empirical analogue $\hat{s}_x$ are

$$s_x = \frac{(p_{1,x,1}/p_{x,1} - p_{1,x,0}/p_{x,0})p_x}{\sum_{x'}(p_{1,x',1}/p_{x',1} - p_{1,x',0}/p_{x',0})p'_x}, \quad \hat{s}_x \equiv \frac{(\hat{p}_{1,x,1}/\hat{p}_{x,1} - \hat{p}_{1,x,0}/\hat{p}_{x,0})\hat{p}_x}{\sum_{x'}(\hat{p}_{1,x',1}/\hat{p}_{x',1} - \hat{p}_{1,x',0}/\hat{p}_{x',0})\hat{p}_{x'}} \tag{1.25}$$

The maps $P_{d|x}$ and their empirical analogues are

$$\begin{aligned} P_{d|x}(f) &= \frac{P(\mathbb{1}_{d,x,d} \times f)/p_{x,d} - P(\mathbb{1}_{d,x,1-d} \times f)/p_{x,1-d}}{p_{d,x,d}/p_{x,d} - p_{d,x,1-d}/p_{x,1-d}}, \\ \hat{P}_{d|x}(f) &\equiv \frac{\mathbb{P}_n(\mathbb{1}_{d,x,d} \times f)/\hat{p}_{x,d} - \mathbb{P}_n(\mathbb{1}_{d,x,1-d} \times f)/\hat{p}_{x,1-d}}{\hat{p}_{d,x,d}/\hat{p}_{x,d} - \hat{p}_{d,x,1-d}/\hat{p}_{x,1-d}}. \end{aligned} \tag{1.26}$$

Computing $\hat{P}_{d|x}(f)$ for a given f is straightforward:

$$\begin{aligned} \hat{P}_{d|x}(f) &= \frac{\frac{1}{\hat{p}_{x,d}}\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{d,x,d}(D_i, X_i, Z_i)f(Y_i) - \frac{1}{\hat{p}_{x,1-d}}\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{d,x,1-d}(D_i, X_i, Z_i)f(Y_i)}{\hat{p}_{d,x,d}/\hat{p}_{x,d} - \hat{p}_{d,x,1-d}/\hat{p}_{x,1-d}} \\ &= \sum_{i=1}^n \omega_{d,x,i} \times f_i \end{aligned}$$

where $f_i = f(Y_i)$ and the weights $\omega_{d,x,i}$ can be computed directly from data:

$$\omega_{d,x,i} \equiv \frac{1}{n} \times \frac{\mathbb{1}_{d,x,d}(D_i, X_i, Z_i)/\hat{p}_{x,d} - \mathbb{1}_{d,x,1-d}(D_i, X_i, Z_i)/\hat{p}_{x,1-d}}{\hat{p}_{d,x,d}/\hat{p}_{x,d} - \hat{p}_{d,x,1-d}/\hat{p}_{x,1-d}}. \tag{1.27}$$

Under assumption 3, $\eta = (\eta_1, \eta_0) = (E_{P_1}[\eta_1(Y_1)], E_{P_0}[\eta_0(Y_0)])$. Each vector $\eta_d \in \mathbb{R}^{K_d}$ has coordinates $\eta_d^{(k)} = \sum_x s_x P_{d|x}(\eta_d^{(k)})$. Empirical analogues $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_0)$ are formed by $\hat{\eta}_d^{(k)} = \sum_x \hat{s}_x \hat{P}_{d|x}(\eta_d^{(k)})$.

The sample analogue estimators of γ^L and γ^H are based on equations (1.22), (1.23), and (1.24):

$$\hat{\theta}_x^L \equiv \theta^L(\hat{P}_{1|x}, \hat{P}_{0|x}), \qquad \hat{\theta}_x^H \equiv \theta^H(\hat{P}_{1|x}, \hat{P}_{0|x}), \tag{1.28}$$

$$\hat{\theta}^L \equiv \sum_x \hat{s}_x \hat{\theta}_x^L, \qquad \hat{\theta}^H \equiv \sum_x \hat{s}_x \hat{\theta}_x^H, \tag{1.29}$$

$$\hat{\gamma}^L \equiv g^L(\hat{\theta}^L, \hat{\theta}^H, \hat{\eta}), \qquad \hat{\gamma}^H \equiv g^H(\hat{\theta}^L, \hat{\theta}^H, \hat{\eta}) \tag{1.30}$$

Where the functions $\theta^L(\cdot,\cdot)$ and $\theta^H(\cdot,\cdot)$ are defined by either (1.19) or (1.20) depending on the cost function. These expressions involve the optimal transport functional $OT_c(P_{1|x}, P_{0|x})$. The sample analogue of the restricted dual problem discussed in section 1.3 is written

$$OT_c(\hat{P}_{1|x}, \hat{P}_{0|x}) = \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \hat{P}_{1|x}(\varphi) + \hat{P}_{0|x}(\psi), \tag{1.31}$$

where the sets $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are defined by displays (1.14) and (1.15) when assumption 2 (i) holds, and by displays (1.16) and (1.17) when assumption 2 (ii) holds.

Computing $OT_c(\hat{P}_{1|x}, \hat{P}_{0|x})$ is especially straightforward when treatment is exogenous. Recall the claim of equation (1.13): the supremum of $P_{1|x}(\varphi) + P_{0|x}(\psi)$ over the larger set Φ_c is the same value when restricted to $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$. The argument behind this claim uses monotonicity of the maps $P_{d|x}$. When treatment is exogenous, $\hat{P}_{d|x}$ corresponds to a probability distribution and is therefore also monotonic. The claim holds replacing $P_{d|x}$ with

$\hat{P}_{d|x}$, implying the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ can be ignored in computation:

$$\begin{aligned} OT_c(\hat{P}_{1|x}, \hat{P}_{0|x}) &= \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \hat{P}_{1|x}(\varphi) + \hat{P}_{0|x}(\psi) = \sup_{(\varphi,\psi)\in\Phi_c} \hat{P}_{1|x}(\varphi) + \hat{P}_{0|x}(\psi) \\ &= \sup_{\{\varphi_i,\psi_j\}_{i,j}} \sum_{i=1}^n \omega_{1,x,i}\varphi_i + \sum_{j=1}^n \omega_{0,x,j}\psi_j \\ &\qquad \text{s.t. } \varphi_i + \psi_j \leq c(Y_i, Y_j) \text{ for all } 1 \leq i, j \leq n. \end{aligned} \tag{1.32}$$

The final problem in this display is a linear programming problem with $2n$ choice variables and n^2 constraints, and can be further simplified by removing choice variables (and the corresponding constraints) whose weights $\omega_{d,x,i}$ equal zero. Many weights do equal zero, as only observations with $X_i = x$ correspond to nonzero weights.

When there is noncompliance in the sample, $\hat{P}_{d|x}$ does not correspond to a probability distribution. This can be seen by noting that for observations i where Z_i differs from D_i, the weight $\omega_{d,x,i}$ defined in display (1.27) is negative. Nonetheless, it remains computationally tractable to search over $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$. For example, when the cost function is continuous $OT_c(\hat{P}_{1|x}, \hat{P}_{0|x})$ remains a linear programming problem with additional linear constraints enforcing $|\varphi_i + \psi_j| \leq L|Y_i - Y_j|$, $-\|c\|_\infty \leq \varphi_i \leq \|c\|_\infty$, and $-2\|c\|_\infty \leq \psi_j \leq 0$.

1.5.1 Weak convergence

The estimators proposed above are especially attractive because they are a (Hadamard directionally) differentiable map of the empirical distribution. Specifically, there exists a collection of functions $\mathcal{F}$ and a map $T : \ell^\infty(\mathcal{F}) \to \mathbb{R}^2$ described by equations (1.25), (1.26), (1.28), (1.29), and (1.30) such that

$$(\hat{\gamma}^L, \hat{\gamma}^H) = T(\mathbb{P}_n), \qquad\qquad (\gamma^L, \gamma^H) = T(P)$$

The set $\mathcal{F}$ consists of the functions in $\mathcal{F}_c$, $\mathcal{F}_c^c$, and the coordinate functions defining η, multiplied by various indicator functions. It is formally defined in appendix 1.9.3. Assumptions 1, 2, and 3, suffice to show $\mathcal{F}$ is Donsker and $T(\cdot)$ is continuous at P, and therefore that the esimators are consistent:

$$(\hat{\gamma}^L, \hat{\gamma}^H) = T(\mathbb{P}_n) \xrightarrow{p} T(P) = (\gamma^L, \gamma^H) \tag{1.33}$$

The map $T(\cdot)$ is not only continuous, but Hadamard directionally differentiable. An application of the functional delta method gives the conclusion $\sqrt{n}((\hat{\gamma}^L, \hat{\gamma}^H) - (\gamma^L, \gamma^H))$ converges in distribution, a result stated formally in theorem 1.5.2 below.

In order to build hypothesis tests or construct confidence intervals based on the asymptotic distribution of $\sqrt{n}((\hat{\gamma}^L, \hat{\gamma}^H) - (\gamma^L, \gamma^H))$, one must be able to estimate the asymptotic distribution. This is possible under assumptions 1, 2, and 3, with a procedure described in section 1.5.2.2. Under an additional assumption, a straightforward bootstrap will do. For each instance of the restricted dual problem used in defining $T(\cdot)$, the set of maximizers

$$\Psi_c(P_{1|x}, P_{0|x}) \equiv \underset{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)}{\arg\max} P_{1|x}(\varphi) + P_{0|x}(\psi) \tag{1.34}$$

is nonempty. If the solutions are suitably unique for each instance, the map $T(\cdot)$ is fully Hadamard differentiable at P and a straightforward bootstrap will consistently estimate the asymptotic distribution. Assumption 4 states this high-level uniqueness condition, while the following lemma 1.5.1 gives low-level sufficient conditions for it to hold. Let $\mathcal{Y}_{d,x}$ be the support of Y conditional on $D = d$ and $X = x$, and $\mathbb{1}_{\mathcal{Y}_{d,x}}(y) = \mathbb{1}\{y \in \mathcal{Y}_{d,x}\}$ be the indicator function for this set.

Assumption 4 (Unique solutions)**.** *For each $x \in \mathcal{X}$, each $c \in \{c_L, c_H\}$, and any*

$(\varphi_1, \psi_1), (\varphi_2, \psi_2) \in \Psi_c(P_{1|x}, P_{0|x})$, *there exists* $s \in \mathbb{R}$ *such that*

$$\mathbb{1}_{\mathcal{Y}_{1,x}} \times \varphi_1 = \mathbb{1}_{\mathcal{Y}_{1,x}} \times (\varphi_2 + s) \qquad \text{and} \qquad \mathbb{1}_{\mathcal{Y}_{0,x}} \times \psi_1 = \mathbb{1}_{\mathcal{Y}_{0,x}} \times (\psi_2 - s)$$

P-almost surely.

Lemma 1.5.1. *Suppose that*

(i) assumption 2 (i) holds, with cost function $c(y_1, y_0)$ *that is continuously differentiable, and*

(ii) for each (d, x)*, the support of* $P_{d|x}$ *is* $\mathcal{Y}_{d,x}$*, which is a bounded interval.*

Then assumption 4 holds.

When treatment is exogenous, condition (ii) of lemma 1.5.1 simplifies to the assumption that the distribution of $Y_d \mid X = x$ has bounded support $[y^\ell_{d,x}, y^u_{d,x}]$. In a setting with instruments, this condition requires the support of Y_1 for compliers is a bounded interval containing the support of Y_1 for always-takers, and the support of Y_0 for compliers is a bounded interval containing the support of Y_0 for never-takers.

Assumption 4 can hold even when the conditions of lemma 1.5.1 do not. For example, when the parameter of interest is the cumulative distribution function of the treatment effects evaluated at a point and assumption 2 (ii) is satisfied, the dual problem is essentially optimizing over the difference of CDFs (see remark 1.4.1). Although the cost functions are not continuously differentiable, it is still plausible for this optimization problem to have a unique solution. For further discussion of uniqueness of the dual solutions of optimal transport, see Staudt et al. (2022).

The following theorem gives the main weak convergence result.

Theorem 1.5.2 (Weak convergence)**.** *Suppose assumptions 1, 2, and 3 hold, and let* $\mathbb{G}$ *be the weak limit of* $\sqrt{n}(\mathbb{P}_n - P)$ *in* $\ell^\infty(\mathcal{F})$*. Then* T *is Hadamard directionally differentiable at*

P tangentially to the support of $\mathbb{G}$, and

$$\sqrt{n}((\hat{\gamma}^L, \hat{\gamma}^H) - (\gamma^L, \gamma^H)) = \sqrt{n}(T(\mathbb{P}_n) - T(P)) \xrightarrow{L} T'_P(\mathbb{G})$$

If assumption 4 also holds, then T'_P is linear on the support of $\mathbb{G}$ and $T'_P(\mathbb{G})$ is bivariate normal.

1.5.2 Inference

To make use of theorem 1.5.2 for inference, this section develops methods of estimating the law of $T'_P(\mathbb{G})$ by utilizing the bootstrap. The "exchangeable bootstrap" procedures discussed in van der Vaart & Wellner (1997) are computationally convenient. These procedures define a new map $\mathbb{P}_n^* \in \ell^\infty(\mathcal{F})$ pointwise with

$$\mathbb{P}_n^*(f) = \frac{1}{n}\sum_{i=1}^{n} W_i f(Y_i, D_i, Z_i, X_i) \tag{1.35}$$

for nonnegative random variables $\{W_i\}_{i=1}^n$ independent of the data $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$, and satisfying technical conditions omitted here. Two notable examples include the nonparametric bootstrap of Efron (1979) and the "Bayesian" bootstrap of Rubin (1981). Either bootstrap can be used to estimate the asymptotic distribution. The Bayesian bootstrap may be preferable in small samples for reasons discussed below.

Definition 1.5.1 (Nonparametric bootstrap)**.** *Let $(W_1, \ldots, W_n) \sim$ Multinomial$(n, (1/n, \ldots, 1/n))$ be independent of the data $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$. Define $\mathbb{P}_n^* \in \ell^\infty(\mathcal{F})$ pointwise with* (1.35)*.*

Definition 1.5.2 (Bayesian bootstrap)**.** *Let $\{\xi_i\}_{i=1}^n$ be i.i.d. exponentially distributed random variables with mean* 1*, independent of the data $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$. Set $W_i = \xi_i / (n^{-1}\sum_{i=1}^n \xi_i)$, and define $\mathbb{P}_n^* \in \ell^\infty(\mathcal{F})$ pointwise with* (1.35)*.*

The map $\mathbb{P}_n^*$ in (1.35) can be used to compute $(\hat{\gamma}^{L*}, \hat{\gamma}^{H*}) = T(\mathbb{P}_n^*)$ in much the same way that $T(\mathbb{P}_n)$ is computed. Specifically, bootstrap analogues of $\hat{p}_{d,x,z}$, $\hat{p}_{x,z}$, and $\hat{p}_x$ are given by

$$\hat{p}_{d,x,z}^* = \frac{1}{n}\sum_{i=1}^n W_i \mathbb{1}_{d,x,z}(D_i, X_i, Z_i), \quad \hat{p}_{x,z}^* = \frac{1}{n}\sum_{i=1}^n W_i \mathbb{1}_{x,z}(X_i, Z_i), \quad \hat{p}_x^* = \frac{1}{n}\sum_{i=1}^n W_i \mathbb{1}_x(X_i),$$

and the bootstrap analogue of $\hat{s}_x$ is

$$\hat{s}_x^* = \frac{(\hat{p}_{1,x,1}^*/\hat{p}_{x,1}^* - \hat{p}_{1,x,0}^*/\hat{p}_{x,0}^*)\hat{p}_x^*}{\sum_{x'}(\hat{p}_{1,x',1}^*/\hat{p}_{x',1}^* - \hat{p}_{1,x',0}^*/\hat{p}_{x',0}^*)\hat{p}_{x'}^*}$$

The maps $\hat{P}_{d|x}$ have bootstrap analogues

$$\hat{P}_{d|x}^*(f) = \frac{\mathbb{P}_n^*(\mathbb{1}_{d,x,d} \times f)/\hat{p}_{x,d}^* - \mathbb{P}_n^*(\mathbb{1}_{d,x,1-d} \times f)/\hat{p}_{x,1-d}^*}{\hat{p}_{d,x,d}^*/\hat{p}_{x,d}^* - \hat{p}_{d,x,1-d}^*/\hat{p}_{x,1-d}^*} = \sum_{i=1}^n \omega_{d,x,i}^* f_i$$

where $f_i = f(Y_i)$ and $\omega_{d,x,i}^*$ are bootstrap versions of the weights in (1.27):

$$\omega_{d,x,i}^* = \frac{W_i}{n} \times \frac{\mathbb{1}_{d,x,d}(D_i, X_i, Z_i)/\hat{p}_{x,d}^* - \mathbb{1}_{d,x,1-d}(D_i, X_i, Z_i)/\hat{p}_{x,1-d}^*}{\hat{p}_{d,x,d}^*/\hat{p}_{x,d}^* - \hat{p}_{d,x,1-d}^*/\hat{p}_{x,1-d}^*} \tag{1.36}$$

Finally, $(\hat{\gamma}^{L*}, \hat{\gamma}^{H*})$ can be computed with

$$\hat{\theta}_x^{L*} = \theta^L(\hat{P}_{1|x}^*, \hat{P}_{0|x}^*), \qquad \hat{\theta}_x^{H*} = \theta^H(\hat{P}_{1|x}^*, \hat{P}_{0|x}^*), \tag{1.37}$$

$$\hat{\theta}^{L*} = \sum_x \hat{s}_x^* \hat{\theta}_x^{L*}, \qquad \hat{\theta}^{H*} = \sum_x \hat{s}_x^* \hat{\theta}_x^{H*}, \tag{1.38}$$

$$\hat{\gamma}^{L*} = g^L(\hat{\theta}^{L*}, \hat{\theta}^{H*}, \hat{\eta}^*), \qquad \hat{\gamma}^{H*} = g^H(\hat{\theta}^{L*}, \hat{\theta}^{H*}, \hat{\eta}^*) \tag{1.39}$$

1.5.2.1 Simple bootstrap with full differentiability

Under assumption 4, estimating the distribution of $T_P'(\mathbb{G})$ is straightforward.

Theorem 1.5.3. *Suppose assumptions 1, 2, 3, and 4 hold, and let $\mathbb{P}_n^*$ be given by definition*

1.5.1 or 1.5.2. Then conditional on $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$,

$$\sqrt{n}(T(\mathbb{P}_n^*) - T(\mathbb{P}_n)) \xrightarrow{L} T'_P(\mathbb{G})$$

in outer probability.

It is worth emphasizing the computationally convenience of the bootstrap $\mathbb{P}_n^*$ given in (1.35) when treatment is exogenous. The weights given in display (1.36) simplify to

$$\omega_{d,x,i}^* = \frac{W_i}{n} \times \frac{\mathbb{1}\{D_i = d, X_i = x\}}{\hat{p}_{x,d}^*} \tag{1.40}$$

As these weights are nonnegative and sum to one, $\hat{P}_{d|x}^*$ is a probability distribution. Accordingly, the function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ can be ignored when computing $\theta^L(\hat{P}_{1|x}^*, \hat{P}_{0|x}^*)$ and $\theta^H(\hat{P}_{1|x}^*, \hat{P}_{0|x}^*)$ for the same reasons discussed above.

A researcher utilizing the nonparametric bootstrap in a small sample runs the risk of a boostrap draw including no observations with $(D_i, X_i) = (d, x)$. This would result in the formula in (1.40) attempting to divide by zero. This problem cannot arise when using the Bayesian bootstrap suggested in 1.5.2; in this procedure $W_i > 0$ for each i, and thus $\hat{p}_{x,d}^* = \frac{1}{n}\sum_{i=1}^n W_i \mathbb{1}\{D_i = d, X_i = x\} > 0$ as long as $\hat{p}_{d,x} > 0$.

1.5.2.2 Alternative for directional differentiability

The solutions to optimal transport may not be unique as assumption 4 requires. As emphasized in the statement of theorem 1.5.2, assumption 4 is not needed to obtain the asymptotic distribution of the estimators – but a straightforward bootstrap may not consistently estimate that limiting distribution. When in doubt, researchers can make use of an alternative procedure based on the results of Fang & Santos (2019) and described below.

Additional notation is needed to describe this alternative. Let $\eta_{d,x}^{(k)} \equiv P_{d|x}(\eta_d^{(k)})$, and $T_1(\cdot)$

denote the "first stage" function computing $P_{1|x}$, $P_{0|x}$, $\eta_{1,x}$, $\eta_{0,x}$, and s_x for each x:

$$T_1(P) = \left(\left\{ P_{1|x}, P_{0|x}, \eta_{1,x}, \eta_{0,x}, s_x \right\}_{x \in \mathcal{X}} \right)$$

Here $\{a_x\}_{x \in \mathcal{X}} = (a_{x_1}, \ldots, a_{x_M})$. Let $\{\kappa_n\}_{n=1}^{\infty}$ be a sequence in $\mathbb{R}$ satisfying $\kappa_n \uparrow \infty$ and $\kappa_n / \sqrt{n} \to 0$. Define the set of empirical approximate maximizers:

$$\widehat{\Psi}_{c,x} \equiv \left\{ (\varphi, \psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c) \; ; \; OT_c(\hat{P}_{1|x}, \hat{P}_{0|x}) \leq \hat{P}_{1|x}(\varphi) + \hat{P}_{0|x}(\psi) + \frac{\kappa_n}{\sqrt{n}} \right\}.$$

Use this set to define the maps

$$\widehat{OT}'_{c,x}(H_1, H_0) = \sup_{(\varphi,\psi) \in \widehat{\Psi}_{c,x}} H_1(\varphi) + H_0(\psi)$$

and

$$\begin{aligned} &\widehat{T}'_{2,T_1(P)} \left(\{ H_{1,x}, H_{0,x}, h_{\eta_1,x}, h_{\eta_0,x}, h_{s,x} \}_{x \in \mathcal{X}} \right) \\ &\quad = \left(\left\{ \widehat{OT}'_{c_L,x}(H_{1,x}, H_{0,x}), -\widehat{OT}'_{c_H,x}(H_{1,x}, H_{0,x}), h_{\eta_1,x}, h_{\eta_0,x}, h_{s,x} \right\}_{x \in \mathcal{X}} \right). \end{aligned}$$

The alternative procedure uses the conditional law of

$$\hat{D}_4 \hat{D}_3 \widehat{T}'_{2,T_1(P)} \left(\sqrt{n} (T_1(\mathbb{P}_n^*) - T_1(\mathbb{P}_n)) \right)$$

given the data, where $\hat{D}_4$ and $\hat{D}_3$ are matrices given by

$$\hat{D}_3 = \underset{(2+d_\eta)\times M(3+d_\eta)}{\left[\hat{D}_{3,x_1} \quad \hat{D}_{s,x_2} \quad \dots \quad \hat{D}_{s,x_M}\right]}, \qquad \hat{D}_{3,x} = \underset{(2+d_\eta)\times(3+d_\eta)}{\begin{bmatrix} \hat{s}_x & 0 & 0 & 0 & \hat{\theta}_x^L \\ 0 & \hat{s}_x & 0 & 0 & \hat{\theta}_x^H \\ 0 & 0 & \hat{s}_x I_{K_1} & 0 & \hat{\eta}_{1,x} \\ 0 & 0 & 0 & \hat{s}_x I_{K_0} & \hat{\eta}_{0,x} \end{bmatrix}},$$

$$D_4 = \underset{2\times(2+d_\eta)}{\begin{bmatrix} \nabla g^L(\hat{\theta}^L, \hat{\theta}^H, \hat{\eta})^\intercal \\ \nabla g^H(\hat{\theta}^L, \hat{\theta}^H, \hat{\eta})^\intercal \end{bmatrix}},$$

Theorem 1.5.4. *Suppose assumptions 1, 2, and 3 hold, let $\mathbb{P}_n^*$ be given by definition 1.5.1 or 1.5.2, and $\{\kappa_n\}_{n=1}^\infty \subseteq \mathbb{R}$ satisfy $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$. Then conditional on $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$,*

$$\hat{D}_4\hat{D}_3\widehat{T}_{2,T_1(P)}(\sqrt{n}(T_1(\mathbb{P}_n^*) - T_1(\mathbb{P}_n))) \xrightarrow{L} T'_P(\mathbb{G})$$

in outer probability.

1.5.2.3 Confidence sets

Theorems 1.5.3 and 1.5.4 make it straightforward to conduct inference. For example, a simple confidence set for the identified set $[\gamma^L, \gamma^H]$ is given by

$$\left[\hat{\gamma}^L - \hat{q}_{1-\alpha}/\sqrt{n}, \hat{\gamma}^H + \hat{q}_{1-\alpha}/\sqrt{n}\right]$$

where $\hat{q}_{1-\alpha}$ is a consistent estimator of the $1-\alpha$ quantile of $\max\{T'_P(\mathbb{G})^{(1)}, -T'_P(\mathbb{G})^{(2)}\}$. When assumptions 1 through 4 hold, let $(\hat{\gamma}^{L*}, \hat{\gamma}^{H*}) = T(\mathbb{P}_n^*)$. When assumptions 1 through 3 hold but assumption 4 is doubtful, let $(\hat{\gamma}^{L*}, \hat{\gamma}^{H*}) = (\hat{\gamma}^L, \hat{\gamma}^H) + \frac{1}{\sqrt{n}}\hat{D}_4\hat{D}_3\widehat{T}_{2,T_1(P)}(\sqrt{n}(T_1(\mathbb{P}_n^*) -$

$T_1(\mathbb{P}_n)$))). In either case, compute

$$\hat{q}_{1-\alpha} = \inf\left\{q \,;\; P\left(\max\left\{\sqrt{n}(\hat{\gamma}^{L*} - \hat{\gamma}^{L}), -\sqrt{n}(\hat{\gamma}^{H*} - \hat{\gamma}^{H})\right\} \leq q \mid \{Y_i, D_i, Z_i, X_i\}_{i=1}^n\right) \geq 1-\alpha\right\}$$

through simulation:

1. Compute $(\hat{\gamma}^L, \hat{\gamma}^H) = T(\mathbb{P}_n)$ and, if necessary, $\hat{D}_4$, and $\hat{D}_3$.
2. Generate B boostrap samples, $\{W_{i,b}\}_{i=1}^n$ for each $b = 1, \ldots, B$ according to definition 1.5.1 or 1.5.2. For each bootstrap sample b, compute $(\hat{\gamma}_b^{L*}, \hat{\gamma}_b^{H*})$ as described above.
3. Let $\hat{q}_{1-\alpha}$ be the $1-\alpha$ quantile of $\{\max\{\sqrt{n}(\hat{\gamma}_b^{L*} - \hat{\gamma}^{L}), -\sqrt{n}(\hat{\gamma}_b^{H*} - \hat{\gamma}^{H})\}_{b=1}^B$.

Under the further assumption that the cumulative distribution function of $\max\{T_P'(\mathbb{G})^{(1)}, -T_P'(\mathbb{G})^{(2)}\}$ is continuous and strictly increasing at its $1-\alpha$ quantile,

$$\lim_{n\to\infty} P\left([\gamma^L, \gamma^H] \subseteq \left[\hat{\gamma}^L - \hat{q}_{1-\alpha}/\sqrt{n}, \hat{\gamma}^H + \hat{q}_{1-\alpha}/\sqrt{n}\right]\right) = 1-\alpha$$

1.6 Simulations

This section explores the finite sample performance of the estimators through simulations, with a focus on coverage rates of confidence sets for the identified set. For simplicity, the data generating process is one of exogenous treatment with no covariates. An observation consists of the vector (Y, D), where $Y = DY_1 + (1-D)Y_0$. Treatment status $D \in \{0, 1\}$ is independent of (Y_1, Y_0), and satisfies $P(D = 1) = 0.5$. Potential outcomes follow with a Kumaraswamy distribution with positive parameters a_d and b_d, having support $[0, 1]$ and cumulative distribution function

$$F_d(y) = P(Y_d \leq y) = 1 - (1 - y^{a_d})^{b_d}$$

The parameter of interest is

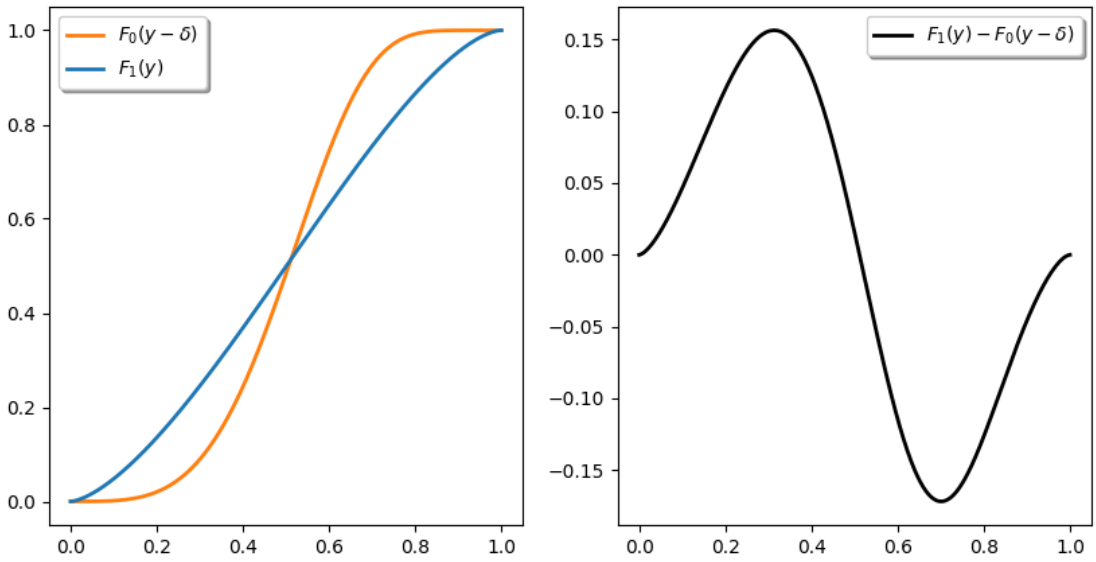
$$\gamma = \theta = P(Y_1 - Y_0 \leq \delta),$$

This parameter is chosen to facilitate computation of the true identified set. As noted in remark 1.4.1, the optimal transport problems involved in characterizing the identified set have simple analytical expressions. Specifically, the identified set for γ is $[\gamma^L, \gamma^H]$, where

$$\gamma^L = \sup_y \{F_1(y) - F_0(y - \delta)\}, \qquad \gamma^H = 1 + \inf_y \{F_1(y) - F_0(y - \delta)\}$$

These expressions and the closed form cdfs F_d allow the true values of γ^L and γ^H to be computed precisely without simulation. The population cumulative distribution functions of Y_1 and Y_0, as well as their difference, are displayed in Figure 1.1.

Figure 1.1: Simulation data generating process, cdfs and dual objective



As is clear from the right panel, there is a unique and well separated maximum and minimum of $F_1(y) - F_0(y - \delta)$ that imply population bounds of $\gamma^L = 0.156$ and $\gamma^H = 0.828$.

The uniqueness of these optimizers indicate that $T(\cdot)$ is fully differentiable, and thus the straightforward bootstrap described by theorem 1.5.3 consistently estimates the asymptotic distribution.

In each simulation, an i.i.d. sample $\{Y_i, D_i\}_{i=1}^n$ is drawn according to the data generating process described above. The estimators are computed as described in section 1.5:

$$\hat{\gamma}^L = OT_{c_L}(\hat{P}_1, \hat{P}_0), \qquad \hat{\gamma}^H = 1 - OT_{c_H}(\hat{P}_1, \hat{P}_0)$$

where the cost functions are $c_L(y_1, y_0) = \mathbb{1}\{y_1 - y_0 < \delta\}$ and $c_H(y_1, y_0) = \mathbb{1}\{y_1 - y_0 > \delta\}$ and optimal transport is computed as

$$\begin{aligned} OT_c(\hat{P}_1, \hat{P}_0) = \sup_{\{\varphi_i, \psi_j\}_{i,j}} & \sum_{i=1}^n \omega_{1,i}\varphi_i + \sum_{j=1}^n \omega_{0,j}\psi_j \\ & \text{s.t. } \varphi_i + \psi_j \leq c(Y_i, Y_j) \text{ for all } 1 \leq i, j \leq n \end{aligned}$$

3,000 bootstrap draws are used to compute the confidence set

$$CI = \left[\hat{\gamma}^L - \hat{c}_{1-\alpha}/\sqrt{n}, \hat{\gamma}^H + \hat{c}_{1-\alpha}/\sqrt{n}\right]$$

with $\alpha = 0.05$, following the procedures outlined in section 1.5.2.3.

It is well known that estimators optimizing over sample averages are biased in small samples (Haile & Tamer, 2003; Kreider & Pepper, 2007; Chernozhukov et al., 2013). Specifically, the expectation of a sup over a sample average is larger than the sup over its popoulation counterpart due to convexity of the sup function. This suggests that in small samples $\hat{\gamma}^L$ is biased upward, and $\hat{\gamma}^H$ biased downward, leading to estimated bounds that are tighter than their population counterparts. Although theorems 1.5.2 and 1.5.3 guarantee correct coverage asymptotically, this finite sample bias can lead to undercoverage in small samples.

Table 1.1 reports the empirical bias and standard deviation of the estimator, as well as

the empirical coverage of the confidence set, from 300 simulations.

Table 1.1: Simulations without bias correction

n	Bias		St. Dev.		Emp. Coverage
	$\hat{\gamma}^L$	$\hat{\gamma}^H$	$\hat{\gamma}^L$	$\hat{\gamma}^H$	CI
100	0.047	-0.051	0.065	0.066	0.900
200	0.031	-0.031	0.049	0.049	0.917
300	0.030	-0.021	0.040	0.040	0.893

The bias is notable in magnitude relative to the standard deviation in these small sample sizes. Empicial coverage is slightly below the nominal value.

The bootstrap bias correction found in Efron & Tibshirani (1994) is simple to implement in the current setting. The finite sample bias of the lower and upper bounds is given by $E[\hat{\gamma}^L] - \gamma^L$ and $E[\hat{\gamma}^H] - \gamma^H$ respectively. These are estimated by $\widehat{bias}^L = B^{-1} \sum_{b=1}^{B} \hat{\gamma}_b^{L*} - \hat{\gamma}^L$ and $\widehat{bias}^H = B^{-1} \sum_{b=1}^{B} \hat{\gamma}_b^{H*} - \hat{\gamma}^H$. The bootstrap bias corrected estimate of the bounds are given by

$$\hat{\gamma}_{BC}^L = \hat{\gamma}^L - \widehat{bias}^L, \qquad \hat{\gamma}_{BC}^H = \hat{\gamma}^H - \widehat{bias}^H$$

The bootstrap bias correction is often found to reduce finite sample bias in simulations and to offer a higher order refinement in various settings (Horowitz, 2001; Hahn et al., 2002). In the context of smooth functions of sample moments, Horowitz (2001) notes that the asymptotic distribution of the corrected estimator is the same as that of the uncorrected estimator when B increases sufficiently quickly with n. The boostrap bias corrected confidence set for the identified set is given by

$$CI_{BC} = \left[\hat{\gamma}_{BC}^L - \hat{c}_{1-\alpha}/\sqrt{n}, \hat{\gamma}_{BC}^H + \hat{c}_{1-\alpha}/\sqrt{n}\right]$$

Table 1.2 reports the results from the same 300 simulations using this bias correction.

Table 1.2: Simulations with bias correction

n	Bias		St. Dev.		Emp. Coverage
	$\hat{\gamma}^L_{BC}$	$\hat{\gamma}^H_{BC}$	$\hat{\gamma}^L_{BC}$	$\hat{\gamma}^H_{BC}$	CI_{BC}
100	0.021	-0.026	0.071	0.071	0.927
200	0.013	-0.015	0.052	0.051	0.953
300	0.015	-0.007	0.042	0.042	0.957

Empirical bias is approximately halved, and coverage is close to the nominal 95%. Efron & Tibshirani (1994) warns that the bootstrap bias correction may increase the variance of the estimator, but in this case the standard deviation increased only marginally.

1.7 Application: National Supported Work Demonstration

This section demonstrates the estimators by revisiting the famous National Supported Work Demonstration program (LaLonde (1986)). This program was implemented in the 1970s with the aim of helping socially and economically disadvantaged workers obtain job skills. Those randomly selected into the program were guaranteed a job lasting six to eighteen months, and frequently met with a counselor to discuss performance. There was no reported noncompliance.

The "LaLonde" sample studied in Diamond & Sekhon (2013) consists of male participants and includes 297 treated and 425 control observations. The outcome of interest is real earnings in 1978. Observed covariates include age, years of education, real earnings in months 13 to 24 prior to randomization, and indicators for whether a participant is a high school dropout, black, hispanic, or married. Averages and standard deviations of these covariates by treatment status are reported in table 1.3:

Table 1.3: Balance table

	base inc.	age	yrs. educ.	HS dropout	black	hispanic	married
control	3672.49	24.45	10.19	0.81	0.80	0.11	0.16
	(6521.53)	(6.59)	(1.62)	(0.39)	(0.40)	(0.32)	(0.36)
treated	3571.00	24.63	10.38	0.73	0.80	0.09	0.17
	(5773.13)	(6.69)	(1.82)	(0.44)	(0.40)	(0.29)	(0.37)

Note: Standard deviations in parentheses.

In this sample, the average treatment effect on 1978 real earnings is \$886. It is natural to ask whether the policy was more beneficial for those who would have low incomes in 1978 without treatment. One parameter addressing this is the OLS slope coefficient of regressing treatment effects on a constant and Y_0:

$$\gamma = \frac{\text{Cov}(Y_1 - Y_0, Y_0)}{\text{Var}(Y_0)} = \frac{E_{P_{1,0}}[(Y_1 - Y_0)Y_0] - (E_{P_1}[Y_1] - E_{P_0}[Y_0])E_{P_0}[Y_0]}{E_{P_0}[Y_0^2] - (E_{P_0}[Y_0])^2}.$$

As described in example 1.2.3, the sign of this parameter describes who receives larger benefits from treatment. Specifically, $\gamma < 0$ implies those with below average untreated outcomes tend to see above average treatment effects.

Discretized versions of baseline income and age are found to be informative covariates. Baseline income is binned as $[0, 0]$, $(0, 4000]$, or $(4000, \infty)$ while age is binned as $[16, 23]$, or $(23, \infty)$. X is the cartesian product of bins. The point estimates for the bounds are $(\hat{\gamma}^L, \hat{\gamma}^H) = (-1.725, -0.003)$.[3] The negative upper bound point estimate suggests that the treatment was especially beneficial for participants who would otherwise have incomes below average (for the eligible population). The bias-corrected point estimates based on 3,000 bootstrap draws are $(\hat{\gamma}^L_{BC}, \hat{\gamma}^H_{BC}) = (-1.731, 0.041)$, and the bias-corrected 95% confidence set for the identified set is $[-1.956, 0.266]$. These suggest that γ may still be zero or slightly

[3]Covariates are found to be informative, especially for the upper bound. Ignoring covariates, the lower bound point estimate is -1.783 and the upper bound point estimate is 0.190.

positive once accounting for sample uncertainty.

This parameter could also be considered conditional on each of the covariate values:

$$\gamma_x \equiv \frac{\text{Cov}(Y_1 - Y_0, Y_0 \mid X = x)}{\text{Var}(Y_0 \mid X = x)}.$$

Bias corrected point estimates and confidence intervals for each γ_x are reported in Table 1.4.

Table 1.4: Estimates conditional on covariate values

age	base inc.	$\hat{\gamma}^L_{BC}$	$\hat{\gamma}^H_{BC}$	CI_{BC}	n
(16, 23]	0	-1.97	0.28	[-2.26, 0.56]	140
	(0, 4000]	-1.74	-0.15	[-1.9, 0.01]	141
	(4000, ∞)	-1.45	-0.44	[-1.63, -0.27]	90
(23, 55]	0	-2.13	0.81	[-2.65, 1.33]	187
	(0, 4000]	-1.39	-0.16	[-1.93, 0.38]	56
	(4000, ∞)	-1.66	0.03	[-2.08, 0.45]	108

It is worth noting the upper bound on the confidence set is negative for young men with baseline income above $\$4,000$, and essentially zero for young men with positive income below $\$4,000$. For these subpopulations, those who would have had below average incomes in 1978 tended to see above average benefits from treatment.

1.8 Conclusion

This paper studies a large class of causal parameters that depend on a moment of the joint distribution of potential outcomes, in a setting with binary treatment, a binary instrument satisfying a monotonicity restriction, and finitely supported covariates. The sharp identified set of such parameters is characterized with the value of optimal transport problems. Estimators based on this identification are $\sqrt{n}$-consistent and converge in distribution under

mild assumptions, and inference procedures based on the bootstrap are straightforward and computationally convenient.

1.9 Appendix

1.9.1 Appendix: identification

Following Kitagawa (2015), let T denote the "type" of a unit:

$$T = \begin{cases} a, \text{ always-taker}, & \text{if } (D_1, D_0) = (1,1) \\ c, \text{ complier}, & \text{if } (D_1, D_0) = (1,0) \\ n, \text{ never-taker}, & \text{if } (D_1, D_0) = (0,0) \\ df, \text{ defier}, & \text{if } (D_1, D_0) = (0,1) \end{cases} \tag{1.41}$$

Note that the primitives $(Y_1, Y_0, D_1, D_0, Z, X)$ are equivalent to (Y_1, Y_0, T, Z, X).

1.9.1.1 Main identification results

Lemma 1.9.1 (Identification of moments)**.** *Suppose assumptions 1 and 2 hold. Then the sharp identified set for θ is $[\theta^L, \theta^H]$.*

Proof. Let T be as defined in (1.41), and note that the primitives of the model $(Y_1, Y_0, D_1, D_0, Z, X)$ are equivalent to (Y_1, Y_0, T, Z, X). Moreover, the event $D_1 > D_0$ is the event $T = c$; thus $P_{d|x}$ is the distribution of $Y_d \mid T = c, X = x$.

In steps:

1. The identified set for $(P_{1,0|x_1}, \ldots, P_{1,0|x_M})$, the conditional distributions of $(Y_1, Y_0) \mid T = c, X = x$ for each $x \in \mathcal{X} = \{x_1, \ldots, x_M\}$, is $\Pi(P_{1|x_1}, P_{0|x_1}) \times \ldots \times \Pi(P_{1|x_M}, P_{0|x_M})$.

That $(P_{1,0|x_1}, \ldots, P_{1,0|x_M}) \in \Pi(P_{1|x_1}, P_{0|x_1}) \times \ldots \times \Pi(P_{1|x_M}, P_{0|x_M})$ is immediate. To see that any element of $\Pi(P_{1|x_1}, P_{0|x_1}) \times \ldots \times \Pi(P_{1|x_M}, P_{0|x_M})$ is possible given the assumptions and distribution of the observables (Y, D, Z, X), fix a distribution of the observables generated by a distribution of the primitives consistent with the assumptions. Note that the distribution of observables is summarized by $P(D = d, Z = z, X = x)$ for each (d, z, x) and the conditional distributions

$$Y \mid D = d, Z = z, X = x$$

Use this observation and the claims of lemma 1.9.5 to see that any two distributions of the primitives (Y_1, Y_0, T, Z, X) (consistent with the assumptions), sharing the same distribution of (T, Z, X), and the same marginal, conditional distributions for

$$\begin{aligned} &Y_1 \mid T = a, X = x && Y_0 \mid T = n, X = x \\ &Y_1 \mid T = c, X = x, && Y_0 \mid T = c, X = x \end{aligned}$$

will produce this distribution of observables. Thus, replacing $(P_{1,0|x_1}, \ldots, P_{1,0|x_M})$ from the distribution of primitives with any

$$(\pi_{x_1}, \ldots, \pi_{x_M}) \in \Pi(P_{1|x_1}, P_{0|x_1}) \times \ldots \times \Pi(P_{1|x_M}, P_{0|x_M})$$

will generate the same observed distribution of (Y, D, Z, X), without violating assumption 1 or 2. The claim follows.

2. The identified set for $(\theta_{x_1}, \ldots, \theta_{x_M}) \in \mathbb{R}^M$ is $[\theta^L_{x_1}, \theta^H_{x_1}] \times \ldots \times [\theta^L_{x_M}, \theta^H_{x_M}]$.

   Recall that $\theta_x = E[c(Y_1, Y_0) \mid X = x]$, and let $\Theta_{I,x}$ denote its identified set. Note that

the previous step implies

$$\Theta_{I,x} = \left\{ t \in \mathbb{R} \; ; \; t = E_{\pi_x}[c(Y_1, Y_0)] \text{ for some } \pi_x \in \Pi(P_{1|x}, P_{0|x}) \right\}$$

$\Pi(P_{1|x}, P_{0|x})$ is convex. Notice that for any $\lambda \in (0,1)$ and $\pi_x^1, \pi_x^0 \in \Pi(P_{1|x}, P_{0|x})$, $E_{\lambda\pi_x^1 + (1-\lambda)\pi_x^0}[c(Y_1, Y_0)] = \lambda E_{\pi_x^1}[c(Y_1, Y_0)] + (1-\lambda) E_{\pi_x^0}[c(Y_1, Y_0)]$. Together these imply $\Theta_{I,x}$ is convex.

It suffices to show that for any x, $\Theta_{I,x} = [\theta_x^L, \theta_x^H]$ There are two cases:

(i) If assumption 2 (i) holds, then for each x,

$$\theta_x^L = OT_c(P_{1|x}, P_{0|x}) = \inf_{\pi_x \in \Pi(P_{1|x}, P_{0|x})} E_{\pi_x}[c(Y_1, Y_0)]$$
$$\theta_x^H = -OT_{-c}(P_{1|x}, P_{0|x}) = \sup_{\pi_x \in \Pi(P_{1|x}, P_{0|x})} E_{\pi_x}[c(Y_1, Y_0)]$$

Since c is continuous, lemma 1.9.30 implies the optimal transport problems are attained, say by π_x^L and π_x^H respectively. It follows that $\theta_x^L, \theta_x^H \in \Theta_{I,x}$, and it is clear from their definitions that they bound $\Theta_{I,x}$. Since $\Theta_{I,x}$ is convex, it follows that $\Theta_{I,x} = [\theta_x^L, \theta_x^H]$.

(ii) If Assumption 2 (ii) holds, then

$$c_L(y_1, y_0) = \mathbb{1}\{y_1 - y_0 < \delta\}, \qquad c_H(y_1, y_0) = \mathbb{1}\{y_1 - y_0 > \delta\},$$
$$\theta_x^L = OT_{c_L}(P_{1|x}, P_{0|x}), \qquad \theta_x^H = 1 - OT_{c_H}(P_{1|x}, P_{0|x})$$

Let $\pi_x^L, \pi_x^H \in \Pi(P_{1|x}, P_{0|x})$ be such that $\theta_x^L = E_{\pi_x^L}[\mathbb{1}\{Y_1 - Y_0 < \delta\}] = P_{\pi_x^L}(Y_1 - Y_0 < \delta)$ and $\theta_x^H = 1 - E_{\pi_x^H}[\mathbb{1}\{Y_1 - Y_0 > \delta\}] = P_{\pi_x^H}(Y_1 - Y_0 \leq \delta)$. Notice that $\theta_x^H \in \Theta_{I,x}$.

Furthermore, $\mathbb{1}\{y_1 - y_0 < \delta\} \leq \mathbb{1}\{y_1 - y_0 \leq \delta\}$ implies

$$\theta_x^L = \inf_{\pi_x \in \Pi(P_{1|x}, P_{0|x})} E_{\pi_x}[\mathbb{1}\{Y_1 - Y_0 < \delta\}] \leq \inf_{\pi_x \in \Pi(P_{1|x}, P_{0|x})} E_{\pi_x}[\mathbb{1}\{Y_1 - Y_0 \leq \delta\}]$$

and thus θ_x^L is a lower bound for $\Theta_{I,x}$. Since $\Theta_{I,x}$ is convex, it suffices to show that $\theta_x^L \in \Theta_{I,x}$.

Corollary 1.9.44 implies $\theta_x^L = P_{\pi_x^L}(Y_1 - Y_0 < \delta) = \sup_y \left\{F_{1|x}(y) - F_{0|x}(y - \delta)\right\}$. Moreover, Villani (2009) theorem 5.10 part (iii) implies the dual problem $\sup_y \left\{F_{1|x}(y) - F_{0|x}(y - \delta)\right\}$ is attained as well, say by y^*. Thus

$$\int \mathbb{1}\{y_1 - y_0 \leq \delta\} d\pi_x^L(y_1, y_0) = \int \mathbb{1}\{y_1 \leq y^*\} dP_{1|x}(y_1) - \int \mathbb{1}\{y_0 \leq y^* - \delta\} dP_{0|x}(y_0) \tag{1.42}$$

Next, notice that

$$\mathbb{1}\{y_1 \leq y^*\} - \mathbb{1}\{y_0 \leq y^* - \delta\} \leq \mathbb{1}\{y_1 - y_0 < \delta\} \tag{1.43}$$

which holds for all (y_1, y_0), must hold with equality π_x^L-almost surely. Indeed, let N be the set where the inequality in (1.43) is strict and suppose N is π_x^L-non-

negligible. Since $\pi_x^L \in \Pi(P_{1|x}, P_{0|x})$,

$$\begin{aligned}
&\int \mathbb{1}\{y_1 \leq y^*\} dP_{1|x}(y_1) - \int \mathbb{1}\{y_0 \leq y^* - \delta\} dP_{0|x}(y_0) \\
&\quad = \int \mathbb{1}\{y_1 \leq y^*\} - \mathbb{1}\{y_0 \leq y^* - \delta\} d\pi_x^L(y_1, y_0) \\
&\quad = \int_N \mathbb{1}\{y_1 \leq y^*\} - \mathbb{1}\{y_0 \leq y^* - \delta\} d\pi_x^L(y_1, y_0) \\
&\qquad\qquad + \int_{N^c} \mathbb{1}\{y_1 \leq y^*\} - \mathbb{1}\{y_0 \leq y^* - \delta\} d\pi_x^L(y_1, y_0) \\
&\quad < \int_N \mathbb{1}\{y_1 - y_0 < \delta\} d\pi_x^L(y_1, y_0) + \int_{N^c} \mathbb{1}\{y_1 - y_0 < \delta\} d\pi_x^L(y_1, y_0) \\
&\quad = \int \mathbb{1}\{y_1 - y_0 \leq \delta\} d\pi_x^L(y_1, y_0)
\end{aligned}$$

contradicts (1.42). This implies that π_x^L concentrates on

$$\begin{aligned}
&\underbrace{\{(y_1, y_0)\ ;\ y_1 \leq y^*, y_0 > y^* - \delta, y_1 - y_0 < \delta\}}_{\text{both sides of (1.43) equal 1}} \\
&\qquad \cup \underbrace{\{(y_1, y_0)\ ;\ y_1 > y^*, y_0 > y^* - \delta, y_1 - y_0 \geq \delta\}}_{\text{both sides of (1.43) equal 0}} \\
&\qquad \cup \underbrace{\{(y_1, y_0)\ ;\ y_1 \leq y^*, y_0 \leq y^* - \delta, y_1 - y_0 \geq \delta\}}_{\text{both sides of (1.43) equal 0}}
\end{aligned}$$

Notice the only point in the set $\{(y_1, y_0)\ ;\ y_1 - y_0 = \delta\}$ where π_x^L could put positive mass is the point $(y_1, y_0) = (y^*, y^* - \delta)$. But since $P_{1|x}$ has a continuous CDF,

$$0 \leq \pi_x^L(\{(y^*, y^* - \delta)\}) \leq \pi_x^L(\{y^*\} \times \mathcal{Y}_0) = P_{1|x}(\{y^*\}) = 0$$

Thus $P_{\pi_x^L}(Y_1 - Y_0 = \delta) = 0$, and so $P_{\pi_x^L}(Y_1 - Y_0 \leq \delta) = P_{\pi_x^L}(Y_1 - Y_0 < \delta) = \theta^L(x)$. Thus $\theta_x^L \in \Theta_{I,x}$, and hence $\Theta_{I,x} = [\theta^L(x), \theta^H(x)]$.

Therefore the identified set for θ_x is $[\theta_x^L, \theta_x^H]$. It follows from this and step one above that the identified set $(\theta_{x_1}, \ldots, \theta_{x_M})$ is $[\theta_{x_1}^L, \theta_{x_1}^H] \times \ldots \times [\theta_{x_M}^L, \theta_{x_M}^H]$.

3. Recall that $\theta = E[c(Y_1, Y_0) \mid T = c] = E[E[c(Y_1, Y_0) \mid T = c, X]] = \sum_x s_x \theta_x$. Since $s_x = P(X = x \mid T = c)$ is point identified for each x, it follows from step two above that the identified set for θ is $[\theta^L, \theta^H]$ where

$$\theta^L = \sum_x s_x \theta_x^L, \qquad \theta^H = \sum_x s_x \theta_x^H$$

This concludes the proof. □

Theorem 1.4.1 (Identification of functions of moments)**.** *Suppose assumptions 1, 2, and 3 are satisfied. Then the sharp identified set for γ is $[\gamma^L, \gamma^H]$.*

Proof. Lemma 1.9.1 shows that under assumptions 1 and 2, the sharp identified set for θ is $[\theta^L, \theta^H]$. Let Γ_I be the identified set for γ, and note that

$$\Gamma_I = \{\gamma \in \mathbb{R} \,;\, \gamma = g(t, \eta) \text{ for some } t \in [\theta^L, \theta^H]\}$$

Assumption 2 implies c is bounded; under assumption 2 (i) the continuous $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ takes a maximum and minimum on the compact set $\mathcal{Y} \times \mathcal{Y}$, while under assumption 2 (ii) the cost function only takes values 0 or 1. It follows that θ^L and θ^H are finite and thus $[\theta^L, \theta^H]$ is compact.

Assumption 3 (ii) is that $g(\cdot, \eta)$ is continuous, and thus the extreme value theorem implies $\gamma^L = \inf_{t \in [\theta^L, \theta^H]} g(t, \eta)$ and $\gamma^H = \sup_{t \in [\theta^L, \theta^H]} g(t, \eta)$ are both elements of Γ_I. The intermediate value theorem then implies $\Gamma_I = [\gamma^L, \gamma^H]$. □

1.9.1.2 Quantiles

Example 1.2.5 considers the parameter q_τ solving

$$P(Y_1 - Y_0 \leq q_\tau \mid D_1 > D_0) = \tau$$

As noted in that example, the sharp identification results for $P(Y_1 - Y_0 \leq \delta \mid D_1 > D_0)$ can be adapted to characterize the sharp identified set for q_τ. First view the bounds on the cumulative distribution function as functions of δ:

$$
\begin{aligned}
&c_{L,\delta}(y_1, y_0) = \mathbb{1}\{y_1 - y_0 < \delta\}, && c_{H,\delta}(y_1, y_0) = \mathbb{1}\{y_1 - y_0 > \delta\}, \\
&\theta_x^L(\delta) = OT_{c_{L,\delta}}(P_{1|x}, P_{0|x}), && \theta_x^H(\delta) = 1 - OT_{c_{H,\delta}}(P_{1|x}, P_{0|x}) \\
&\theta^L(\delta) = \sum_x s_x \theta_x^L(\delta) && \theta^H(\delta) = \sum_x s_x \theta_x^H(\delta)
\end{aligned}
$$

Let $Q_{I,\tau}$ denote the sharp identified set for q_τ.

Lemma 1.9.2 (Identification of q_τ)**.** *Suppose assumptions 1 and 2 (ii) hold. Then $q \in Q_{I,\tau}$ if and only if $\theta^L(q) \leq \tau \leq \theta^H(q)$.*

Proof. By definition, $q \in \Gamma_{I,\tau}$ if and only if there exists a distribution of the primitives, π, consistent with the observed distribution, such that $P_\pi(Y_1 - Y_0 \leq q) = \tau$. Lemma 1.9.1 shows that $\theta^L(q) \leq \tau \leq \theta^H(q)$ if and only if there exists a distribution of the primitives, π, such that $P_\pi(Y_1 - Y_0 \leq q) = \tau$. This concludes the proof. □

Lemma 1.9.2 implies that inverting a test of $H_0 : \theta^L(q) \leq \tau \leq \theta^H(q)$ against the alternative $H_1 : \tau < \theta^L(q)$ or $\theta^H(q) < \tau$ will lead to valid confidence sets for q_τ.

Consider instead defining q_τ to be the closed subset of $\mathbb{R}$ given by

$$
q_\tau = [\inf\{y \,;\, P(Y_1 - Y_0 \leq y) \geq \tau\}, \inf\{y \,;\, P(Y_1 - Y_0 \leq y) > \tau\}]
$$

Note that this q_τ is the singleton $\inf\{y \,;\, P(Y_1 - Y_0 \leq y) \geq \tau\}$, unless $P(Y_1 - Y_0 \leq \cdot)$ is flat when equal to τ, in which case it equals the τ-level set $\{y \,;\, P(Y_1 - Y_0 \leq y) = \tau\}$. (Compare Ehm et al. (2016), who define the τ-th quantile equivalently as $q_\tau = [\sup\{y \,;\, P(Y_1 - Y_0 \leq y) < \tau\}, \sup\{y \,;\, P(Y_1 - Y_0 \leq y) \leq \tau\}]$.)

Lemma 1.9.3 (Identification: τ-th quantile)**.** *Let q_τ be defined as*

$$q_\tau \equiv [\inf\{y\ ;\ P(Y_1 - Y_0 \leq y) \geq \tau\}, \inf\{y\ ;\ P(Y_1 - Y_0 \leq y) > \tau\}] \tag{1.44}$$

Suppose assumption 1 and 2 (ii) hold, and let $Q_{I,\tau}$ denote the identified set of q_τ defined by (1.44)*. Then $q \in Q_{I,\tau}$ if and only if $\theta^L(q) \leq \tau \leq \theta^H(q)$.*

Proof. Suppose $\theta^L(q) \leq \tau \leq \theta^H(q)$. Lemma 1.9.1 implies there exists a distribution π of the primitives consistent with assumption 2 (ii) such that $P_\pi(Y_1 - Y_0 \leq q) = \tau$. Thus $q \in [\inf\{y\ ;\ P_\pi(Y_1 - Y_0 \leq y) \geq \tau\}, \inf\{y\ ;\ P_\pi(Y_1 - Y_0 \leq y) > \tau\}]$ and hence $q \in Q_{I,\tau}$.

Before showing the other direction, we next show that assumption 2 (ii) implies $\theta^L(\delta)$ is continuous. Specifically, apply corollary 1.9.44 to find $\theta^L_x(\delta) = \sup_y\{F_{1|x}(y) - F_{0|x}(y-\delta)\}$. So for any δ, δ',

$$\begin{aligned}
\theta^L_x(\delta) - \theta^L_x(\delta') &= \sup_y\{F_{1|x}(y) - F_{0|x}(y-\delta)\} - \sup_y\{F_{1|x}(y) - F_{0|x}(y-\delta')\} \\
&\leq \sup_y \left\{F_{0|x}(y-\delta') - F_{0|x}(y-\delta)\right\} \\
&\leq \sup_y \left|F_{0|x}(y-\delta') - F_{0|x}(y-\delta)\right|
\end{aligned}$$

and thus $|\theta^L_x(\delta) - \theta^L_x(\delta')| \leq \sup_y \left|F_{0|x}(y-\delta') - F_{0|x}(y-\delta)\right|$. Recall that any continuous CDF is in fact uniformly continuous, and so $F_{0|x}$ is in fact uniformly continuous. Let $\varepsilon > 0$, choose $\eta > 0$ such that for any $y, y' \in \mathbb{R}$ with $|y - y'| < \eta$, one has $|F_{0|x}(y) - F_{0|x}(y')| < \varepsilon/2$, and notice that

$$|\delta - \delta'| < \eta \implies \sup_y \left|F_{0|x}(y-\delta') - F_{0|x}(y-\delta)\right| \leq \varepsilon/2 < \varepsilon$$

This shows $\theta^L_x(\delta)$ is continuous, and so $\theta^L(\delta) = \sum_x s_x \theta^L_x$ is continuous.

Return to showing the other direction, through the contrapositive. Suppose it is not the

case that $\theta^L(q) \leq \tau \leq \theta^H(q)$. There are two possibilities:

1. Suppose $\theta^H(q) < \tau$. Then there is no distribution π of the primitives such that $P_\pi(Y_1 - Y_0 \leq q) \geq \tau$, hence there is no distribution where $q \in [\inf\{y \ ; \ P(Y_1 - Y_0 \leq y) \geq \tau\}, \inf\{y \ ; \ P(Y_1 - Y_0 \leq y) > \tau\}]$ and thus $q \notin Q_{I,\tau}$.

2. Suppose $\tau < \theta^L(q)$. If one further supposes that $q \in Q_{I,\tau}$, then $\theta^L(\cdot)$ would have a jump discontinuity at q, contradicting the continuity shown above.

   Specifically, if $\tau < \theta^L(q)$ and $q \in Q_{I,\tau}$, then there exists a distribution π of the primitives such that $P_\pi(Y_1 - Y_0 \leq q) > \tau$ and $q \in [\inf\{y \ ; \ P_\pi(Y_1 - Y_0 \leq y) \geq \tau\}, \inf\{y \ ; \ P_\pi(Y_1 - Y_0 \leq y) > \tau\}]$, implying that $P_\pi(Y_1 - Y_0 \leq \cdot)$ jumps at q from below τ to above $\theta^L(q)$:

$$\lim_{\epsilon \to 0} P_\pi(Y_1 - Y_0 \leq q - \epsilon) < \tau < \theta^L(q) \leq P_\pi(Y_1 - Y_0 \leq q)$$

   This jump discontinuity at q is at least of size $\varepsilon = \theta^L(q) - \tau > 0$. But then $\theta^L(\cdot)$ would have a jump discontinuity of at least size ε at q as well, a contradiction of the continuity of $\theta^L(\cdot)$ shown above.

   Thus if $\tau < \theta^L(q)$, then $q \notin Q_{I,\tau}$.

In either case, $q \notin Q_{I,\tau}$. This completes the proof. □

1.9.1.3 Additional identification lemmas

The lemmas below contain results well known in the literature. They are included here with proofs for completeness.

Lemma 1.9.4. *Let P_1 be any distribution and P_0 be degenerate at $\tilde{y}_0 \in \mathbb{R}$. Then the only*

possible coupling of P_1 and P_0 is characterized by the cumulative distribution function

$$P(Y_1 \leq y_1, Y_0 \leq y_0) = \begin{cases} P(Y_1 \leq y_1) & \text{if } y_0 \geq \tilde{y}_0 \\ 0 & \text{if } y_0 < \tilde{y}_0 \end{cases}$$

Proof. First suppose $y_0 < \tilde{y}_0$. Then $0 \leq P(Y_1 \leq y_1, Y_0 \leq y_0) \leq P(Y_0 \leq y_0) = 0$.

Next suppose $y_0 \geq \tilde{y}_0$. Then $1 \geq P(\{Y_1 \leq y_1\} \cup \{Y_0 \leq y_0\}) \geq P(Y_0 \leq y_0) = 1$ implies that

$$\begin{aligned} P(Y_1 \leq y_1, Y_0 \leq y_0) &= P(Y_1 \leq y_1) + \underbrace{P(Y_0 \leq y_0)}_{=1} - \underbrace{P(\{Y_1 \leq y_1\} \cup \{Y_0 \leq y_0\})}_{=1} \\ &= P(Y_1 \leq y_1) \end{aligned}$$

which completes the proof. □

Lemma 1.9.5 below summarizes the empirical content of the model described in assumption 1. In particular, it implies that any two distributions of the primitives consistent with assumption 1 that share the same marginal distribution of (T, Z, X) and marginal, conditional distributions of

$$\begin{aligned} &Y_1 \mid T = a, X = x && Y_0 \mid T = n, X = x \\ &Y_1 \mid T = c, X = x, && Y_0 \mid T = c, X = x \end{aligned}$$

will produce the same distribution of observables.

Lemma 1.9.5. *Suppose assumpion 1 holds. Then*

$$P(D = 1 \mid Z = 0, X = x) = P(T = a \mid X = x)$$
$$P(D = 0 \mid Z = 1, X = x) = P(T = n \mid X = x)$$
$$P(D = 1 \mid Z = 1, X = x) = P(T \in \{a, c\} \mid X = x)$$
$$P(D = 0 \mid Z = 0, X = x) = P(T \in \{c, n\} \mid X = x)$$

and for any integrable function f,

$$E[f(Y) \mid D = 1, Z = 1, X = x] = E[f(Y_1) \mid T \in \{a, c\}, X = x]$$
$$E[f(Y) \mid D = 0, Z = 0, X = x] = E[f(Y_0) \mid T \in \{c, n\}, X = x]$$

Furthermore,

if $P(D = 1 \mid Z = 0, X = x) > 0$,
then $E[f(Y) \mid D = 1, Z = 0, X = x] = E[f(Y_1) \mid T = a, X = x]$
if $P(D = 0 \mid Z = 1, X = x) > 0$,
then $E[f(Y) \mid D = 0, Z = 1, X = x] = E[f(Y_0) \mid T = n, X = x]$

Proof. Assumption 1 (ii) implies $\mathbb{1}\{D_1 = 0, D_0 = 1\} = 0$. The definition of T in (1.41) then implies

$$\mathbb{1}\{D_0 = 1\} = \mathbb{1}\{D_1 = 1, D_0 = 1\} + \cancel{\mathbb{1}\{D_1 = 0, D_0 = 1\}} = \mathbb{1}\{T = a\}$$
$$\mathbb{1}\{D_1 = 0\} = \mathbb{1}\{D_1 = 0, D_0 = 0\} + \cancel{\mathbb{1}\{D_1 = 0, D_0 = 1\}} = \mathbb{1}\{T = n\}$$
$$\mathbb{1}\{D_1 = 1\} = \mathbb{1}\{D_1 = 1, D_0 = 1\} + \mathbb{1}\{D_1 = 1, D_0 = 0\} = \mathbb{1}\{T \in \{a, c\}\}$$
$$\mathbb{1}\{D_0 = 0\} = \mathbb{1}\{D_1 = 1, D_0 = 0\} + \mathbb{1}\{D_1 = 0, D_0 = 0\} = \mathbb{1}\{T \in \{c, n\}\}$$

These observations, equation (1.2), and assumption 1 (i) imply

$$
\begin{aligned}
&P(D=1 \mid Z=0, X=x) = P(D_0=1 \mid X=x) = P(T=a \mid X=x),\\
&P(D=0 \mid Z=1, X=x) = P(D_1=0 \mid X=x) = P(T=n \mid X=x),\\
&P(D=1 \mid Z=1, X=x) = P(D_1=1 \mid X=x) = P(T \in \{a,c\} \mid X=x), \text{ and}\\
&P(D=0 \mid Z=0, X=x) = P(D_0=0 \mid X=x) = P(T \in \{c,n\} \mid X=x)
\end{aligned}
$$

Note the first two equalities can be summarized as $P(D=d \mid Z=z, X=x) = P(D_z=d \mid X=x)$.

Next, let $f:\mathbb{R}\to\mathbb{R}$ be integrable. Assumption 1 (i) and equations (1.1) and (1.2) imply that for any (d,z,x),

$$
\begin{aligned}
&P(D=d \mid Z=z, X=x)E[f(Y) \mid D=d, Z=z, X=x]\\
&\quad = P(D_z=d \mid X=x)E[f(Y_d) \mid D_z=d, X=x]
\end{aligned}
$$

and since $P(D=d \mid Z=z, X=x) = P(D_z=d \mid X=x)$, this implies

$$
0 = P(D=d \mid Z=z, X=x)\Big(E[f(Y) \mid D=d, Z=z, X=x] - E[f(Y_d) \mid D_z=d, X=x]\Big) \tag{1.45}
$$

Assumption 1 (iii) implies

$$
\begin{aligned}
&P(D=1 \mid Z=1, X=x) = P(T \in \{a,c\} \mid X=x) \geq P(T=c \mid X=x) > 0\\
&P(D=0 \mid Z=0, X=x) = P(T \in \{c,n\} \mid X=x) \geq P(T=c \mid X=x) > 0
\end{aligned}
$$

Use strict positivity of $P(D=1 \mid Z=1, X=x)$ and $P(D=0 \mid Z=0, X=x)$ to see that

$$E[f(Y) \mid D=1, Z=1, X=x] = E[f(Y_1) \mid D_1=1, X=x] = E[f(Y_1) \mid T \in \{a,c\}, X=x]$$
$$E[f(Y) \mid D=0, Z=0, X=x] = E[f(Y_0) \mid D_0=0, X=x] = E[f(Y_0) \mid T \in \{c,n\}, X=x]$$

Similarly, (1.45) implies

$$\begin{aligned}
&\text{if } P(D=1 \mid Z=0, X=x) > 0, \\
&\qquad \text{then } E[f(Y) \mid D=1, Z=0, X=x] = E[f(Y_1) \mid T=a, X=x] \\
&\text{if } P(D=0 \mid Z=1, X=x) > 0, \\
&\qquad \text{then } E[f(Y) \mid D=0, Z=1, X=x] = E[f(Y_0) \mid T=n, X=x]
\end{aligned}$$

this concludes the proof. □

Lemma 1.2.1 (Abadie (2003))**.** *Suppose assumption 1 holds. Then the marginal distributions of Y_d conditional on $D_1 > D_0$ and $X = x$, denoted $P_{d|x}$, are identified by*

$$\begin{aligned}
E_{P_{d|x}}[f(Y_d)] &\equiv E[f(Y_d) \mid D_1 > D_0, X=x] \\
&= \frac{E[f(Y)\mathbb{1}\{D=d\} \mid Z=d, X=x] - E[f(Y)\mathbb{1}\{D=d\} \mid Z=1-d, X=x]}{P(D=d \mid Z=d, X=x) - P(D=d \mid Z=1-d, X=x)}
\end{aligned} \tag{1.4}$$

for any integrable function f. Furthermore, the distribution of X conditional on $D_1 > D_0$ is identified by

$$\begin{aligned}
s_x &\equiv P(X=x \mid D_1 > D_0) \\
&= \frac{[P(D=1 \mid Z=1, X=x) - P(D=1 \mid Z=0, X=x)]\, P(X=x)}{\sum_{x'} [P(D=1 \mid Z=1, X=x') - P(D=1 \mid Z=0, X=x')]\, P(X=x')}
\end{aligned} \tag{1.5}$$

Proof. First notice that using T as defined in (1.41),

$$E[f(Y_d) \mid D_1 > D_0, X = x] = E[f(Y_d) \mid T = c, X = x] = \frac{E[f(Y_d)\mathbb{1}\{T = c\} \mid X = x]}{P(T = c \mid X = x)} \tag{1.46}$$

Now notice that

$$D_1 - D_0 = (1 - D_0) - (1 - D_1) = \mathbb{1}\{D_d = d\} - \mathbb{1}\{D_{1-d} = d\}$$

for either $d \in \{1, 0\}$. Monotonicity (assumption 1 (ii)) implies that this is an indicator for $T = c$:

$$D_1 - D_0 = \mathbb{1}\{D_1 = 1, D_0 = 0\} = \mathbb{1}\{T = c\}$$

So,

$$\begin{aligned} &E[f(Y)\mathbb{1}\{D = d\} \mid Z = d, X = x] - E[f(Y)\mathbb{1}\{D = d\} \mid Z = 1 - d, X = x] \\ &\quad = E[f(Y_d)\mathbb{1}\{D_d = d\} \mid X = x] - E[f(Y_d)\mathbb{1}\{D_{1-d} = d\} \mid X = x] \\ &\quad = E[f(Y_d)(\mathbb{1}\{D_d = d\} - \mathbb{1}\{D_{1-d} = d\}) \mid X = x] \\ &\quad = E[f(Y_d)\mathbb{1}\{T = c\} \mid X = x] \end{aligned} \tag{1.47}$$

Lemma 1.9.5 shows that

$$\begin{aligned} &P(D = 1 \mid Z = 1, X = x) - P(D = 1 \mid Z = 0, X = x) \\ &\quad = P(T \in \{a, c\} \mid X = x) - P(T = a \mid X = x) = P(T = c \mid X = x) \end{aligned}$$

and similarly,

$$\begin{aligned} &P(D=0 \mid Z=0, X=x) - P(D=0 \mid Z=1, X=x) \\ &\quad = P(T \in \{c,n\} \mid X=x) - P(T=n \mid X=x) = P(T=c \mid X=x) \end{aligned}$$

Thus for either $d \in \{1,0\}$,

$$P(D=d \mid Z=d, X=x) - P(D=d \mid Z=1-d, X=x) = P(T=c \mid X=x). \quad (1.48)$$

It follows from (1.46), (1.47), and (1.48) that

$$\begin{aligned} E_{P_{d|x}}[f(Y_d)] &= E[f(Y_d) \mid D_1 > D_0, X=x] \\ &= \frac{E[f(Y)\mathbb{1}\{D=d\} \mid X=x, Z=d] - E[f(Y)\mathbb{1}\{D=d\} \mid X=x, Z=1-d]}{P(D=d \mid X=x, Z=d) - P(D=d \mid X=x, Z=1-d)}, \end{aligned}$$

and from (1.48) that

$$\begin{aligned} s_x = P(X=x \mid D_1 > D_0) = P(X=x \mid T=c) &= \frac{P(T=c \mid X=x)P(X=x)}{\sum_{x'} P(T=c \mid X=x')P(X=x')} \\ = \frac{\left[P(D=1 \mid X=x, Z=1) - P(D=1 \mid X=x, Z=0)\right] P(X=x)}{\sum_{x'} \left[P(D=1 \mid X=x', Z=1) - P(D=1 \mid X=x', Z=0)\right] P(X=x')}&. \end{aligned}$$

This concludes the proof. □

1.9.2 Appendix: properties of optimal transport

Suppose that strong duality holds:

$$\inf_{\pi \in \Pi(P_1,P_0)} \int c(y_1,y_0) d\pi(y_1,y_0) = \sup_{(\varphi,\psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)} \int \varphi(y_1) dP_1(y_1) + \int \psi(y_0) dP_0(y_0) \quad (1.49)$$

for sets of universally bounded functions $\mathcal{F}_c \subseteq L^1(P_1)$ and $\mathcal{F}_c^c \subseteq L^1(P_0)$. See lemmas 1.9.38 and 1.9.42 for examples.[4] Then for suitable sets $\mathcal{F}_1$ and $\mathcal{F}_0$ with $\mathcal{F}_c \subseteq \mathcal{F}_1$ and $\mathcal{F}_c^c \subseteq \mathcal{F}_0$, the map $OT_c(P_1, P_0) = \inf_{\pi \in \Pi(P_1,P_0)} \int c(y_1, y_0) d\pi(y_1, y_0)$ can be viewed as

$$OT_c : \ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0) \to \mathbb{R}, \qquad OT_c(P_1, P_0) = \sup_{(\varphi,\psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)} P_1(\varphi) + P_0(\psi) \tag{1.50}$$

where $P_d(f) = \int f(y_d) dP_d(y_d) = E_{P_d}[f(Y_d)]$.

The functional in (1.50) is defined over the familiar Banach space $\ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0)$. This makes it straightforward to show that optimal transport, as a functional from this space to $\mathbb{R}$, has certain desirable properties.

1.9.2.1 Continuity

Lemma 1.9.6 (Optimal transport is uniformly continuous)**.** *Suppose that for some universally bounded $\mathcal{F}_c \subseteq L^1(P_1)$ and $\mathcal{F}_c^c \subseteq L^1(P_0)$,* (1.49) *holds. Then the optimal transport functional, given by* (1.50)*, is uniformly continuous.*

Proof. Define

$$\mathcal{S} : \ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0) \to \ell^\infty(\mathcal{F}_1 \times \mathcal{F}_0), \qquad \mathcal{S}(H_1, H_0)(\varphi, \psi) = H_1(\varphi) + H_0(\psi)$$
$$\Xi_c : \ell^\infty(\mathcal{F}_1 \times \mathcal{F}_0) \to \mathbb{R}, \qquad \Xi_c[G] = \sup_{(\varphi,\psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)} G(\varphi, \psi)$$

and notice that $OT_c(H_1, H_0) = \Xi_c(\mathcal{S}(H_1, H_0))$. Since $s : \mathbb{R}^2 \to \mathbb{R}$ given by $s(h_1, h_2) = h_1 + h_2$

[4] $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are typically found with the following steps:

(i) Start with a known strong duality result; for some $\Phi_{cs} \subseteq \Phi_c$,

$$\inf_{\pi \in \Pi(P_1,P_0)} \int c(y_1, y_0) d\pi(y_1, y_0) = \sup_{(\varphi,\psi) \in \Phi_{cs}} \int \varphi(y_1) dP_1(y_1) + \int \psi(y_0) dP_0(y_0)$$

(ii) Compute $\mathcal{F}_c(\Phi_{cs})$ and $\mathcal{F}_c^c(\Phi_{cs})$ defined by (1.85).

(iii) Notice that $\mathcal{F}_c(\Phi_{cs}) \subseteq \mathcal{F}_c$ and $\mathcal{F}_c^c(\Phi_{cs}) \subseteq \mathcal{F}_c^c$ for known and easy to study sets $\mathcal{F}_c$, $\mathcal{F}_c^c$

Lemma 1.9.36 and remark 1.9.2 are useful to ensure $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are universally bounded.

is uniformly continuous, we have that $\mathcal{S}$ is uniformly continuous (see lemma 1.9.47). Lemma 1.9.49 shows that Ξ_c is uniformly continuous. The composition of uniformly continuous functions is uniformly continuous, implying OT_c is uniformly continuous. This completes the proof. □

1.9.2.2 Directional Differentiability

The optimal transport functional given by (1.50) is Hadamard directionally differentiable. The formal result, stated below, requires that $\mathcal{F}_c$ and $\mathcal{F}_c^c$ each be equipped with a semimetric. The semimetrics chosen must be such that $P_1 \in \ell^\infty(\mathcal{F}_c)$ and $P_0 \in \ell^\infty(\mathcal{F}_c^c)$ are continuous and the product space $\mathcal{F}_c \times \mathcal{F}_c^c$ and its subset $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ are compact.

The setting suggests a very convenient semimetric. Let P be the distribution of an observation, i.e. $(Y, D, Z, X) \sim P$. Note that under assumption 1, the distributions $P_{d|x}$ are dominated by P with bounded densities $\frac{dP_{d|x}}{dP}$. Specifically, recall that

$$\begin{aligned} E_{P_{d|x}}[f(Y_d)] &= E[f(Y_d) \mid D_1 > D_0, X = x] \\ &= \frac{E[f(Y)\mathbb{1}\{D = d\} \mid Z = d, X = x] - E[f(Y)\mathbb{1}\{D = d\} \mid Z = 1-d, X = x]}{P(D = d \mid Z = d, X = x) - P(D = d \mid Z = 1-d, X = x)} \end{aligned}$$

Let $\mathbb{1}_{d,x,z}(D, X, Z) = \mathbb{1}\{D = d, X = x, Z = z\}$, $p_{d,x,z} = P(D = d, X = x, Z = z)$, and $p_{x,z} = P(X = x, Z = z)$. Observe that

$$\begin{aligned} E[f(Y_d) \mid D_1 > D_0, X = x] &= E\left[f(Y)\frac{\mathbb{1}_{d,x,d}(D, X, Z)/p_{x,d} - \mathbb{1}_{d,x,1-d}(D, X, Z)/p_{x,1-d}}{p_{d,x,d}/p_{x,d} - p_{d,x,1-d}/p_{x,1-d}}\right] \\ &= E\left[f(Y)E\left[\frac{\mathbb{1}_{d,x,d}(D, X, Z)/p_{x,d} - \mathbb{1}_{d,x,1-d}(D, X, Z)/p_{x,1-d}}{p_{d,x,d}/p_{x,d} - p_{d,x,1-d}/p_{x,1-d}} \mid Y\right]\right] \end{aligned}$$

reveals the densities to be $\frac{dP_{d|x}}{dP}(Y) = E\left[\frac{\mathbb{1}_{d,x,d}(D,X,Z)/p_{x,d} - \mathbb{1}_{d,x,1-d}(D,X,Z)/p_{x,1-d}}{p_{d,x,d}/p_{x,d} - p_{d,x,1-d}/p_{x,1-d}} \mid Y\right]$.

We now drop the subscript x for the remainder of this appendix. Because P dominates

both P_1 and P_0 with bounded densities, the $L_{2,P}$ semimetric works very well:

$$L_{2,P}(f_1, f_2) = \sqrt{P((f_1 - f_2)^2)} = \sqrt{E_P[(f_1(Y) - f_2(Y))^2]} \tag{1.51}$$

Equip the product space $\mathcal{F}_1 \times \mathcal{F}_0$ with the product semimetric:

$$L_2((f_1, g_1), (f_2, g_2)) = \sqrt{L_{2,P}(f_1, f_2)^2 + L_{2,P}(g_1, g_2)^2} \tag{1.52}$$

To apply the $L_{2,P}$ semimetric, each $f \in \mathcal{F}_1$ and $f \in \mathcal{F}_0$ are defined on whole domain $\mathcal{Y}$.

Lemma 1.9.7 (Hadamard directional differentiability of optimal transport)**.** *Let $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be lower semicontinuous, $\mathcal{F}_1, \mathcal{F}_0$ be sets of measurable functions mapping $\mathcal{Y}$ to $\mathbb{R}$, and $\mathcal{F}_c \subseteq \mathcal{F}_1$ and $\mathcal{F}_c^c \subseteq \mathcal{F}_0$ be universally bounded subsets. Suppose that*

1. *Strong duality holds:*

$$\inf_{\pi \in \Pi(P_1, P_0)} \int c(y_1, y_0) d\pi(y_1, y_0) = \sup_{(\varphi, \psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)} \int \varphi(y_1) dP_1(y_1) + \int \psi(y_0) dP_0(y_0),$$

2. *P dominates P_1 and P_0 with bounded densities,*

3. *$\mathcal{F}_d$ is P-Donsker and $\sup_{f \in \mathcal{F}_d} |P(f)| < \infty$ for each $d = 1, 0$, and*

4. *$(\mathcal{F}_1 \times \mathcal{F}_0, L_2)$ and the subset*

$$\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c) = \{(\varphi, \psi) \in \mathcal{F}_c \times \mathcal{F}_c^c \; ; \; \varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)\}$$

   are complete.

Then $OT_c : \ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0) \to \mathbb{R}$ defined by

$$OT_c(P_1, P_0) = \sup_{(\varphi, \psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)} P_1(\varphi) + P_0(\psi)$$

is Hadamard directionally differentiable at (P_1, P_0) *tangentially to*

$$\mathbb{D}_{Tan} = \mathcal{C}(\mathcal{F}_1, L_{2,P}) \times \mathcal{C}(\mathcal{F}_0, L_{2,P}). \tag{1.53}$$

The set of maximizers $\Psi_c(P_1, P_0) = \arg\max_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} P_1(\varphi) + P_0(\psi)$ *is nonempty, and the derivative* $OT'_{c,(P_1,P_0)} : \mathbb{D}_{Tan} \to \mathbb{R}$ *is given by*

$$OT'_{c,(P_1,P_0)}(H_1, H_0) = \sup_{(\varphi,\psi)\in\Psi_c(P_1,P_0)} H_1(\varphi) + H_0(\psi)$$

Proof. For legibility, the proof is broken down into four steps:

1. Define

$$\mathcal{S} : \ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0) \to \ell^\infty(\mathcal{F}_1 \times \mathcal{F}_0), \qquad \mathcal{S}(H_1, H_0)(\varphi, \psi) = H_1(\varphi) + H_0(\psi)$$
$$\Xi_c : \ell^\infty(\mathcal{F}_1 \times \mathcal{F}_0) \to \mathbb{R}, \qquad \Xi_c[G] = \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} G(\varphi, \psi)$$

   and notice that $OT_c(H_1, H_0) = \Xi_c(\mathcal{S}(H_1, H_0))$. This suggests application of the chain rule.

2. $\mathcal{S}$ is linear and continuous at every point of $\ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0)$, which implies it is (fully) Hadamard differentiable at any $(H_1, H_0) \in \ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0)$ tangentially to $\ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0)$, and is its own derivative. Indeed, for any $(H_{1n}, H_{0n}) \to (H_1, H_0) \in \ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0)$ and any $t_n \downarrow 0$,

$$\lim_{n\to\infty} \left\| \frac{\mathcal{S}((H_1, H_0) + t_n(H_{1n}, H_{0n})) - \mathcal{S}(H_1, H_0)}{t_n} - \mathcal{S}(H_1, H_0) \right\|_{\mathcal{F}_c\times\mathcal{F}_c^c}$$
$$= \lim_{n\to\infty} \|\mathcal{S}(H_{1n}, H_{0n}) - \mathcal{S}(H_1, H_0)\|_{\mathcal{F}_c\times\mathcal{F}_c^c} = 0$$

3. Consider Ξ_c. Verify the conditions of lemma 1.9.55:

   (a) $(\mathcal{F}_1 \times \mathcal{F}_0, L_2)$ and the subset $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ are compact.

First recall that a subset of semimetric space is compact if and only if it is totally bounded and complete.[5] Completeness of both sets is assumed, so it suffices to show they are totally bounded. Since $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ is a subset of $\mathcal{F}_1 \times \mathcal{F}_0$, it suffices to show the latter set is totally bounded.

Using the assumption that $\mathcal{F}_d$ is P-Donsker and $\sup_{f \in \mathcal{F}_d} |P(f)| < \infty$, we have that $\sup_{\varphi \in \mathcal{F}_c} |P(\varphi)| < \infty$ and $(\mathcal{F}_d, L_{2,P})$ is totally bounded (see van der Vaart & Wellner (1997) problem 2.1.2.). It follows that the product space $(\mathcal{F}_1 \times \mathcal{F}_0, L_2)$ is totally bounded.[6]

(b) $\mathcal{S}(P_1, P_0) \in \mathcal{C}(\mathcal{F}_1 \times \mathcal{F}_0, L_2)$.

Notice that

$$|P_1(f_1) - P_1(f_2)| \leq P_1(|f_1 - f_2|) \leq \sqrt{P_1((f_1 - f_2)^2)} = L_{2,P_1}(f_1, f_2)$$

where the second inequality is an applications of Jensen's inequality. This implies $P_1 \in \mathcal{C}(\mathcal{F}_1, L_{2,P_1})$. Moreover, since $P_1 \ll P$ and $\frac{dP_1}{dP} \leq K_1 < \infty$ for some $K_1 \in \mathbb{R}$,

$$\begin{aligned} L_{2,P_1}(f_1, f_2) &= \left(\int (f_1 - f_2)^2 \frac{dP_1}{dP} dP \right)^{1/2} \\ &\leq K_1^{1/2} \left(\int (f_1 - f_2)^2 dP \right)^{1/2} = K_1^{1/2} L_{2,P}(f_1, f_2) \end{aligned}$$

shows that $\mathcal{C}(\mathcal{F}_1, L_{2,P_1}) \subseteq \mathcal{C}(\mathcal{F}_1, L_{2,P})$ and so $P_1 \in \mathcal{C}(\mathcal{F}_1, L_{2,P})$. A similar argument shows $P_0 \in \mathcal{C}(\mathcal{F}_0, L_{2,P})$.

[5] See van der Vaart & Wellner (1997), footnote on p. 17.

[6] For $\varepsilon > 0$, let $(f_1, \ldots, f_K)$ be the centers of $L_{2,P}$-balls of radius $\varepsilon/\sqrt{2}$ that cover $\mathcal{F}_1$, and $(g_1, \ldots, g_M)$ be the center of $L_{2,P}$-balls of radius $\varepsilon/\sqrt{2}$ that cover $\mathcal{F}_0$. Then for any $(f, g) \in \mathcal{F}_1 \times \mathcal{F}_0$, there exists f_k and g_m such that $L_{2,P}(f, f_k) < \varepsilon/\sqrt{2}$ and $L_{2,P}(g, g_m) < \varepsilon/\sqrt{2}$, and so

$$L_2((f, g), (f_k, g_m) = \sqrt{L_{2,P}(f, f_k)^2 + L_{2,P}(g, g_m)^2} < \sqrt{(\varepsilon/\sqrt{2})^2 + (\varepsilon/\sqrt{2})^2} = \varepsilon$$

and thus the KM balls in $(\mathcal{F}_1 \times \mathcal{F}_0)$ of radius ε centered at (f_k, g_m) for some k, m cover $\mathcal{F}_1 \times \mathcal{F}_0$.

Use the inequalities above to see that

$$\begin{aligned}
|\mathcal{S}(P_1,P_0)(f_1,g_1)-\mathcal{S}(P_1,P_0)(f_2,g_2)| &= |P_1(f_1)-P_1(f_2)+P_0(g_1)-P_0(g_2)| \\
&\leq L_{2,P_1}(f_1,f_2)+L_{2,P_0}(g_1,g_2) \leq K_1^{1/2}L_{2,P}(f_1,f_2)+K_0^{1/2}L_{2,P}(\psi_1,\psi_2) \\
&\leq 2\max\{K_1^{1/2},K_0^{1/2}\}\max\{L_{2,P}(f_1,f_2),L_{2,P}(g_1,g_2)\} \\
&= 2\max\{K_1^{1/2},K_0^{1/2}\}\sqrt{\max\{L_{2,P}(f_1,f_2)^2,L_{2,P}(g_1,g_2)^2\}} \\
&\leq 2\max\{K_1^{1/2},K_0^{1/2}\}\sqrt{L_{2,P}(f_1,f_2)^2+L_{2,P}(g_1,g_2)^2} \\
&= 2\max\{K_1^{1/2},K_0^{1/2}\}L_2((f_1,g_1),(f_2,g_2))
\end{aligned}$$

hence $L_2((f_1,g_1),(f_2,g_2)) < \varepsilon/(2\max\{K_1^{1/2},K_0^{1/2}\})$ implies

$$|\mathcal{S}(P_1,P_0)(f_1,g_1)-\mathcal{S}(P_1,P_0)(f_2,g_2)| < \varepsilon$$

and therefore $\mathcal{S}(P_1,P_0) \in \mathcal{C}(\mathcal{F}_1 \times \mathcal{F}_0, L_2)$.

Lemma 1.9.55 shows that Ξ_c is Hadamard directionally differentiable at $\mathcal{S}(P_1,P_0)$ tangentially to $\mathcal{C}(\mathcal{F}_1 \times \mathcal{F}_0, L_2)$, with derivative

$$\Xi'_{c,\mathcal{S}(P_1,P_0)} : \mathcal{C}(\mathcal{F}_1 \times \mathcal{F}_0, L_2) \to \mathbb{R}, \qquad \Xi'_{c,\mathcal{S}(P_1,P_0)}(H) = \sup_{(\varphi,\psi)\in\Psi_c(P_1,P_0)} H(\varphi,\psi)$$

where $\Psi_c(P_1,P_0) = \arg\max_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} P_1(\varphi)+P_0(\psi)$ is nonempty, because $P_1+P_0 = \mathcal{S}(P_1,P_0)$ is continuous and $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ is compact.

4. Now consider the tangent spaces to ensure the composition of the derivatives is well defined. Observe that if $(H_1,H_0) \in \mathcal{C}(\mathcal{F}_1,L_{2,P})\times\mathcal{C}(\mathcal{F}_0,L_{2,P})$ then $\mathcal{S}(H_1,H_0) = H_1+H_0 \in \mathcal{C}(\mathcal{F}_1\times\mathcal{F}_0,L_2)$.[7] It follows from the chain rule (lemma 1.9.50) that OT_c is Hadamard directionally differentiable at (P_1,P_0) tangentially to $\mathcal{C}(\mathcal{F}_1,L_{2,P})\times\mathcal{C}(\mathcal{F}_0,L_{2,P})$ with

[7]Fix $(f,g) \in \mathcal{F}_1\times\mathcal{F}_0$ and let $\delta_1>0$ and $\delta_0>0$ be such that $L_{2,P_1}(f,\tilde{f})<\delta_1$ implies $H_1(f,\tilde{f})<\varepsilon/2$ and

derivative $OT_c : \mathcal{C}(\mathcal{F}_1, L_{2,P}) \times \mathcal{C}(\mathcal{F}_0, L_{2,P}) \to \mathbb{R}$ given by

$$OT'_{c,(P_1,P_0)}(H_1, H_0) = \Xi'_{c,\mathcal{S}(P_1,P_0)}(\mathcal{S}'_{(P_1,P_0)}(H_1, H_0)) = \sup_{(\varphi,\psi)\in\Psi_c(P_1,P_0)} H_1(\varphi) + H_0(\psi)$$

□

1.9.2.3 Full differentiability

The property distinguishing directional from full differentiability on a subspace is linearity of the derivative (Fang & Santos (2019), proposition 2.1). In the case of optimal transport, the derivative found in lemma 1.9.7 is linear on a large subspace of the tangent space when the solution to the dual problem is suitably unique. When it holds, this is sufficient for simpler bootstrap procedures to work for inference.

The dual solutions

$$(\varphi, \psi) \in \Psi_c(P_1, P_0) = \operatorname*{arg\,max}_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} P_1(\varphi) + P_0(\psi)$$

are referred to as *Kantorovich potentials*. Notice that for any $s \in \mathbb{R}$,

$$P_1(\varphi + s) + P_0(\psi - s) = P_1(\varphi) + P_0(\psi)$$

shows the most one can hope for is uniqueness up to a constant; if $(\varphi, \psi) \in \Psi_c(P_1, P_0)$, then $(\varphi + s, \psi - s) \in \Psi_c(P_1, P_0)$ as well.[8] It is well known in the optimal transport literature that

$L_{2,P_0}(g, \tilde{g}) < \delta_0$ implies $H_0(g, \tilde{g}) < \varepsilon/2$. The inequality

$$\begin{aligned} L_{2,P}(f, \tilde{f}) + L_{2,P}(g, \tilde{g}) &\leq 2\max\{L_{2,P}(f, \tilde{f}), L_{2,P}(g, \tilde{g})\} \\ &= 2\sqrt{\max\{L_{2,P}(f, \tilde{f})^2, L_{2,P}(g, \tilde{g})^2\}} = 2L_2((f, g), (\tilde{f}, \tilde{g})) \end{aligned}$$

implies that if $L_2((f, g), (\tilde{f}, \tilde{g})) < \min\{\delta_1, \delta_2\}/2$ then $|\mathcal{S}(H_1, H_0)(f, g) - \mathcal{S}(H_1, H_0)(\tilde{f}, \tilde{g})| \leq |H_1(f) - H_1(\tilde{f})| + |H_0(g) - H_0(\tilde{g})| < \varepsilon$.

[8] See Staudt et al. (2022) for extended discussion on uniqueness of Kantorovich potentials.

when the distributions P_1, P_0 have full support on a convex, compact subset of $\mathbb{R}$ and c is differentiable, the Kantorovich potential is indeed unique in this way on the supports of P_1 and P_0.

Lemma 1.9.8. *Suppose that*

1. *$c(y_1, y_0)$ is continuously differentiable.*
2. *P_d has compact support $\mathcal{Y}_d = [y_d^\ell, y_d^u] \subseteq \mathbb{R}$, and*

Let $\mathcal{F}_c$ and $\mathcal{F}_c^c$ be defined by (1.14) *and* (1.15) *respectively, and*

$$\Psi_c(P_1, P_0) = \underset{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)}{\arg\max} P_1(\varphi) + P_0(\psi)$$

Then for any $(\varphi_1, \psi_1), (\varphi_2, \psi_2) \in \Psi_c(P_1, P_0)$, there exists $s \in \mathbb{R}$ such that for all $(y_1, y_0) \in \mathcal{Y}_1 \times \mathcal{Y}_0$

$$\varphi_1(y_1) - \varphi_2(y_1) = s, \qquad \psi_1(y_0) - \psi_2(y_0) = -s$$

Proof. The proof is quite similar to that of Santambrogio (2015) proposition 7.18.

Let $(\varphi_1, \psi_1), (\varphi_2, \psi_2) \in \Psi_c(P_1, P_0)$. For $k = 1, 2$, φ_k and ψ_k (being elements of $\mathcal{F}_c$ and $\mathcal{F}_c^c$ respectively) are L-Lipschitz and hence absolutely continuous. This implies all four functions are differentiable Lebesgue-almost everywhere, and that for any $(y_1, y_0) \in \mathcal{Y}_1 \times \mathcal{Y}_0$,

$$\varphi_k(y_1) = \varphi_k(y_1^\ell) + \int_{y_1^\ell}^{y_1} \varphi_k'(y)dy \qquad \psi_k(y_0) = \psi_k(y_0^\ell) + \int_{y_0^\ell}^{y_0} \psi_k'(y)dy$$

Notice that the subset of $\mathcal{Y}_1$ where both φ_1 and φ_2 are differentiable also has full Lebesgue measure. It suffices to show that $\varphi_1'(y_1) = \varphi_2'(y_1)$ on this set (and $\psi_1'(y_0) = \psi_2'(y_0)$ on the subset of $\mathcal{Y}_0$ where both ψ_1 and ψ_2 are differentiable, which also has full Lebesgue measure),

from which it will follow that for any $(y_1, y_0) \in \mathcal{Y}_1 \times \mathcal{Y}_0$,

$$\varphi_1(y_1) - \varphi_2(y_1) = \varphi_1(y_1^\ell) - \varphi_2(y_1^\ell) + \int_{y_1^\ell}^{y_1} (\varphi_1'(y) - \varphi_2'(y))dy = \underbrace{\varphi_1(y_1^\ell) - \varphi_2(y_1^\ell)}_{:=s_\varphi}$$
$$\psi_1(y_0) - \psi_2(y_0) = \psi_1(y^\ell) - \psi_2(y^\ell) + \int_{y_0^\ell}^{y_0} (\psi_1'(y) - \psi_2'(y))dy = \underbrace{\psi_1(y_0^\ell) - \psi_2(y_0^\ell)}_{:=s_\psi}$$

Finally, observe that $P_1(\varphi_2) + P_0(\varphi_2) = P_1(\varphi_1) + P_0(\psi_1) = P_1(\varphi_2 + s_\varphi) + P_0(\psi_2 + s_\psi) = P_1(\varphi_2) + P_0(\psi_2) + s_\varphi + s_\psi$ implies $s_\varphi = -s_\psi$.

The remainder of the proof shows that for any $\bar{y}_1$ in the set where both φ_1 and φ_2 are differentiable, $\varphi_1'(\bar{y}_1) = \varphi_2'(\bar{y}_1)$. The same arguments work to show the corresponding claim regarding ψ_1 and ψ_2.

There exists $\pi \in \Pi(P_1, P_0)$ that solves the primal problem (see lemma 1.9.30). For any such π,

1. $\text{Supp}(P_1) = \{y_1 \in \mathcal{Y}_1 \ ; \ \exists y_0 \in \mathcal{Y}_0 \text{ s.t. } (y_1, y_0) \in \text{Supp}(\pi)\}$

   This follows because $\text{Pr}_1(\text{Supp}(\pi)) \coloneqq \{y_1 \in \mathcal{Y}_1 \ ; \ \exists y_0 \in \mathcal{Y}_0 \text{ s.t. } (y_1, y_0) \in \text{Supp}(\pi)\}$ is dense in $\text{Supp}(P_1)$, and $\text{Pr}_1(\text{Supp}(\pi))$ is closed because $\mathcal{Y}_0$ is compact.[9]

[9]Specifically, for any $A \subseteq \mathcal{Y}_1 \times \mathcal{Y}_0 \subseteq \mathbb{R}^2$, let $\text{Pr}_1(A) = \{y_1 \in \mathcal{Y}_1 \ ; \ \exists y_0 \in \mathcal{Y}_0 \text{ s.t. } (y_1, y_0) \in A\}$ be the cartesian projection of the set A onto the first coordinate. Let $P_1 \in \mathcal{P}(\mathcal{Y}_1)$, $P_0 \in \mathcal{P}(\mathcal{Y}_0)$, and $\pi \in \Pi(P_1, P_0)$. As noted in Staudt et al. (2022) (Remark 1), $\text{Pr}_1(\text{Supp}(\pi)) \subseteq \text{Supp}(P_1)$ with the possibility that inclusion is strict.

However, $\text{Pr}_1(\text{Supp}(\pi))$ is always dense in $\text{Supp}(P_1)$: let $y_1 \in \text{Supp}(P_1)$ and $\delta > 0$ be arbitrary, and suppose for contradiction that $B_\delta(y_1) \cap \text{Pr}_1(\text{Supp}(\pi)) = \varnothing$. Then $\big(B_\delta(y_1) \times \mathcal{Y}_0\big) \cap \text{Supp}(\pi) = \varnothing$ follows from the definition of $\text{Pr}_1(\text{Supp}(\pi))$, and thus

$$\begin{aligned} 0 &= \pi\left(\big(B_\delta(y_1) \times \mathcal{Y}_0\big) \cap \text{Supp}(\pi)\right) = \pi\left(\big(B_\delta(y_1) \times \mathcal{Y}_0\big)\right) + \pi\left(\text{Supp}(\pi)\right) - \pi\left(\big(B_\delta(y_1) \times \mathcal{Y}_0\big) \cup \text{Supp}(\pi)\right) \\ &= \pi\left(\big(B_\delta(y_1) \times \mathcal{Y}_0\big)\right) = P_1(B_\delta(y_1)) > 0 \end{aligned}$$

a contradiction showing $B_\delta(y_1) \cap \text{Pr}_1(\text{Supp}(\pi)) \neq \varnothing$. Thus $\text{Pr}_1(\text{Supp}(\pi))$ is dense in $\text{Supp}(P_1)$.

Moreover, if $\mathcal{Y}_0$ is compact then the map Pr_1 is closed: suppose $A \subseteq \mathcal{Y}_1 \times \mathcal{Y}_0 \subseteq \mathbb{R}^2$ is closed, and $\{y_{1n}\}_{n=1}^\infty \subseteq \text{Pr}_1(A)$ converges to y_1. Then there exists $\{y_{0n}\}_{n=1}^\infty \subseteq \mathcal{Y}_0$ such that $(y_{1n}, y_{0n}) \in A$ for each n. Since $\mathcal{Y}_0$ is compact, there exists a subsequence $\{y_{0n_k}\}_{k=1}^\infty$ and y_0 such that $\lim_{k\to\infty} y_{0n_k} = y_0$. Then notice that $\lim_{k\to\infty}(y_{1n_k}, y_{0n_k}) = (y_1, y_0)$. Since A is closed, $(y_1, y_0) \in A$.

$\text{Supp}(\pi)$ is closed by definition, hence $\text{Pr}_1(\text{Supp}(\pi))$ is closed and dense in $\text{Supp}(P_1)$, from which it follows that $\text{Supp}(\pi) = \text{Supp}(P_1)$.

2. For all $(y_1, y_0) \in \text{Supp}(\pi)$, $\varphi_k(y_1) + \psi_k(y_0) = c(y_1, y_0)$.

   It is easy to see that the equality holds π-almost surely. To see it holds specifically on the support, notice that optimality of π and (φ_k, ψ_k) implies that

$$\int c(y_1, y_0) d\pi(y_1, y_0) = \int \varphi_k(y_1) dP(y_1) + \int \psi_k(y_0) dP_0(y_0)$$

   and recall that $\varphi_k(y_1) + \psi_k(y_0) \leq c(y_1, y_0)$ holds for all $(y_1, y_0) \in \mathcal{Y} \times \mathcal{Y}$. If the inequality were strict for some $(y_1', y_0') \in \text{Supp}(\pi)$, then continuity of φ_k, ψ_k, and c would imply the inequality is sharp on a ball centered at (y_1, y_0) of some positive radius, denoted B, leading to the contradiction

$$\begin{aligned}
\int c(y_1, y_0) d\pi(y_1, y_0) &= \int_B c(y_1, y_0) d\pi(y_1, y_0) + \int_{B^c} c(y_1, y_0) d\pi(y_1, y_0) \\
&> \int_B \varphi_k(y_1) + \psi_k(y_0) d\pi(y_1, y_0) + \int_{B^c} \varphi_k(y_1) + \psi_k(y_0) d\pi(y_1, y_0) \\
&= \int \varphi_k(y_1) + \psi_k(y_0) d\pi(y_1, y_0) = \int \varphi_k(y_1) dP_1(y_1) + \int \psi_k(y_0) dP_0(y_0)
\end{aligned}$$

3. For any $\bar{y}_1 \in \text{Supp}(P_1)$, the above implies there there exists $\bar{y}_0 \in \mathcal{Y}_0$ such that $(\bar{y}_1, \bar{y}_0) \in \text{Supp}(\pi)$, and hence $\varphi_k(\bar{y}_1) + \psi_k(\bar{y}_0) = c(\bar{y}_1, \bar{y}_0)$. For any such $\bar{y}_0$,

$$y_1 \mapsto \varphi_k(y_1) - c(y_1, \bar{y}_0) \text{ is maximized at } \bar{y}_1 \tag{1.54}$$

   Indeed, if there were $y_1' \in \mathcal{Y}_1$ such that $\varphi_k(y_1') - c(y_1', \bar{y}_0) > \varphi_k(\bar{y}_1) - c(\bar{y}_1, \bar{y}_0)$, then by adding $\psi_k(\bar{y}_0)$ to both sides we find

$$\varphi_k(y_1') + \psi_k(\bar{y}_0) - c(y_1', \bar{y}_0) > \varphi_k(\bar{y}_1) + \psi_k(\bar{y}_0) - c(\bar{y}_1, \bar{y}_0) = 0$$

   This implies $\varphi_k(y_1') + \psi_k(\bar{y}_0) > c(y_1', \bar{y}_0)$, which contradicts $\varphi_k(y_1') + \psi_k(\bar{y}_0) \leq c(y_1', \bar{y}_0)$ for all $(y_1, y_0) \in \mathcal{Y}_1 \times \mathcal{Y}_0$.

4. Now observe that if $\bar{y}_1 \in (y_1^\ell, y_1^u)$ is a point at which φ_k is differentiable, then (1.54) implies $\varphi_k'(\bar{y}_1) = \frac{\partial c}{\partial y_1}(\bar{y}_1, \bar{y}_0)$.[10] Thus if $\bar{y}_1 \in (y_1^\ell, y_1^u)$ is a point at which both φ_1 and φ_2 are differentiable, then

$$\varphi_1(\bar{y}_1) = \frac{\partial c}{\partial y_1}(\bar{y}_1, \bar{y}_0) = \varphi_2(\bar{y}_1)$$

This completes the proof. □

To specify the subset of the tangent space on which $OT'_{c,(P_1,P_0)}$ is linear, let $\mathcal{Y}_d \subseteq \mathcal{Y}$ and $\mathbb{1}_{\mathcal{Y}_d}(y) = \mathbb{1}\{y \in \mathcal{Y}_d\}$. Let $\mathcal{G}$ denote a set of real-valued functions $g : \mathcal{Y} \to \mathbb{R}$ with the following property: if $g \in \mathcal{G}$, then $\mathbb{1}_{\mathcal{Y}_d} \times g \in \mathcal{G}$.[11] Let $\ell^\infty_{\mathcal{Y}_d}(\mathcal{G})$ be the set of bounded, linear functions $H : \mathcal{G} \to \mathbb{R}$ that evaluate constant functions to zero and "ignore" the value of functions outside of $\mathcal{Y}_d$. Specifically, define

$$\begin{aligned}\ell^\infty_{\mathcal{Y}_d}(\mathcal{G}) = \Big\{ H \in \ell^\infty(\mathcal{G}) \,;\ \ &\text{for all } a, b \in \mathbb{R} \text{ and } f, g \in \mathcal{G}, \\ &(i)\ H(f) = H(\mathbb{1}_{\mathcal{Y}_d} \times f),\ \ (ii)\ \text{ if } a \in \mathcal{G} \text{ then } H(a) = 0, \text{ and} \\ &(iii)\ \text{ if } af + bg \in \mathcal{G} \text{ then } H(af + bg) = aH(f) + bH(g) \Big\} \quad (1.55)\end{aligned}$$

Here we slightly abuse notation; $a \in \mathcal{G}$ refers to the function mapping each point in $\mathcal{Y}$ to the constant $a \in \mathbb{R}$. Equip $\ell^\infty_{\mathcal{Y}_d}(\mathcal{G})$ with the supremum norm, $\|H\|_{\mathcal{G}} = \|H\|_\infty = \sup_{g \in \mathcal{G}} |H(g)|$. As shown in appendix 1.9.3, first stage estimators of (P_1, P_0) based on the empirical distribution have weak limits concentrated on $\ell^\infty_{\mathcal{Y}_1}(\mathcal{F}_c) \times \ell^\infty_{\mathcal{Y}_0}(\mathcal{F}_c^c)$ where $\mathcal{Y}_d$ is the support of P_d.

Lemma 1.9.9. *$\ell^\infty_{\mathcal{Y}_d}(\mathcal{G})$ defined by (1.55) is closed.*

Proof. Let $\{H_n\}_{n=1}^\infty \subseteq \ell^\infty_{\mathcal{Y}_d}(\mathcal{G})$ be Cauchy, and let H be its limit in the Banach space $\ell^\infty(\mathcal{G})$. It suffices to show $H \in \ell^\infty_{\mathcal{Y}_d}(\mathcal{G})$.

Toward this end, first notice that $\|H_n - H\|_{\mathcal{G}} \to 0$ implies that for any $f \in \mathcal{G}$, $|H_n(f) - H(f)| \to 0$. Next observe that if the constant function $a \in \mathcal{G}$, then $0 = \lim_{n\to\infty} |H_n(a) -$

[10] Notice that the "choice" of π or $\bar{y}_0$ doesn't matter, because $\varphi_k'(\bar{y}_1)$ can take only one value.

[11] If we have a set $\tilde{\mathcal{G}}$ that does not satisfy this property, the set $\mathcal{G} = \tilde{\mathcal{G}} \cup \left\{ \mathbb{1}_{\mathcal{Y}_d} \times g \,;\ g \in \tilde{\mathcal{G}} \right\}$ will satisfy it.

$H(a)| = \lim_{n\to\infty}|H(a)| = |H(a)|$. For any function $f \in \mathcal{G}$, since $H_n(f) = H_n(\mathbb{1}_{\mathcal{Y}_d} \times f)$,

$$0 \le |H(f) - H(\mathbb{1}_{\mathcal{Y}_d} \times f)| \le |H(f) - H_n(f)| + |H(\mathbb{1}_{\mathcal{Y}_d} \times f) - H_n(\mathbb{1}_{\mathcal{Y}_d} \times f)| \to 0$$

and thus $H(\mathbb{1}_{\mathcal{Y}_d} \times f) = H(f)$. Finally, suppose $a, b \in \mathbb{R}$ and $f, g \in \mathcal{G}$ are such that $af + bg \in \mathcal{G}$. Similar to the argument above, since $H_n(af + bg) = aH_n(f) + bH_n(g)$,

$$\begin{aligned} 0 &\le |H(af + bg) - aH(f) - bH(g)| \\ &\le |H(af + bg) - H_n(af + bg)| + |aH_n(f) + bH_n(f) - aH(f) - bH_n(g)| \\ &\le |H(af + bg) - H_n(af + bg)| + |a||H_n(f) - H(f)| + |b||H_n(g) - H_n(g)| \to 0 \end{aligned}$$

and thus $H(af + bg) = aH(f) + bH(g)$.

This shows $H \in \ell^\infty_{\mathcal{Y}_d}(\mathcal{G})$, and completes the proof. □

Lemma 1.9.10 (Full differentiability of optimal transport)**.** *Let $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be lower semicontinuous, $\mathcal{F}_1, \mathcal{F}_0$ be sets of measurable functions mapping $\mathcal{Y}$ to $\mathbb{R}$, and $\mathcal{F}_c \subseteq \mathcal{F}_1$ and $\mathcal{F}_c^c \subseteq \mathcal{F}_0$ be universally bounded subsets. Suppose that*

1. *Strong duality holds:*

$$\inf_{\pi\in\Pi(P_1,P_0)} \int c(y_1, y_0)d\pi(y_1, y_0) = \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \int \varphi(y_1)dP_1(y_1) + \int \psi(y_0)dP_0(y_0),$$

2. *P dominates P_1 and P_0 with bounded densities,*

3. *$\mathcal{F}_d$ is P-Donsker and $\sup_{f\in\mathcal{F}_d}|P(f)| < \infty$ for each $d = 1, 0$, and*

4. *$(\mathcal{F}_1 \times \mathcal{F}_0, L_2)$ and the subset*

$$\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c) = \{(\varphi, \psi) \in \mathcal{F}_c \times \mathcal{F}_c^c \,;\, \varphi(y_1) + \psi(y_0) \le c(y_1, y_0)\}$$

are complete.

Let $\mathcal{Y}_1, \mathcal{Y}_0 \subseteq \mathcal{Y}$ *and* $\Psi_c(P_1, P_0) = \arg\max_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} P_1(\varphi) + P_0(\psi)$*, and further assume*

5. *For any* $(\varphi_1, \psi_1), (\varphi_2, \psi_2) \in \Psi_c(P_1, P_0)$*, there exists* $s \in \mathbb{R}$ *such that*

$$\mathbb{1}_{\mathcal{Y}_1} \times \varphi_1 = \mathbb{1}_{\mathcal{Y}_1} \times (\varphi_2 + s),\ P\text{-a.s.} \qquad \text{and} \qquad \mathbb{1}_{\mathcal{Y}_0} \times \psi_1 = \mathbb{1}_{\mathcal{Y}_0} \times (\psi_2 - s),\ P\text{-a.s.}$$

Then $OT_c : \ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0) \to \mathbb{R}$ *defined by*

$$OT_c(P_1, P_0) = \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} P_1(\varphi) + P_0(\psi)$$

is fully Hadamard differentiable at (P_1, P_0) *tangentially to*

$$\mathbb{D}_{Tan,Full} = \Big(\ell^\infty_{\mathcal{Y}_1}(\mathcal{F}_c) \times \ell^\infty_{\mathcal{Y}_0}(\mathcal{F}_c^c)\Big) \cap \Big(\mathcal{C}(\mathcal{F}_1, L_{2,P}) \times \mathcal{C}(\mathcal{F}_0, L_{2,P})\Big) \tag{1.56}$$

with derivative $OT'_{c,(P_1,P_0)} : \mathbb{D}_{Tan,Full} \to \mathbb{R}$ *given by*

$$OT'_{c,(P_1,P_0)}(H_1, H_0) = \sup_{(\varphi,\psi)\in\Psi_c(P_1,P_0)} H_1(\varphi) + H_0(\psi)$$

Proof. The first four assumptions allow application of lemma 1.9.7 to find that $OT_c : \ell^\infty(\mathcal{F}_1) \times \ell^\infty(\mathcal{F}_0) \to \mathbb{R}$ given by

$$OT_c(P_1, P_0) = \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} P_1(\varphi) + P_0(\psi)$$

is Hadamard directionally differentiable at (P_1, P_0) tangentially to $\mathbb{D}_{Tan} = \mathcal{C}(\mathcal{F}_1, L_{2,P}) \times \mathcal{C}(\mathcal{F}_0, L_{2,P})$. The set of maximizers $\Psi_c(P_1, P_0) = \arg\max_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} P_1(\varphi) + P_0(\psi)$ is nonempty, and the derivative $OT'_{c,(P_1,P_0)} : \mathbb{D}_{Tan} \to \mathbb{R}$ is given by

$$OT'_{c,(P_1,P_0)}(H_1, H_0) = \sup_{(\varphi,\psi)\in\Psi_c(P_1,P_0)} H_1(\varphi) + H_0(\psi)$$

Next observe that for any $(H_1, H_0) \in \mathbb{D}_{Tan,Full}$, $H_1 + H_0$ is flat on $\Psi_c(P_1, P_0)$. Specifically,

for any $(\varphi_1, \psi_1), (\varphi_2, \psi_2) \in \Psi_c(P_1, P_0)$, let s be such that

$$\mathbb{1}_{\mathcal{Y}_1} \times \varphi_1 = \mathbb{1}_{\mathcal{Y}_1} \times (\varphi_2 + s),\ P\text{-a.s.} \qquad \text{and} \qquad \mathbb{1}_{\mathcal{Y}_0} \times \psi_1 = \mathbb{1}_{\mathcal{Y}_0} \times (\psi_2 - s),\ P\text{-a.s.}$$

Then

$$\begin{aligned} H_1(\varphi_1) + H_0(\psi_1) &= H_1(\mathbb{1}_{\mathcal{Y}_1} \times \varphi_1) + H_0(\mathbb{1}_{\mathcal{Y}_0} \times \psi_1) \\ &= H_1(\mathbb{1}_{\mathcal{Y}_1} \times (\varphi_2 + s)) + H_0(\mathbb{1}_{\mathcal{Y}_0} \times (\psi_2 - s)) \\ &= H_1(\varphi_2 + s) + H_0(\psi_2 - s) \\ &= H_1(\varphi_2) + H_1(s) + H_0(\psi_2) - H_0(s) \\ &= H_1(\varphi_2) + H_0(\psi_2) \end{aligned}$$

where the first, third, fourth, and fifth equalities hold because $(H_1, H_0) \in \ell^\infty_{\mathcal{Y}_1}(\mathcal{F}_c) \times \ell^\infty_{\mathcal{Y}_0}(\mathcal{F}_c^c)$, and the second because $(H_1, H_0) \in \mathcal{C}(\mathcal{F}_1, L_{2,P}) \times \mathcal{C}(\mathcal{F}_0, L_{2,P})$.

Now use this "flatness" to observe the derivative is linear. Let $(H_1, H_0), (G_1, G_0) \in \mathbb{D}_{Tan,Full}$, $a, b \in \mathbb{R}$, and $(\tilde{\varphi}, \tilde{\psi}) \in \Psi_c(P_1, P_0)$, and notice that

$$\begin{aligned} OT'_{c,(P_1,P_0)}(a(H_1, H_0) + b(G_1, G_0)) &= \sup_{(\varphi,\psi)\in\Psi(P_1,P_0)} (aH_1 + bG_1)(\varphi) + (aH_0 + bG_0)(\psi) \\ &= aH_1(\tilde{\varphi}) + bG_1(\tilde{\varphi}) + aH_0(\tilde{\psi}) + bG_0(\tilde{\psi}) = a(H_1(\tilde{\varphi}) + H_0(\tilde{\psi})) + b(G_1(\tilde{\varphi}) + G_0(\tilde{\psi}) \\ &= a \times \sup_{(\varphi,\psi)\in\Psi(P_1,P_0)} \{H_1(\varphi) + H_0(\psi)\} + b \times \sup_{(\varphi,\psi)\in\Psi(P_1,P_0)} \{G_1(\varphi) + G_0(\psi)\} \\ &= aOT'_{c,(P_1,P_0)}(H_1, H_0) + bOT'_{c,(P_1,P_0)}(G_1, G_0) \end{aligned}$$

Since $OT'_{c,(P_1,P_0)}$ is linear on the subspace $\mathbb{D}_{Tan,Full}$, Fang & Santos (2019) proposition 2.1 implies OT_c is fully Hadamard differentiable at (P_1, P_0) tangentially to $\mathbb{D}_{Tan,Full}$. □

1.9.3 Appendix: weak convergence

Recall that

$$\begin{aligned}
\theta_x^L &= \theta^L(P_{1|x}, P_{0|x}), & \theta_x^H &= \theta^H(P_{1|x}, P_{0|x}) \\
\theta^L &= \sum_x s_x \theta_x^L, & \theta^L &= \sum_x s_x \theta_x^H \\
\gamma^L &= \inf_{t \in [\theta^L, \theta^H]} g(t, \eta) & \gamma^H &= \sup_{t \in [\theta^L, \theta^H]} g(t, \eta)
\end{aligned}$$

where $\eta = (\eta_1, \eta_0)$, with $\eta_d \in \mathbb{R}^{K_d}$ having coordinates

$$\eta_d^{(k)} = \sum_x P(X = x \mid D_1 > D_0) E[\eta_d^{(k)}(Y_d) \mid D_1 > D_0, X = x] = \sum_x s_x \eta_{d,x}^{(k)}$$

Here $\eta_{d,x}^{(k)} = P_{d|x}(\eta_d^{(k)})$, which are collected as $\eta_{d,x} = (\eta_{d,x}^{(1)}, \ldots, \eta_{d,x}^{(K_d)})$.

Define the following sets of functions:

$$\begin{aligned}
\tilde{\mathcal{F}}_1 &= \left\{ f : \mathcal{Y} \to \mathbb{R} \,;\; f = \varphi \text{ for some } \varphi \in \mathcal{F}_c, \text{ or } f = \eta_1^{(k)} \text{ for some } k = 1, \ldots, K_1 \right\} \quad (1.57) \\
\tilde{\mathcal{F}}_0 &= \left\{ f : \mathcal{Y} \to \mathbb{R} \,;\; f = \psi \text{ for some } \psi \in \mathcal{F}_c^c, \text{ or } f = \eta_0^{(k)} \text{ for some } k = 1, \ldots, K_0 \right\} \\
\mathcal{F}_{d,x} &= \left\{ f : \mathcal{Y} \to \mathbb{R} \,;\; f = g \text{ or } \mathbb{1}_{\mathcal{Y}_{d,x}} \times g \text{ for some } g \in \tilde{\mathcal{F}}_d \right\}
\end{aligned}$$

where $\mathcal{Y}_{d,x}$ is the support of $Y \mid D = d, X = x$, and $\mathbb{1}_{\mathcal{Y}_{d,x}}(y) = \mathbb{1}\{y \in \mathcal{Y}_{d,x}\}$. The additional functions of the form $f(y) = \mathbb{1}_{\mathcal{Y}_{d,x}}(y) g(y)$ are used to characterize the support of the weak limit of $\sqrt{n}(\hat{P}_{d|x} - P_{d|x})$ in $\ell^\infty(\mathcal{F}_{d,x})$. The maps $P_{d|x}$ can be written as

$$P_{d|x} : \mathcal{F}_{d,x} \to \mathbb{R}, \quad P_{d|x}(f) = \frac{P(\mathbb{1}_{d,x,d} \times f)/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d} \times f)/P(\mathbb{1}_{x,1-d})}{P(\mathbb{1}_{d,x,d})/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d})/P(\mathbb{1}_{x,1-d})} \quad (1.58)$$

and finally, define the set

$$\mathcal{F} = \bigcup_{d,x,z} \{\mathbb{1}_{d,x,z} \times f \; ; \; f \in \mathcal{F}_{d,x}\} \cup \{\mathbb{1}_{d,x,z}, \mathbb{1}_{x,z}, \mathbb{1}_x\}. \tag{1.59}$$

This appendix defines and studies the map $T : \mathbb{D}_C \subseteq \ell^\infty(\mathcal{F}) \to \mathbb{R}^2$ given by $(\gamma^L, \gamma^H) = T(P)$. The coming results show that $\mathcal{F}$ is P-Donsker, and the map T is Hadamard directionally differentiable at P. Together these imply, through the functional delta method, the weak convergence of $\sqrt{n}(T(\mathbb{P}_n) - T(P))$ (Fang & Santos (2019)).

Several operations in the definition of the map T are repeated for each $x \in \mathcal{X} = \{x_1, \ldots, x_M\}$, leading to large expressions. These are shortened with the notation $\{a_x\}_{x\in\mathcal{X}}$, which refers to $(a_{x_1}, \ldots, a_{x_M})$. For example,

$$\begin{aligned}&\left(\left\{P_{1|x}, P_{0|x}, \eta_{1,x}, \eta_{0,x}, s_x\right\}_{x\in\mathcal{X}}\right) \\ &\qquad = (P_{1|x_1}, P_{0|x_1}, \eta_{1,x_1}, \eta_{0,x_1}, s_{x_1}, \ldots, P_{1|x_M}, P_{0|x_M}, \eta_{1,x_M}, \eta_{0,x_M}, s_{x_M})\end{aligned}$$

is an element of $\prod_{m=1}^M \ell^\infty(\mathcal{F}_{1,x_m}) \times \ell^\infty(\mathcal{F}_{0,x_m}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$.

The function T is viewed as the composition of four functions: $T(P) = T_4(T_3(T_2(T_1(P))))$.

1. T_1 is the map to the conditional distributions and $\eta_{d,x}$:

$$T_1(P) = (\{P_{1|x}, P_{0|x}, \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}}),$$

2. T_2 involves optimal transport:

$$T_2(\{(P_{1|x}, P_{0|x}, \eta_{1,x}, \eta_{0,x}, s_x)\}_{x\in\mathcal{X}}) = (\{\theta_x^L, \theta_x^H, \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}}),$$

3. T_3 takes expectations over covariates: $T_3(\{(\theta_x^L, \theta_x^H, \eta_{1,x}, \eta_{0,x}, s_x)\}_{x\in\mathcal{X}}) = (\theta^L, \theta^H, \eta)$,

4. T_4 optimizes over $t \in [\theta^L, \theta^H]$: $T^4(\theta^L, \theta^H, \eta) = (\gamma^L, \gamma^H)$.

1.9.3.1 Verifying Donsker conditions

Before studying this map, this subsection shows the relevant sets are Donsker. The function classes $\mathcal{F}_c$ and $\mathcal{F}_c^c$ given by (1.14) and (1.15), or by (1.16) and (1.17), are well known Donsker classes as noted below. The results of van der Vaart & Wellner (1997) chapter 2.10 allow these to be extended to show $\mathcal{F}_{1,x}$ and $\mathcal{F}_{0,x}$ are Donsker. It follows quickly that $\mathcal{F}$ is Donsker.

Lemma 1.9.11. *Suppose that $\mathcal{Y} \subset \mathbb{R}$ is compact and $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is L-Lipschitz. Let $\mathcal{F}_c$, $\mathcal{F}_c^c$ be given by* (1.14) *and* (1.15) *respectively. Then $\mathcal{F}_c$ and $\mathcal{F}_c^c$ are universally Donsker.*

Proof. Note that any distribution defined on the compact $\mathcal{Y}$ has a finite $2+\delta$ moment. The result follows from the bracketing number bound given by van der Vaart & Wellner (1997) corollary 2.7.4. □

Lemma 1.9.12. *$\mathcal{F}_c$ and $\mathcal{F}_c^c$ given by* (1.16) *and* (1.17) *are universally Donsker.*

Proof. The intervals (convex subsets of $\mathbb{R}$) form a well-known VC class with VC-dimension at most 3. Consider an arbitrary set of three real numbers $\{y_1, y_2, y_3\}$ with $y_1 < y_2 < y_3$, and notice that no interval can pick out the set $\{y_1, y_3\}$; that is, there does not exist an interval I with $\{y_1, y_3\} = \{y_1, y_2, y_3\} \cap I$. Since the intervals cannot shatter finite sets of size 3, the VC-dimension of the intervals is at most 3.

Similarly, the complements of intervals form a VC class of VC-dimension at most 4. Consider $\{y_1, y_2, y_3, y_4\}$ with $y_1 < y_2 < y_3 < y_4$ and notice that no complement of an interval can pick out $\{y_1, y_3\}$. Since the complements of intervals cannot shatter finite sets of size 4, the VC-dimension of the complements of intervals is at most 4.

The claim follows, because any (suitably measurable) VC class is Donsker for any probability measure (van der Vaart & Wellner (1997) section 2.6.1). □

Lemma 1.9.13. *Let $\mathcal{G}$ be P-Donsker and $\mathbb{1}_A$ be the indicator function for the set A. Then the set $\{\mathbb{1}_A \times g \ ; \ g \in \mathcal{G}\}$ is P-Donsker.*

Proof. The proof is an application of van der Vaart & Wellner (1997) theorem 2.10.6. Specifically, let $\phi : \mathcal{G} \times \{\mathbb{1}_A\} \to \mathbb{R}$ be the map $\phi(g, \mathbb{1}_a) = \mathbb{1}_A \times g$. Notice that for any $f, g \in \mathcal{G}_1 \times \{\mathbb{1}_A\}$,

$$\begin{aligned} |\phi \circ f(w) - \phi \circ g(w)|^2 &= |\mathbb{1}_A(w) \times f_1(w) - \mathbb{1}_A(w) \times g_1(w)|^2 \\ &= \mathbb{1}_A(w) \times |f_1(w) - g_2(w)|^2 \\ &\leq |f_1(w) - g_1(w)|^2 = \sum_{\ell=1}^{k} (f_\ell(w) - g_\ell(w))^2 \end{aligned}$$

and thus van der Vaart & Wellner (1997) condition (2.10.5) holds. Moreover, notice that for any $g \in \mathcal{G}$, $(\mathbb{1}_A \times g)^2 \leq g^2$ and P-square integrability of $g \in \mathcal{G}$ implies $\mathbb{1}_A \times g$ is P-square integrable. Thus van der Vaart & Wellner (1997) theorem 2.10.6 implies $\{\mathbb{1}_A \times g \; ; \; g \in \mathcal{G}\}$ is P-Donsker. □

Lemma 1.9.14 ($\mathcal{F}_{d,x}$ are P-Donsker)**.** *Suppose assumptions 1, 2, and 3 hold. Let $\mathcal{F}_c$ and $\mathcal{F}_c^c$ be given by* (1.14) *and* (1.15)*, or by* (1.16) *and* (1.17)*. Let $\mathcal{F}_{d,x}$ be as defined in* (1.57)*. Then $\mathcal{F}_{d,x}$ is P-Donsker and $\sup_{f \in \mathcal{F}_{d,x}} |P(f)| < \infty$.*

Proof. 1. We first show $\tilde{\mathcal{F}}_d$ is P-Donsker and $\sup_{g \in \tilde{F}_d} |P(f)| < \infty$. The argument shows the argument for $\tilde{\mathcal{F}}_1$, as the same argument works when applied to $\tilde{\mathcal{F}}_0$.

Begin by noticing that

$$\begin{aligned} \tilde{\mathcal{F}}_1 &= \left\{ f : \mathcal{Y} \to \mathbb{R} \; ; \; f = \varphi \text{ for some } \varphi \in \mathcal{F}_c, \text{ or } f = \eta_1^{(k)} \text{ for some } k = 1, \ldots, K_1 \right\} \\ &= \mathcal{F}_c \cup \left\{ \eta_1^{(1)}, \ldots, \eta_1^{(K_1)} \right\} \end{aligned}$$

Since $\left\{ \eta_1^{(1)}, \ldots, \eta_1^{(K_1)} \right\}$ is a finite number of functions which, by assumption 3 (i), have finite second P-moment: $P((\eta_1^{(k)})^2) < \infty$. Thus $\left\{ \eta_1^{(1)}, \ldots, \eta_1^{(K_1)} \right\}$ is Donsker. $\mathcal{F}_c$ is Donsker by lemma 1.9.11 or 1.9.12, and so $\tilde{\mathcal{F}}_1 = \mathcal{F}_c \cup \left\{ \eta_1^{(1)}, \ldots, \eta_1^{(K_1)} \right\}$ is the union of

two P-Donsker sets. Since

$$\|P\|_{\tilde{\mathcal{F}}_1} = \max\{\sup_{\varphi\in\mathcal{F}_c}|P(\varphi)|, |P(\eta_1^{(1)})|, \ldots, |P(\eta_1^{(K_1)})|\} < \infty$$

van der Vaart & Wellner (1997) example 2.10.7 shows $\tilde{\mathcal{F}}_1$ is P-Donsker. Note we have also shown that $\sup_{g\in\tilde{F}_1}|P(f)| < \infty$.

2. Now notice that

$$\begin{aligned}\mathcal{F}_{d,x} &= \left\{f : \mathcal{Y} \to \mathbb{R}\ ;\ f = g \text{ or } \mathbb{1}_{\mathcal{Y}_{d,x}} \times g \text{ for some } g \in \tilde{\mathcal{F}}_d\right\} \\ &= \tilde{\mathcal{F}}_d \cup \left\{\mathbb{1}_{\mathcal{Y}_{d,x}} \times g\ ;\ g \in \tilde{\mathcal{F}}_d\right\}\end{aligned}$$

Lemma 1.9.13 shows $\left\{\mathbb{1}_{\mathcal{Y}_{d,x}} \times g\ ;\ g \in \tilde{\mathcal{F}}_d\right\}$ is P-Donsker. Moreover, since $\mathcal{F}_c$ is uniformly bounded,

$$\begin{aligned}&\|P\|_{\left\{\mathbb{1}_{\mathcal{Y}_{d,x}}\times g\ ;\ g\in\tilde{\mathcal{F}}_d\right\}} \\ &\quad= \max\left\{\sup_{\varphi\in\mathcal{F}_c}|P(\mathbb{1}_{\mathcal{Y}_{d,x}} \times \varphi)|, |P(\mathbb{1}_{\mathcal{Y}_{d,x}} \times \eta_1^{(1)})|, \ldots, |P(\mathbb{1}_{\mathcal{Y}_{d,x}} \times \eta_1^{(K_1)})|\right\} < \infty\end{aligned}$$

It follows that

$$\|P\|_{\mathcal{F}_{d,x}} = \sup_{f\in\mathcal{F}_{d,x}}|P(f)| = \max\left\{\sup_{f\in\tilde{\mathcal{F}}_d}|P(f)|, \sup_{f\in\{\mathbb{1}_{\mathcal{Y}_{d,x}}\times g\ ;\ g\in\tilde{\mathcal{F}}_d\}}|P(f)|\right\} < \infty$$

Thus van der Vaart & Wellner (1997) example 2.10.7 implies $\mathcal{F}_1$ is P-Donsker.

□

Lemma 1.9.15 ($\mathcal{F}$ is P-Donsker)**.** *Suppose assumptions 1, 2 and 3 hold. Then $\mathcal{F}$ is P-*

Donsker, implying

$$\sqrt{n}(\mathbb{P}_n - P) \xrightarrow{L} \mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{F}),$$

where $\mathbb{G}$ *is a tight, mean-zero Gaussian process with* $P(\mathbb{G} \in \mathcal{C}(\mathcal{F}, L_{2,P}) = 1$.

Proof. Lemma 1.9.13 shows $\{\mathbb{1}_{d,x,z} \times f \ ; \ f \in \mathcal{F}_{d,x}\}$ is P-Donsker. Moreover, $\mathcal{F}_{d,x}$ is the union of a subset of universally bounded functions (in either $\mathcal{F}_c$ or $\mathcal{F}_c^c$) and a finite subset of square integrable functions. It follows that

$$\|P\|_{\left\{\mathbb{1}_{d,x,z} \times g \ ; \ g \in \mathcal{F}_{d,x}\right\}} = \sup_{f \in \left\{\mathbb{1}_{d,x,z} \times g \ ; \ g \in \mathcal{F}_{d,x}\right\}} |P(f)| < \infty$$

Next notice that

$$\mathcal{F} = \bigcup_{d,x,z} \{\mathbb{1}_{d,x,z} \times f \ ; \ f \in \mathcal{F}_{d,x}\} \cup \{\mathbb{1}_{d,x,z}, \mathbb{1}_{x,z}, \mathbb{1}_x\}$$

is the union of a finite number of P-Donsker sets, with

$$\|P\|_{\mathcal{F}} = \max_{d,x,z} \left\{ \max \left\{ \sup_{f \in \left\{\mathbb{1}_{d,x,z} \times g \ ; \ g \in \mathcal{F}_{d,x}\right\}} |P(f)|, |P(\mathbb{1}_{d,x,z})|, |P(\mathbb{1}_{x,z})|, |P(\mathbb{1}_x)|, \right\} \right\} < \infty$$

It follows from van der Vaart & Wellner (1997) example 2.10.7 that $\mathcal{F}$ is P-Donsker, which implies $\sqrt{n}(\mathbb{P}_n - P) \xrightarrow{L} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$, where $\mathbb{G}$ is a tight, mean-zero Gaussian process. Moreover, van der Vaart & Wellner (1997) section 2.1.2 and problem 2.1.2 imply that $P(\mathbb{G} \in \mathcal{C}(\mathcal{F}, L_{2,P}) = 1$. □

1.9.3.2 Conditional Distributions, $T_1(P) = (\{P_{1|x}, P_{0|x}, \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}})$

Lemma 1.2.1 shows that the distributions of $Y_d \mid D_1 > D_0, X = x$, denoted $P_{d|x}$, are identified by

$$\begin{aligned} P_{d|x}(f) &= E_{P_{d|x}}[f(Y_d)] = E[f(Y_d) \mid D_1 > D_0, X = x] \\ &= \frac{E[f(Y)\mathbb{1}\{D = d\} \mid Z = d, X = x] - E[f(Y)\mathbb{1}\{D = d\} \mid Z = 1 - d, X = x]}{P(D = d \mid Z = d, X = x) - P(D = d \mid Z = 1 - d, X = x)} \end{aligned}$$

and the distribution of X conditional on $D_1 > D_0$ is identified by

$$\begin{aligned} s_x &= P(X = x \mid D_1 > D_0) \\ &= \frac{[P(D = 1 \mid Z = 1, X = x) - P(D = 1 \mid Z = 0, X = x)]\, P(X = x)}{\sum_{x'} [P(D = 1 \mid Z = 1, X = x') - P(D = 1 \mid Z = 0, X = x')]\, P(X = x')} \end{aligned}$$

Recall the notation shortening indicators

$$\mathbb{1}_{d,x,z}(D, X, Z) = \mathbb{1}\{D = d, X = x, Z = z\},$$
$$\mathbb{1}_{x,z}(X, Z) = \mathbb{1}\{X = x, Z = z\}, \qquad \mathbb{1}_x(X) = \mathbb{1}\{X = x\}$$

and notice that $P_{d|x} : \ell^\infty(\mathcal{F}_d) \to \mathbb{R}$ and $s_x \in \mathbb{R}$, given by

$$\begin{aligned} P_{d|x}(f) &= \frac{P(\mathbb{1}_{d,x,d} \times f)/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d} \times f)/P(\mathbb{1}_{x,0})}{P(\mathbb{1}_{d,x,d})/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d})/P(\mathbb{1}_{x,1-d})}, \\ s_x &= \frac{[P(\mathbb{1}_{1,x,1})/P(\mathbb{1}_{x,1}) - P(\mathbb{1}_{1,x,0})/P(\mathbb{1}_{x,0})]P(\mathbb{1}_x)}{\sum_{x'}[P(\mathbb{1}_{1,x',1})/P(\mathbb{1}_{x',1}) - P(\mathbb{1}_{1,x',0})/P(\mathbb{1}_{x',0})]P(\mathbb{1}_{x'})}, \end{aligned}$$

are functions of $P \in \ell^\infty(\mathcal{F})$. Moreover, $\eta_{d,x}^{(k)} = E[\eta_d^{(k)}(Y_d) \mid D_1 > D_0, X = x] = P_{d|x}(\eta_d^{(k)})$ and $\eta_{d,x} = (\eta_{d,x}^{(1)}, \ldots, \eta_{d,x}^{(K_1)})$ is simply an evaluation of $P_{d|x}$ at the points $\eta_d^{(k)} \in \mathcal{F}_{d,x}$.

This map is given by

$$
\begin{aligned}
T_1 : \mathbb{D}_C \subseteq \ell^\infty(\mathcal{F}) &\to \prod_{m=1}^{M} \ell^\infty(\mathcal{F}_{1,x_m}) \times \ell^\infty(\mathcal{F}_{0,x_m}) \times \mathbb{R}^{(K_1)} \times \mathbb{R}^{(K_0)} \times \mathbb{R} \\
T_1(P) &= \left(\left\{ P_{1|x}, P_{0|x}, \eta_{1,x}, \eta_{0,x}, s_x \right\}_{x \in \mathcal{X}} \right) \\
&= (P_{1|x_1}, P_{0|x_1}, \eta_{1,x_1}, \eta_{0,x_1}, s_{x_1}, \ldots, P_{1|x_M}, P_{0|x_M}, \eta_{1,x_M}, \eta_{0,x_M}, s_{x_M})
\end{aligned}
$$

where the domain, $\mathbb{D}_C \subseteq \ell^\infty(\mathcal{F})$, ensures the map never divide by zero:

$$
\begin{aligned}
\mathbb{D}_C = \big\{ G \in \ell^\infty(\mathcal{F}) \,;\ \ &\text{for all } (d,x,z),\ G(\mathbb{1}_x) > 0,\ G(\mathbb{1}_{x,z}) > 0, \text{ and} \\
&G(\mathbb{1}_{d,x,d})/G(\mathbb{1}_{x,d}) - G(\mathbb{1}_{d,x,1-d})/G(\mathbb{1}_{x,1-d}) > 0 \big\}
\end{aligned} \tag{1.60}
$$

Note that assumption 1 implies $P \in \mathbb{D}_C$, a claim shown in the proof of lemma 1.9.17 below.

Lemma 1.9.51 shows that Hadamard differentiable functions with the same domain can be "stacked". Moreover, the coordinates corresponding to the η terms are evaluations of $P_{d|x}$ at specific coordinates; since evaluation is linear and continuous, the map defining these terms is fully Hadamard differentiable if the other maps are fully Hadamard differentiable. Thus it suffices to ensure the maps $C_{d,x} : \mathbb{D}_C \to \mathbb{R}$ and $C_{s,x} : \mathbb{D}_C \to \mathbb{R}$ given by $C_{d,x}(P) = P_{d|x}$ and $C_{s,x}(P) = s_x$ are fully Hadamard differentiable at P tangentially to $\ell^\infty(\mathcal{F})$.

Lemma 1.9.16 (Maps to conditional distributions are fully Hadamard differentiable)**.** *Let $\mathcal{F}$ be defined by* (1.59)*, and $\mathbb{D}_C$ be defined by* (1.60)*. Define the functions $C_{1,x}$, $C_{0,x}$, and $C_{s,x}$ with*

$$
\begin{aligned}
&C_{d,x} : \mathbb{D}_C \to \ell^\infty(\mathcal{F}_{d,x}), \quad C_{d,x}(G)(f) = \frac{G(\mathbb{1}_{d,x,d} \times f)/G(\mathbb{1}_{x,d}) - G(\mathbb{1}_{d,x,1-d} \times f)/G(\mathbb{1}_{x,1-d})}{G(\mathbb{1}_{d,x,d})/G(\mathbb{1}_{x,d}) - G(\mathbb{1}_{d,x,1-d})/G(\mathbb{1}_{x,1-d})}, \\
&C_{s,x} : \mathbb{D}_C \to \mathbb{R}, \qquad\qquad C_{s,x}(G) = \frac{[G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})]G(\mathbb{1}_x)}{\sum_{x'}[G(\mathbb{1}_{1,x',1})/G(\mathbb{1}_{x',1}) - G(\mathbb{1}_{1,x',0})/G(\mathbb{1}_{x',0})]G(\mathbb{1}_{x'})}
\end{aligned}
$$

All three functions are fully Hadamard differentiable at any $G \in \mathbb{D}_C$ tangentially to $\ell^\infty(\mathcal{F})$,

with derivatives $C'_{d,x,G} : \ell^\infty(\mathcal{F}) \to \ell^\infty(\mathcal{F}_{d,x})$ *and* $C'_{s,x,G} : \ell^\infty(\mathcal{F}) \to \mathbb{R}$ *described in the proof.*

Proof. In steps:

1. We first show differentiability of $C_{1,x}$. The argument applies the chain rule. An inner function "rearranges" elements of $\mathbb{D}_C \subseteq \ell^\infty(\mathcal{F})$, which can be viewed as a fully Hadamard differentiable mapping (see lemma 1.9.52). An outer function maps that rearrangment to $\ell^\infty(\mathcal{F}_1)$, and is shown fully Hadamard differentiable at $G \in \mathbb{D}_C$ by applying corollary 1.9.54.

   In detailed steps:

   (a) Define $\mathbb{D}_q = \{(n_1, p_{11}, p_1, n_0, p_{10}, p_0) \in \mathbb{R}^6 \; ; \; p_1 > 0, \; p_0 > 0, \; p_{11}/p_1 - p_{10}/p_0 > 0\}$ and

   $$q : \mathbb{D}_q \to \mathbb{R}, \qquad q(n_1, p_{11}, p_1, n_0, p_{10}, p_0) = \frac{n_1/p_1 - n_0/p_0}{p_{11}/p_1 - p_{10}/p_0}$$

   Recall the following notation from corollary 1.9.54:

   $$\ell^\infty(\mathcal{F}_1, \mathbb{D}_q) = \left\{ r : \mathcal{F}_1 \to \mathbb{R}^6 \; ; \; r(\varphi) \in \mathbb{D}_q, \; \sup_{\varphi \in \mathcal{F}_1} \|r(f)\| < \infty \right\} \subseteq \ell^\infty(\mathcal{F}_1)^6$$
   $$\ell^\infty_q(\mathcal{F}_1, \mathbb{D}_q) = \left\{ r \in \ell^\infty(\mathcal{F}_1, \mathbb{D}_q) \; ; \; \sup_{f \in \mathcal{F}_1} |q(r(f))| < \infty \right\}$$

   For elements $r \in \ell^\infty(\mathcal{F}_1, \mathbb{D}_q)$, the composition $q(r(\varphi))$ is well defined for any $\varphi \in \mathcal{F}_1$. For elements $r \in \ell^\infty_q(\mathcal{F}_1, \mathbb{D}_q)$, composition defines a bounded map; that is, $\varphi \mapsto q(r(\varphi))$ defines an element of $\ell^\infty(\mathcal{F}_1)$. Finally, define

   $$Q : \ell^\infty_q(\mathcal{F}_1, \mathbb{D}_q) \to \ell^\infty(\mathcal{F}_1), \qquad Q(r)(\varphi) = q(r(\varphi))$$

   (b) For the rearrangement, define $\tilde{\mathcal{F}}_{1,x,1} \equiv \{\mathbb{1}_{1,x,1} \times f \; ; \; f \in \mathcal{F}_1\}$,

$\tilde{\mathcal{F}}_{1,x,0} \equiv \{\mathbb{1}_{1,x,0} \times f \,;\, f \in \mathcal{F}_1\}$, and

$$\begin{aligned}\tilde{R}_{1,x} : \mathbb{D}_C \to \ell^\infty(\tilde{\mathcal{F}}_{1,x,1}) \times \ell^\infty(\{\mathbb{1}_{1,x,1}\}) \times \ell^\infty(\{\mathbb{1}_{x,1}\}) \\ \times \ell^\infty(\tilde{\mathcal{F}}_{1,x,0}) \times \ell^\infty(\{\mathbb{1}_{1,x,0}\}) \times \ell^\infty(\{\mathbb{1}_{x,0}\})\end{aligned}$$

$$\begin{aligned}&\tilde{R}_{1,x}(G)(\mathbb{1}_{1,x,1} \times f, \mathbb{1}_{1,x,1}, \mathbb{1}_{x,1}, \mathbb{1}_{1,x,0} \times f, \mathbb{1}_{1,x,0}, \mathbb{1}_{x,0}) \\ &\quad = (G(\mathbb{1}_{1,x,1} \times f), G(\mathbb{1}_{1,x,1}), G(\mathbb{1}_{x,1}), G(\mathbb{1}_{1,x,0} \times f), G(\mathbb{1}_{1,x,0}), G(\mathbb{1}_{x,0}))\end{aligned}$$

Lemma 1.9.52 shows that $\tilde{R}_{1,x}$ is fully Hadamard differentiable tangentially to $\ell^\infty(\mathcal{F})$ and is its own derivative; i.e. $\tilde{R}'_{1,x,g} = \tilde{R}_{1,x}$. Now view $\tilde{R}_{1,x}$ as a map from $\mathbb{D}_C \subseteq \ell^\infty(\mathcal{F})$ to $\ell_q^\infty(\mathcal{F}_1, \mathbb{D}_q)$, i.e. define $R_{1,x} : \mathbb{D}_C \to \ell_q^\infty(\mathcal{F}_1, \mathbb{D}_q)$ pointwise with

$$\begin{aligned}R_{1,x}(G)(f) &= \tilde{R}_{1,x}(G)(\mathbb{1}_{1,x,1} \times f, \mathbb{1}_{1,x,1}, \mathbb{1}_{x,1}, \mathbb{1}_{1,x,0} \times g, \mathbb{1}_{1,x,0}, \mathbb{1}_{x,0}) \\ &= (G(\mathbb{1}_{1,x,1} \times f), G(\mathbb{1}_{1,x,1}), G(\mathbb{1}_{x,1}), G(\mathbb{1}_{1,x,0} \times f), G(\mathbb{1}_{1,x,0}), G(\mathbb{1}_{x,0}))\end{aligned}$$

Note that $G \in \mathbb{D}_C$ implies

$$\sup_{f \in \mathcal{F}_1} |q(R_{1,x}(G)(f))| = \sup_{f \in \mathcal{F}_1} \left| \frac{G(\mathbb{1}_{1,x,1} \times f)/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0} \times f)/G(\mathbb{1}_{x,0})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})} \right| < \infty$$

and thus $R_{1,x}(G) \in \ell_q^\infty(\mathcal{F}_1, \mathbb{D}_q)$.

(c) To apply corollary 1.9.54, observe that $q(n_1, p_{11}, p_1, n_0, p_{10}, p_0) = \frac{n_1/p_1 - n_0/p_0}{p_{11}/p_1 - p_{10}/p_0}$ is

continuously differentiable on $\mathbb{D}_q$ with gradient $\nabla q : \mathbb{D}_q \to \mathbb{R}^6$ given by

$$\begin{aligned}
&\nabla q(n_1, p_{11}, p_1, n_0, p_{10}, p_0) = \left(\tfrac{\partial q}{\partial n_1},\ \tfrac{\partial q}{\partial p_{11}},\ \tfrac{\partial q}{\partial p_1},\ \tfrac{\partial q}{\partial n_0},\ \tfrac{\partial q}{\partial p_{10}},\ \tfrac{\partial q}{\partial p_0}\right)^{\intercal},\\
&\frac{\partial q}{\partial n_1} = \frac{1/p_1}{p_{11}/p_1 - p_{10}/p_0}\\
&\frac{\partial q}{\partial p_{11}} = -\frac{n_1/p_1 - n_0/p_0}{(p_{11}/p_1 - p_{10}/p_0)^2}\frac{1}{p_1} = \left[\frac{1/p_1}{p_{11}/p_1 - p_{10}/p_0}\right](-q)\\
&\frac{\partial q}{\partial p_1} = \frac{(p_{11}/p_1 - p_{10}/p_0)(-n_1/p_1^2) - (n_1/p_1 - n_0/p_0)(-p_{11}/p_1^2)}{(p_{11}/p_1 - p_{10}/p_0)^2}\\
&\qquad = \frac{-n_1/p_1^2}{p_{11}/p_1 - p_{10}/p_0} + \frac{q(p_{11}/p_1^2)}{p_{11}/p_1 - p_{10}/p_0} = \left[\frac{1/p_1}{p_{11}/p_1 - p_{10}/p_0}\right]\frac{qp_{11} - n_1}{p_1}\\
&\frac{\partial q}{\partial n_0} = \frac{-1/p_0}{p_{11}/p_1 - p_{10}/p_0}\\
&\frac{\partial q}{\partial p_{10}} = -\frac{n_1/p_1 - n_0/p_0}{(p_{11}/p_1 - p_{10}/p_0)^2}\left(-\frac{1}{p_0}\right) = \left[\frac{-1/p_0}{p_{11}/p_1 - p_{10}/p_0}\right](-q)\\
&\frac{\partial q}{\partial p_0} = \frac{(p_{11}/p_1 - p_{10}/p_0)(n_0/p_0^2) - (n_1/p_1 - n_0/p_0)(p_{10}/p_0^2)}{(p_{11}/p_1 - p_{10}/p_0)^2}\\
&\qquad = \frac{n_0/p_0^2}{p_{11}/p_1 - p_{10}/p_0} - \frac{q(p_{10}/p_0^2)}{p_{11}/p_1 - p_{10}/p_0} = \left[\frac{-1/p_0}{p_{11}/p_1 - p_{10}/p_0}\right]\frac{qp_{10} - n_0}{p_0}
\end{aligned}$$

Furthermore, there exists $\delta > 0$ such that

$$R_{1,x}(G)(\mathcal{F}_1) = \left\{ r \in \mathbb{R}^6 \ ;\ \inf_{f \in \mathcal{F}_1} \|r - R_{1,x}(G)(\varphi)\| \leq \delta \right\} \subseteq \mathbb{D}_q$$

and so lemma 1.9.54 implies Q is fully Hadamard differentiable at $R_{1,x}(G)$ tangentially to $\ell^\infty(\mathcal{F}_1)^6$ with derivative $Q'_{R_{1,x}(G)} : \ell^\infty(\mathcal{F}_1)^6 \to \ell^\infty(\mathcal{F}_1)$ given pointwise by

$$Q'_{R_{1,x}(G)}(J)(f) = \left[\nabla q(R_{1,x}(G)(\varphi))\right]^{\intercal} J(f)$$

(d) Finally, observe that $C_{1,x}(G) = Q(R_{1,x}(G))$ and apply the chain rule (lemma 1.9.50) to find that $C_{1,x}$ is fully Hadamard differentiable at G tangentially to

$\ell^\infty(\mathcal{F})$ with derivative

$$C'_{1,x,G} : \ell^\infty(\mathcal{F}) \to \ell^\infty(\mathcal{F}_{1,x}), \qquad C'_{1,x,G}(H) = Q'_{R_{1,x}(G)}(R_{1,x}(H))$$

Writing out an evaluation clarifies the notation of the derivative:

$$\begin{aligned}
C'_{1,x,G}(H)(f) &= Q'_{R_{1,x}(G)}(R_{1,x}(H))(f) = [\nabla q(R_{1,x}(G)(f))]^\intercal R_{1,x}(H)(f) \qquad (1.61)\\
&= \left[\frac{1/G(\mathbb{1}_{x,1})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}\right] H(\mathbb{1}_{1,x,1} \times f)\\
&\quad + \left[\frac{1/G(\mathbb{1}_{x,1})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}\right] (-C_{1,x}(G)(f))H(\mathbb{1}_{1,x,1})\\
&\quad + \left[\frac{1/G(\mathbb{1}_{x,1})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}\right]\\
&\qquad\qquad \times \frac{C_{1,x}(G)(f) \times G(\mathbb{1}_{1,x,1}) - G(\mathbb{1}_{1,x,1} \times f)}{G(\mathbb{1}_{x,1})} H(\mathbb{1}_{x,1})\\
&\quad + \left[\frac{-1/G(\mathbb{1}_{x,0})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}\right] H(\mathbb{1}_{1,x,0} \times f)\\
&\quad + \left[\frac{-1/G(\mathbb{1}_{x,0})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}\right] (-C_{1,x}(G)(f))H(\mathbb{1}_{1,x,0})\\
&\quad + \left[\frac{-1/G(\mathbb{1}_{x,0})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}\right]\\
&\qquad\qquad \times \frac{C_{1,x}(G)(f) \times G(\mathbb{1}_{1,x,0}) - G(\mathbb{1}_{1,x,0} \times f)}{G(\mathbb{1}_{x,0})} H(\mathbb{1}_{x,0})
\end{aligned}$$

2. The same arguments imply the claim regarding $C_{0,x}$.

   Specifically, notice that $C_{0,x}$ is the same outer transformation applied to a different rearrangement: let

$$\begin{aligned}
R_{1,x}(G)(\varphi) &= (G(\mathbb{1}_{1,x,1} \times \varphi), G(\mathbb{1}_{1,x,1}), G(\mathbb{1}_{x,1}), G(\mathbb{1}_{1,x,0} \times \varphi), G(\mathbb{1}_{1,x,0}), G(\mathbb{1}_{x,0}))\\
R_{0,x}(G)(\varphi) &= (G(\mathbb{1}_{0,x,0} \times \psi), G(\mathbb{1}_{0,x,0}), G(\mathbb{1}_{x,0}), G(\mathbb{1}_{0,x,1} \times \psi), G(\mathbb{1}_{0,x,1}), G(\mathbb{1}_{x,1}))
\end{aligned}$$

observe that

$$C_{1,x}(G)(f) = \frac{G(\mathbb{1}_{1,x,1} \times f)/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0} \times f)/G(\mathbb{1}_{x,0})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})} = q(R_{1,x}(G)(f))$$
$$C_{0,x}(G)(f) = \frac{G(\mathbb{1}_{0,x,0} \times f)/G(\mathbb{1}_{x,0}) - G(\mathbb{1}_{0,x,1} \times f)/G(\mathbb{1}_{x,1})}{G(\mathbb{1}_{0,x,0})/G(\mathbb{1}_{x,0}) - G(\mathbb{1}_{0,x,1})/G(\mathbb{1}_{x,1})} = q(R_{0,x}(G)(f))$$

Thus, the same argument shows $C_{0,x} : \mathbb{D}_C \to \ell^\infty(\mathcal{F}_{0,x})$ is fully Hadamard differentiable at any $G \in \mathbb{D}_C$ tangentially to $\ell^\infty(\mathcal{F})$, and $C'_{0,x,G}(H)(f)$ can be found with the appropriate substitutions in (1.61) above.

3. Finally consider $C_{s,x}$. Notice that

$$\mathbb{D}_{q_{s,x}} = \Big\{ \{p_{1,x,1}, p_{x,1}, p_{1,x,0}, p_{x,0}, p_x\}_{x \in \mathcal{X}} \in \mathbb{R}^{5M} ; \\ p_{x,1} > 0, p_{x,0} > 0, p_{1,x,1}/p_{x,1} - p_{1,x,0}/p_{x,0} > 0, p_x > 0 \text{ for all } x \in \mathcal{X} \Big\}$$
$$q_{s,x} : \mathbb{D}_{q_{s,x}} \to \mathbb{R},$$
$$q_{s,x}(\{p_{1,x_m,1}, p_{x_m,1}, p_{1,x_m,0}, p_{x_m,0}\}_{m=1}^M) = \frac{(p_{1,x,1}/p_{x,1} - p_{1,x,0}/p_{x,0})p_x}{\sum_{m=1}^M (p_{1,x_m,1}/p_{x_m,1} - p_{1,x_m,0}/p_{x_m,0})p_{x_m}}$$

is continuously differentiable at any point in $\mathbb{D}_{q_{s,x}}$ with gradient

$$\nabla q(\{p_{1,x_m,1}, p_{x_m,1}, p_{1,x_m,0}, p_{x_m,0}, p_{x_m}\}_{m=1}^M) \in \mathbb{R}^{5M}$$

Furthermore, notice that for any $G \in \mathbb{D}_C$, $C_{s,x}(G) = q_{s,x}(R_{s,x}(G))$, where

$$R_{s,x} : \ell^\infty(\mathcal{F}) \to \mathbb{R}^{5M},$$
$$R_{s,x}(G) = (\{G(\mathbb{1}_{1,x_m,1}), G(\mathbb{1}_{x_m,1}), G(\mathbb{1}_{1,x_m,0}), G(\mathbb{1}_{x_m,0}), G(\mathbb{1}_{x_m})\}_{m=1}^M)$$

It follows that $C_{s,x} : \mathbb{D}_C \to \mathbb{R}$ is fully Hadamard differentiable at any $G \in \mathbb{D}_C$ tangen-

tially to $\ell^\infty(\mathcal{F})$. The derivative is

$$\begin{aligned} C'_{s,x,G}(H) = \sum_{m=1}^{M} &\frac{\partial q_{s,x}}{\partial p_{1,x_m,1}}(R_{s,x}(G)) \times H(\mathbb{1}_{1,x_m,1}) + \frac{\partial q_{s,x}}{\partial p_{x_m,1}}(R_{s,x}(G)) \times H(\mathbb{1}_{x_m,1}) \\ &+ \frac{\partial q_{s,x}}{\partial p_{1,x_m,0}}(R_{s,x}(G)) \times H(\mathbb{1}_{1,x_m,0}) + \frac{\partial q_{s,x}}{\partial p_{x_m,0}}(R_{s,x}(G)) \times H(\mathbb{1}_{x_m,0}) \\ &+ \frac{\partial q_{s,x}}{\partial p_{x_m}}(R_{s,x}(G)) \times H(\mathbb{1}_{x_m}) \end{aligned}$$

This completes the proof. □

Lemma 1.9.17 (T_1 is fully Hadamard differentiable)**.** *Let $\mathcal{F}$ be defined by* (1.59) *and $\mathbb{D}_C$ by* (1.60)*. Let $C_{d,x}$ and $C_{s,x}$ be as defined in lemma 1.9.16, and*

$$\tilde{\eta}_{d,x} : \mathbb{D}_C \to \mathbb{R}^{K_d}, \qquad \tilde{\eta}_{d,x}(G) = \left(C_{d,x}(G)(\eta_{d,x}^{(1)}), \ldots, C_{d,x}(G)(\eta_{d,x}^{(K_d)})\right)$$

Further define

$$\begin{aligned} T_1 &: \mathbb{D}_C \to \prod_{m=1}^{M} \ell^\infty(\mathcal{F}_{1,x_m}) \times \ell^\infty(\mathcal{F}_{0,x_m}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R} \\ T_1(G) &= \left(\{C_{1,x}(G), C_{0,x}(G), \tilde{\eta}_{1,x}(G), \tilde{\eta}_{0,x}(G), C_{s,x}(G)\}_{x\in\mathcal{X}}\right) \end{aligned}$$

T_1 is fully Hadamard differentiable at any $G \in \mathbb{D}_C$ tangentially to $\ell^\infty(\mathcal{F})$.

Proof. Lemma 1.9.16 shows that $C_{d,x}$ and $C_{s,x}$ are fully Hadamard differentiable at any $G \in \mathbb{D}_C$ tangentially to $\ell^\infty(\mathcal{F})$.

Define the evaluation maps

$$ev_{\eta_d^{(k)}} : \ell^\infty(\mathcal{F}_{d,x}) \to \mathbb{R}, \qquad ev_{\eta_d^{(k)}}(H) = H(\eta_d^{(k)})$$

Note that each $ev_{\eta_d^{(k)}}$ is continuous and linear, and is therefore fully Hadamard differentiable

at any $H \in \ell^\infty(\mathcal{F}_{d,x})$ tangentially to $\ell^\infty(\mathcal{F}_{d,x})$ (and is its own derivative). Moreover,

$$\tilde{\eta}_{d,x}(G) = (ev_{\eta_d^{(1)}}(C_{d,x}(G)), \ldots, ev_{\eta_d^{(K_1)}}(C_{d,x}(G)))$$

is the composition of an inner function that is fully Hadamard differentiable at any $G \in \mathbb{D}_C$, and an other function that is fully differentiable at any $H \in \ell^\infty(\mathcal{F}_{d,x})$. Therefore $\tilde{\eta}_{d,x}$ is fully Hadamard differentiable at any $G \in \mathbb{D}_C$ tangentially to $\ell^\infty(\mathcal{F})$.

Next apply lemma 1.9.51 to find that

$$T_1 : \mathbb{D}_C \to \prod_{m=1}^{M} \ell^\infty(\mathcal{F}_{1,x_m}) \times \ell^\infty(\mathcal{F}_{0,x_m}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$$
$$T_1(G) = \left(\{C_{1,x}(G), C_{0,x}(G), \tilde{\eta}_{1,x}(G), \tilde{\eta}_{0,x}(G), C_{s,x}(G)\}_{x\in\mathcal{X}}\right)$$

is fully Hadamard differentiable at any $G \in \mathbb{D}_C$ tangentially to $\ell^\infty(\mathcal{F})$. □

Support of the weak limit of $\sqrt{n}(T_1(\mathbb{P}_n) - T_1(P))$

The next few lemmas study the support of the asymptotic distribution of $\sqrt{n}(T_1(\mathbb{P}_n) - T_1(P))$; in particular, it concentrates on the tangent set of the next map studied in appendix 1.9.3.3.

Lemma 1.9.18 (Continuity of $C'_{d,x,G}(H)(\cdot)$)**.** *Let $C_{d,x}$ be as defined in lemma 1.9.16. If $G, H \in \mathcal{C}(\mathcal{F}, L_{2,P})$, then $C'_{d,x,G}(H) \in \mathcal{C}(\mathcal{F}_{d,x}, L_{2,P})$.*

Proof. Consider $C'_{1,x,G}(H)$ first. Fix $f \in \mathcal{F}_{1,x}$ and let $\varepsilon > 0$. Let

$$\mathrm{Coef}_1(G) = \left[\frac{1/G(\mathbb{1}_{x,1})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}\right]$$
$$\mathrm{Coef}_2(G) = \left[\frac{-1/G(\mathbb{1}_{x,0})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}\right]$$

and use display (1.61) to see that

$$
\begin{aligned}
&|C'_{1,x,G}(H)(f) - C'_{1,x,G}(H)(g)| \\
&= \Bigg| \mathrm{Coef}_1(G) \times [H(\mathbb{1}_{1,x,1} \times f) - H(\mathbb{1}_{1,x,1} \times g)] \\
&\quad + \mathrm{Coef}_1(G) \times (-[C_{1,x}(G)(f) - C_{1,x}(G)(g)])\, H(\mathbb{1}_{1,x,1}) \\
&\quad + \mathrm{Coef}_1(G) \\
&\qquad \times \frac{[C_{1,x}(G)(f) - C_{1,x}(G)(g)] \times G(\mathbb{1}_{1,x,1}) - [G(\mathbb{1}_{1,x,1} \times f) - G(\mathbb{1}_{1,x,1} \times g)]}{G(\mathbb{1}_{x,1})} H(\mathbb{1}_{x,1}) \\
&\quad + \mathrm{Coef}_2(G) \times [H(\mathbb{1}_{1,x,0} \times f) - H(\mathbb{1}_{1,x,0} \times g)] \\
&\quad + \mathrm{Coef}_2(G) \times (-[(C_{1,x}(G)(f)) - C_{1,x}(G)(g)])\, H(\mathbb{1}_{1,x,0}) \\
&\quad + \mathrm{Coef}_2(G) \\
&\qquad \times \frac{[C_{1,x}(G)(f) - C_{1,x}(G)(g)] \times G(\mathbb{1}_{1,x,0}) - [G(\mathbb{1}_{1,x,0} \times f) - G(\mathbb{1}_{1,x,0} \times g)]}{G(\mathbb{1}_{x,0})} H(\mathbb{1}_{x,0}) \Bigg|
\end{aligned}
$$

Recall that $C_{1,x}(G)(f) = \frac{G(\mathbb{1}_{1,x,1} \times f)/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0} \times f)/G(\mathbb{1}_{x,0})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}$, and thus

$$
\begin{aligned}
&C_{1,x}(G)(f) - C_{1,x}(G)(g) \\
&= \frac{[G(\mathbb{1}_{1,x,1} \times f) - G(\mathbb{1}_{1,x,1} \times g)]/G(\mathbb{1}_{x,1}) - [G(\mathbb{1}_{1,x,0} \times f) - G(\mathbb{1}_{1,x,0} \times g)]/G(\mathbb{1}_{x,0})}{G(\mathbb{1}_{1,x,1})/G(\mathbb{1}_{x,1}) - G(\mathbb{1}_{1,x,0})/G(\mathbb{1}_{x,0})}
\end{aligned}
$$

use this to see that

$$
\begin{aligned}
&|C'_{1,x,G}(H)(f) - C'_{1,x,G}(H)(g)| \\
&\qquad \leq A_1 \times |H(\mathbb{1}_{1,x,1} \times f) - H(\mathbb{1}_{1,x,1} \times g)| + A_2 \times |G(\mathbb{1}_{1,x,1} \times f) - G(\mathbb{1}_{1,x,1} \times g)| \\
&\qquad\qquad + A_3 \times |H(\mathbb{1}_{1,x,0} \times f) - H(\mathbb{1}_{1,x,0} \times g)| + A_4 \times |G(\mathbb{1}_{1,x,0} \times f) - G(\mathbb{1}_{1,x,0} \times g)|
\end{aligned} \tag{1.62}
$$

for finite constants A_1, A_2, A_3, and A_4 that depend on G and H, but not on f or g. Now

use $G, H \in \mathcal{C}(\mathcal{F}, L_{2,P})$ to choose $\delta_{z,H} > 0$ and $\delta_{z,G} > 0$ such that

$$\begin{aligned}
L_{2,P}(\mathbb{1}_{1,x,1} \times f, \mathbb{1}_{1,x,1} \times g) < \delta_{1,H} \quad &\Longrightarrow \quad |H(\mathbb{1}_{1,x,1} \times f) - H(\mathbb{1}_{1,x,1} \times g)| < \varepsilon/(4A_1) \\
L_{2,P}(\mathbb{1}_{1,x,1} \times f, \mathbb{1}_{1,x,1} \times g) < \delta_{1,G} \quad &\Longrightarrow \quad |G(\mathbb{1}_{1,x,1} \times f) - G(\mathbb{1}_{1,x,1} \times g)| < \varepsilon/(4A_2) \\
L_{2,P}(\mathbb{1}_{1,x,0} \times f, \mathbb{1}_{1,x,0} \times g) < \delta_{0,H} \quad &\Longrightarrow \quad |H(\mathbb{1}_{1,x,0} \times f) - H(\mathbb{1}_{1,x,0} \times g)| < \varepsilon/(4A_3) \\
L_{2,P}(\mathbb{1}_{1,x,0} \times f, \mathbb{1}_{1,x,0} \times g) < \delta_{0,G} \quad &\Longrightarrow \quad |G(\mathbb{1}_{1,x,0} \times f) - G(\mathbb{1}_{1,x,0} \times g)| < \varepsilon/(4A_4) \quad (1.63)
\end{aligned}$$

Finally, notice that

$$\begin{aligned}
L_{2,P}(\mathbb{1}_{1,x,z} \times f, \mathbb{1}_{1,x,z} \times g) &= \sqrt{P((\mathbb{1}_{1,x,z} \times f - \mathbb{1}_{1,x,z} \times g)^2)} = \sqrt{P(\mathbb{1}_{1,x,z} \times (f-g)^2)} \\
&\leq \sqrt{P((f-g)^2)} = L_{2,P}(f,g) \qquad (1.64)
\end{aligned}$$

It follows from (1.62), (1.63), and (1.64) that

$$L_{2,P}(f,g) < \min\{\delta_{1,H}, \delta_{1,G}, \delta_{0,H}, \delta_{0,G}\} \qquad \Longrightarrow \qquad |C'_{1,x,G}(H)(f) - C'_{1,x,G}(H)(g)| < \varepsilon$$

i.e., $C'_{1,x,G}(H)(\cdot)$ is continuous at f. Since $f \in \mathcal{F}_{1,x}$ and $G, H \in \mathcal{C}(\mathcal{F}, L_{2,P})$ were arbitrary, this shows that $G, H \in \mathcal{C}(\mathcal{F}, L_{2,P})$ implies $C'_{1,x,G}(H) \in \mathcal{C}(\mathcal{F}_{1,x}, L_{2,P})$.

The same argument shows that $G, H \in \mathcal{C}(\mathcal{F}, L_{2,P})$ implies $C'_{0,x,G}(H) \in \mathcal{C}(\mathcal{F}_{0,x}, L_{2,P})$. This completes the proof. □

Lemma 1.9.19 (Support of $T'_{1,P}(\mathbb{G})$)**.** *Let $\mathcal{F}$ be defined by* (1.59) *and T_1 be as defined in lemma 1.9.17.*

1. *If assumption 1 holds, $P \in \mathbb{D}_C$ and hence T_1 is fully Hadamard differentiable at P tangentially to $\ell^\infty(\mathcal{F})$.*

2. *If assumptions 1, 2, and 3 hold,*

$$\sqrt{n}(T_1(\mathbb{P}_n) - T_1(P)) \xrightarrow{L} T'_{1,P}(\mathbb{G})$$

where $\mathbb{G}$ is the Gaussian limit of $\sqrt{n}(\mathbb{P}_n - P)$ in $\ell^\infty(\mathcal{F})$ discussed in lemma 1.9.15.

3. *If assumptions 1, 2, and 3, then $P(T'_{1,P}(\mathbb{G}) \in \mathbb{D}_{Tan,Full}) = 1$ where*

$$\mathbb{D}_{Tan,Full} = \prod_{m=1}^{M} \left(\ell^\infty_{\mathcal{Y}_{1,x_m}}(\mathcal{F}_{1,x_m}) \times \ell^\infty_{\mathcal{Y}_{0,x_m}}(\mathcal{F}_{0,x_m})\right) \cap \left(\mathcal{C}(\mathcal{F}_{1,x_m}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x_m}, L_{2,P})\right) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R} \tag{1.65}$$

Proof. In steps:

1. $P \in \mathbb{D}_C$ and differentiability of T_1 at P.

   Assumption 1 implies $P \in \mathbb{D}_C$, given by (1.60). To see this, recall that assumption 1 (iv) is that $P(\mathbb{1}_{x,z}) = P(X = x, Z = z) > 0$ (implying $P(\mathbb{1}_x) = P(X = x) = P(X = x, Z = 1) + P(X = x, Z = 0) > 0$). Furthermore,

$$\begin{aligned} &P(\mathbb{1}_{d,x,d})/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d})/P(\mathbb{1}_{x,1-d}) \\ &\quad = P(D = d \mid X = x, Z = d) - P(D = d \mid X = x, Z = 1 - d) \\ &\quad = P(D_1 > D_0 \mid X = x) > 0 \end{aligned}$$

   The second equality is shown in the proof of lemma 1.2.1, and the inequality is assumption 1 (iii). Lemma 1.9.17 thus shows that T_1 is fully Hadamard differentiable at P tangentially to $\ell^\infty(\mathcal{F})$.

2. Functional delta method.

   Under assumptions 1, 2, and 3, lemma 1.9.15 shows that $\sqrt{n}(\mathbb{P}_n - P) \xrightarrow{L} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$.

The functional delta method (van der Vaart (2007) theorem 20.8) then implies

$$\sqrt{n}(T_1(\mathbb{P}_n) - T_1(P)) \xrightarrow{L} T'_{1,P}(\mathbb{G}), \quad \text{in } \prod_{m=1}^{M} \ell^\infty(\mathcal{F}_{1,x_m}) \times \ell^\infty(\mathcal{F}_{0,x_m}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$$

3. Support of $T'_{1,P}(\mathbb{G})$.

   Notice that $T'_P(\mathbb{G}) = \left(\left\{C'_{1,x,P}(\mathbb{G}), C'_{0,x,P}(\mathbb{G}), \tilde{\eta}'_{1,x,P}(\mathbb{G}), \tilde{\eta}'_{0,x,P}(\mathbb{G}), C'_{s,x,P}(\mathbb{G})\right\}_{x\in\mathcal{X}}\right)$, where $\tilde{\eta}_{d,x}$ are defined in lemma 1.9.17. Let

$$\mathbb{S}_x = \left(\ell^\infty_{\mathcal{Y}_{1,x_m}}(\mathcal{F}_{1,x_m}) \times \ell^\infty_{\mathcal{Y}_{0,x_m}}(\mathcal{F}_{0,x_m})\right) \cap \left(\mathcal{C}(\mathcal{F}_{1,x_m}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x_m}, L_{2,P})\right) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$$

   and note that it suffices to show

$$P\left(C'_{1,x,P}(\mathbb{G}), C'_{0,x,P}(\mathbb{G}), \tilde{\eta}'_{1,x,P}(\mathbb{G}), \tilde{\eta}'_{0,x,P}(\mathbb{G}), C'_{s,x,P}(\mathbb{G}) \in \mathbb{S}_x\right) = 1$$

   for each x. Moreover,

$$P\left(\left(\tilde{\eta}'_{1,x,P}(\mathbb{G}), \tilde{\eta}'_{0,x,P}(\mathbb{G}), C'_{s,x,P}(\mathbb{G})\right) \in \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}\right) = 1$$

   is immediate. To complete the proof we must show $P(C'_{d,x,P}(\mathbb{G}) \in \ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{d,x})) = P(C'_{d,x,P}(\mathbb{G}) \in \mathcal{C}(\mathcal{F}_{d,x}, L_{2,P})) = 1$.

   (a) To see that $P(C'_{d,x,P}(\mathbb{G}) \in \mathcal{C}(\mathcal{F}_{d,x}, L_{2,P})) = 1$, first note that for any functions $f_1, f_2 \in \mathcal{F}$,

$$\begin{aligned} |P(f_1) - P(f_2)| &\leq P(|f_1 - f_2|) = P(\sqrt{(f_1 - f_2)}) \\ &\leq \sqrt{P((f_1 - f_2)^2)} = L_{2,P}(f_1, f_2) \end{aligned}$$

   where the second inequality is an application of Jensen's inequality. Thus $P \in \mathcal{C}(\mathcal{F}, L_{2,P})$.

Next apply lemma 1.9.18 to see that $G \in \mathcal{C}(\mathcal{F}, L_{2,P})$ implies $C'_{d,x,P}(G) \in \mathcal{C}(\mathcal{F}_{d,x}, L_{2,P})$. It follows that

$$1 = P\left(\mathbb{G} \in \mathcal{C}(\mathcal{F}, L_{2,P})\right) \leq P\left(C'_{d,x,P}(\mathbb{G}) \in \mathcal{C}(\mathcal{F}_{d,x}, L_{2,P})\right)$$

(b) To see that $P(C'_{d,x,P}(\mathbb{G}) \in \ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{d,x})) = 1$, we show that $P(\sqrt{n}(C_{d,x}(\mathbb{P}_n) - C_{d,x}(P)) \in \ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{d,x})) = 1$.

First recall the definition given in (1.55):

$$\ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{d,x}) = \Big\{ H \in \ell^\infty(\mathcal{F}_{d,x}) \,;\ \text{for all } a, b \in \mathbb{R} \text{ and } f, g \in \mathcal{F}_{d,x},$$
$$H(f) = H(\mathbb{1}_{\mathcal{Y}_{d,x}} \times f),\ \text{if } a \in \mathcal{F}_{d,x} \text{ then } H(a) = 0,\ \text{and}$$
$$\text{if } af + bg \in \mathcal{F}_{d,x} \text{ then } H(af + bg) = aH(f) + bH(g) \Big\}$$

i. $\sqrt{n}(C_{d,x}(\mathbb{P}_n) - C_{d,x}(P))$ is linear and evaluates constants to zero.

This follows because $C_{d,x}(\mathbb{P}_n)$ and $C_{d,x}(P)$ are linear and "return constants". To see this, recall that $C_{d,x}(P) \in \ell^\infty(\mathcal{F}_{d,x})$ is given pointwise by

$$C_{d,x}(P)(f) = \frac{P(\mathbb{1}_{d,x,d} \times f)/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d} \times f)/P(\mathbb{1}_{x,1-d})}{P(\mathbb{1}_{d,x,d})/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d})/P(\mathbb{1}_{x,1-d})}$$

Use this to see that for any $a, b \in \mathbb{R}$ and $f, g \in \mathcal{F}_{d,x}$. if $af + bg \in \mathcal{F}_{d,x}$, then linearity of P implies $C_{d,x}(P)(af + bg) = aC_{d,x}(P)(f) + bC_{d,x}(P)(g)$ and $C_{d,x}(\mathbb{P}_n)(af + bg) = aC_{d,x}(\mathbb{P}_n)(f) + bC_{d,x}(\mathbb{P}_n)(g)$. Similarly, if $a \in \mathcal{F}_{d,x}$ is the constant function always returning a, then $C_{d,x}(P)(a) = a$. The same observations apply to $C_{d,x}(\mathbb{P}) \in \ell^\infty(\mathcal{F}_{d,x})$.

Therefore

$$\begin{aligned}
&\sqrt{n}(C_{d,x}(\mathbb{P}_n) - C_{d,x}(P))(af + bg) \\
&\quad = \sqrt{n}(C_{d,x}(\mathbb{P}_n)(af + bg) = -C_{d,x}(P)(af + bg)) \\
&\quad = \sqrt{n}(aC_{d,x}(\mathbb{P}_n)(f) + bC_{d,x}(\mathbb{P}_n)(g) - aC_{d,x}(P)(f) - bC_{d,x}(P)(g)) \\
&\quad = a \times \sqrt{n}(C_{d,x}(\mathbb{P}_n) - C_{d,x}(P))(f) + b \times \sqrt{n}(C_{d,x}(\mathbb{P}_n) - C_{d,x}(P))(g)
\end{aligned}$$

and furthermore, if $a \in \mathcal{F}_{d,x}$, then

$$\sqrt{n}(C_{d,x}(\mathbb{P}_n) - C_{d,x}(P))(a) = \sqrt{n}(a - a) = 0$$

ii. $C_{d,x}(P)$ "ignores values outside $\mathcal{Y}_{d,x}$"; i.e. $C_{d,x}(P)(f) = C_{d,x}(P)(\mathbb{1}_{\mathcal{Y}_{d,x}} \times f)$. To see this, notice

$$\begin{aligned}
&C_{d,x}(P)(f) \qquad (1.66)\\
&= \frac{E[f(Y)\mathbb{1}\{D = d\} \mid X = x, Z = d] - E[f(Y)\mathbb{1}\{D = d\} \mid X = x, Z = 1 - d]}{P(\mathbb{1}_{d,x,d})/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d})/P(\mathbb{1}_{x,1-d})} \\
&= \frac{P(D = d \mid X = x, Z = d)E[f(Y) \mid D = d, X = x, Z = d]}{P(\mathbb{1}_{d,x,d})/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d})/P(\mathbb{1}_{x,1-d})} \\
&\qquad - \frac{P(D = d \mid X = x, Z = 1 - d)E[f(Y) \mid D = d, X = x, Z = 1 - d]}{P(\mathbb{1}_{d,x,d})/P(\mathbb{1}_{x,d}) - P(\mathbb{1}_{d,x,1-d})/P(\mathbb{1}_{x,1-d})}.
\end{aligned}$$

Since $\mathcal{Y}_{d,x}$ is the support of $Y \mid D = d, X = x$,

$$\begin{aligned}
&E[f(Y) \mid D = d, X = x, Z = z] \\
&\quad = E[f(Y)\mathbb{1}\{Z = z\} \mid D = d, X = x]/P(Z = z \mid D = d, X = x) \\
&\quad = \frac{E[\mathbb{1}\{Y \in \mathcal{Y}_{d,x}\}f(Y)\mathbb{1}\{Z = z\} \mid D = d, X = x]}{P(Z = z \mid D = d, X = x)} \\
&\quad = E[\mathbb{1}\{Y \in \mathcal{Y}_{d,x}\}f(Y) \mid D = d, X = x, Z = z]
\end{aligned}$$

Along with (1.66), this implies $C_{d,m}(P)(f) = C_{d,m}(P)(\mathbb{1}_{\mathcal{Y}_{d,x}} \times f)$.

iii. Now notice that with probability one the sample is a subset of the support, and when this is so, $C_{d,x}(\mathbb{P}_n)$ ignores values outside of $\mathcal{Y}_{d,x}$.

Specifically, observe that

$$
\begin{aligned}
&C_{d,x}(\mathbb{P}_n)(f) \qquad (1.67)\\
&= \frac{\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{D_i = d, X_i = x\}\mathbb{1}\{Z_i = d\}f(Y_i)\right] / \left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i = x, Z_i = d\}\right]}{\mathbb{P}_n(\mathbb{1}_{d,x,d})/\mathbb{P}_n(\mathbb{1}_{x,d}) - \mathbb{P}_n(\mathbb{1}_{d,x,1-d})/\mathbb{P}_n(\mathbb{1}_{x,1-d})}\\
&\quad - \frac{\frac{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{D_i=d,X_i=x\}\mathbb{1}\{Z_i=1-d\}f(Y_i)}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i=x,Z_i=1-d\}}}{\mathbb{P}_n(\mathbb{1}_{d,x,d})/\mathbb{P}_n(\mathbb{1}_{x,d}) - \mathbb{P}_n(\mathbb{1}_{d,x,1-d})/\mathbb{P}_n(\mathbb{1}_{x,1-d})}
\end{aligned}
$$

Note that because $\mathcal{Y}_{d,x}$ is the support of $Y \mid D = d, X = x$, we have that with probability one, $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n \subseteq \mathcal{S} \equiv \bigcup_{d,z,x} \mathcal{Y}_{d,x} \times \{d\} \times \{z\} \times \{x\}$. Indeed, since $\mathcal{Y}_{d,x} \times \{d\} \times \{z\} \times \{x\} \subseteq \mathbb{R}^4$ are disjoint for each distinct (d, z, x),

$$
\begin{aligned}
P((Y_i, D_i, Z_i, X_i) \in \mathcal{S}) &= P\left((Y_i, D_i, Z_i, X_i) \in \bigcup_{d,z,x} \mathcal{Y}_{d,x} \times \{d\} \times \{z\} \times \{x\}\right)\\
&= \sum_{d,z,x} P(Y_i \in \mathcal{Y}_{d,x}, D_i = d, X_i = x, Z_i = z)\\
&= \sum_{d,z,x} P(D_i = d, X_i = x, Z_i = z) \times \frac{\overbrace{P(Y_i \in \mathcal{Y}_{d,x}, Z_i = z \mid D_i = d, X_i = x)}^{=P(Z_i=z|D_i=d,X_i=x)}}{P(Z_i = z \mid D_i = d, X_i = x)}\\
&= \sum_{d,z,x} P(D_i = d, X_i = x, Z_i = z) = 1
\end{aligned}
$$

Since $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$ is i.i.d.,

$$
\begin{aligned}
P\left(\{Y_i, D_i, Z_i, X_i\}_{i=1}^n \subseteq \mathcal{S}\right) &= P\left(\bigcap_{i=1}^n \{(Y_i, D_i, Z_i, X_i) \in \mathcal{S}\}\right)\\
&= \prod_{i=1}^n P\left((Y_i, D_i, Z_i, X_i) \in \mathcal{S}\right) = 1
\end{aligned}
$$

When $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n \subseteq \mathcal{S}$ holds, $\mathbb{1}\{D_i = d, X_i = x\} \leq \mathbb{1}\{Y_i \in \mathcal{Y}_{d,x}\} = \mathbb{1}_{\mathcal{Y}_{d,x}}(Y_i)$ and thus $\mathbb{1}_{\mathcal{Y}_{d,x}}(Y_i) \times \mathbb{1}\{D_i = d, X_i = x\} = \mathbb{1}\{D_i = d, X_i = x\}$. This and (1.67) implies that when $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n \subseteq \mathcal{S}$ holds,

$$C_{d,x}(\mathbb{P}_n)(f) = C_{d,x}(\mathbb{P}_n)(\mathbb{1}_{\mathcal{Y}_{d,x}} \times f)$$

iv. Use the facts established above to see that

$$\begin{aligned}
&P(\sqrt{n}(C_{d,x}(\mathbb{P}_n) - C_{d,x}(P)) \in \ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{1,x})) \\
&\quad = P(\sqrt{n}(C_{d,x}(\mathbb{P}_n) - C_{d,x}(P)) \in \ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{d,x}) \mid \{Y_i, D_i, Z_i, X_i\}_{i=1}^n \subseteq \mathcal{S}) \\
&\quad = 1
\end{aligned}$$

Lemma 1.9.9 is that $\ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{1,x})$ is closed, so Portmanteau (van der Vaart & Wellner (1997) theorem 1.3.4) implies

$$\begin{aligned}
1 &= \limsup_{n\to\infty} P(\sqrt{n}(C_{d,x}(\mathbb{P}_n) - C_{d,x}(P)) \in \ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{1,x})) \\
&\leq P(C'_{d,x,P}(\mathbb{G}) \in \ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{1,x}))
\end{aligned}$$

In summary, we have

$$\begin{aligned}
1 &= P\left(\left(\tilde{\eta}'_{1,x,P}(\mathbb{G}), \tilde{\eta}'_{0,x,P}(\mathbb{G}), C'_{s,x,P}(\mathbb{G})\right) \in \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}\right) \\
&= P(C'_{d,x,P}(\mathbb{G}) \in \ell^\infty_{\mathcal{Y}_{d,x}}(\mathcal{F}_{d,x})) \\
&= P\left(C'_{d,x,P}(\mathbb{G}) \in \mathcal{C}(\mathcal{F}_{d,x}, L_{2,P})\right)
\end{aligned}$$

From which it follows that

$$1 = P\left(C'_{1,x,P}(\mathbb{G}), C'_{0,x,P}(\mathbb{G}), \tilde{\eta}'_{1,x,P}(\mathbb{G}), \tilde{\eta}'_{0,x,P}(\mathbb{G}), C'_{s,x,P}(\mathbb{G}) \in \mathbb{S}_x\right)$$

for each x, and therefore

$$P(T'_{1,P}(\mathbb{G}) \in \mathbb{D}_{Tan,Full})$$
$$= P\left(\bigcap_{x\in\mathcal{X}} \left\{C'_{1,x,P}(\mathbb{G}), C'_{0,x,P}(\mathbb{G}), \tilde{\eta}'_{1,x,P}(\mathbb{G}), \tilde{\eta}'_{0,x,P}(\mathbb{G}), C'_{s,x,P}(\mathbb{G}) \in \mathbb{S}_x\right\}\right) = 1$$

This completes the proof. □

1.9.3.3 Optimal transport, $T_2(\{P_{1|x}, P_{0|x}, \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}}) = (\{\theta_x^L, \theta_x^H, \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}})$

The second map applies the directional differentiability of optimal transport shown in appendix 1.9.2.2. There are three assumptions in lemma 1.9.7 to verify: strong duality, Donsker conditions, and completeness. Strong duality is shown by lemmas 1.9.38 and 1.9.42, and the Donsker conditions are shown by lemma 1.9.14. It remains to verify the completeness assumptions.

Verifying completeness

Lemma 1.9.20 (Completeness of dual problem feasible set in L_2 for smooth cost functions)**.** *Suppose $\mathcal{Y} \subset \mathbb{R}$ is compact and $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is L-Lipschitz. Let $\mathcal{F}_c$, $\mathcal{F}_c^c$ be given by* (1.14) *and* (1.15) *respectively:*

$$\mathcal{F}_c = \{\varphi : \mathcal{Y} \to \mathbb{R} \,;\; -\|c\|_\infty \le \varphi(y_1) \le \|c\|_\infty,\; |\varphi(y) - \varphi(y')| \le L|y - y'|\}\,,$$
$$\mathcal{F}_c^c = \{\psi : \mathcal{Y} \to \mathbb{R} \,;\; -2\|c\|_\infty \le \psi(y) \le 0,\; |\psi(y) - \psi(y')| \le L|y - y'|\}\,,$$

Further let Φ_c be defined by (1.80)*, and $\mathcal{F}_d$ defined by* (1.57)*. Let $L_{2,P}$ be given by* (1.51)*, and L_2 be given by* (1.52)*. Then $(\mathcal{F}_{1,x} \times \mathcal{F}_{0,x}, L_2)$ and its subset $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ are complete.*

Proof. In steps:

1. $(\mathcal{F}_c, L_{2,P})$ and $(\mathcal{F}_c^c, L_{2,P})$ are complete.

   The proof that $(\mathcal{F}_c, L_{2,P})$ is complete is broken into steps:

   (a) Let $\{\varphi_n\}_{n=1}^{\infty} \subseteq \mathcal{F}_c$ be $L_{2,P}$-Cauchy. The L_p semimetrics are complete for any probability distribution (Pollard (2002) section 2.7 and chapter 2 problem [19]), thus there exists $\tilde{\varphi}$ such that $L_{2,P}(\varphi_n, \tilde{\varphi}) \to 0$. Convergence in $L_{2,P}$ implies convergence almost surely along a subsequence (Pollard (2002) section 2.8). Thus there exists a subsequence $\{\varphi_{n_k}\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty} \varphi_{n_k}(y) = \tilde{\varphi}(y)$ for P-almost every y. Let $N_1 \subseteq \mathcal{Y}$ be the P-negligible set where this fails.

   (b) Observe that on $N_1^c = \mathcal{Y} \setminus N_1$, $\tilde{\varphi}$ obeys the bounds and Lipschitz continuity of $\mathcal{F}_c$. Specifically,

$$-\|c\|_\infty \leq \lim_{k\to\infty} -\|c\|_\infty \leq \underbrace{\lim_{k\to\infty} \varphi_{n_k}(y)}_{\tilde{\varphi}(y)} \leq \lim_{k\to\infty} \|c\|_\infty \leq \|c\|_\infty$$

   Furthermore, for any $y, y' \in N_1^c$,

$$\begin{aligned} |\tilde{\varphi}(y) - \tilde{\varphi}(y')| &= |\lim_{k\to\infty} \varphi_{n_k}(y) - \lim_{k\to\infty} \varphi_{n_k}(y')| = \lim_{k\to\infty} |\varphi_{n_k}(y) - \varphi_{n_k}(y')| \\ &\leq \lim_{k\to\infty} L|y - y'| = L|y - y'| \end{aligned}$$

   (c) Now define functions $\bar{\varphi}, \varphi : \mathcal{Y} \to \mathbb{R}$ with

$$\bar{\varphi}(y_1) = \sup_{y_1' \in N_1^c} \{\tilde{\varphi}(y_1') - L|y_1 - y_1'|\}, \qquad \varphi(y_1) = \max\{\bar{\varphi}(y_1), -\|c\|_\infty\}$$

   Then $L_{2,P}(\varphi_n, \varphi) \to 0$ and $\varphi \in \mathcal{F}_c$, which shows $(\mathcal{F}_c, L_{2,P})$ is complete.

   i. $L_{2,P}(\varphi_n, \varphi) \to 0$ follows from $\varphi(y) = \tilde{\varphi}(y)$ for all $y \in N_1^c$. To see this, let $y \in N_1^c$. Since $\tilde{\varphi}$ is L-Lipschitz on N_1^c, it follows that for any $y' \in N_1^c$,

$$\tilde{\varphi}(y') - L|y - y'| \leq \tilde{\varphi}(y)$$

and thus $\bar{\varphi}(y) = \tilde{\varphi}(y)$. This implies $\bar{\varphi}(y) = \tilde{\varphi}(y) \geq -\|c\|_\infty$, and thus $\varphi(y) = \bar{\varphi}(y) = \tilde{\varphi}(y)$. Thus $\varphi(y) = \tilde{\varphi}(y)$ for P-almost all y, implying $L_{2,P}(\tilde{\varphi}, \varphi) = 0$ and thus $L_{2,P}(\varphi_n, \varphi) \to 0$.

ii. To see that $\varphi \in \mathcal{F}_c$, first notice that $\bar{\varphi}(y) = \sup_{y' \in N_1^c}\{\tilde{\varphi}(y') - L|y - y'|\} \leq \sup_{y' \in N_1^c} \tilde{\varphi}(y) \leq \|c\|_\infty$, and hence $\bar{\varphi}$ obeys the upper bound for $\mathcal{F}_c$. It then follows easily that $\varphi(y) = \max\{\bar{\varphi}(y), -\|c\|_\infty\}$ obeys both the upper and lower bound. Next notice that $\bar{\varphi}$ is L-Lipschitz on all of $\mathcal{Y}$:

$$\begin{aligned}
\bar{\varphi}(y) - \bar{\varphi}(y') = &= \sup_{\tilde{y} \in N_1^c} \{\tilde{\varphi}(\tilde{y}) - L|y - \tilde{y}|\} - \sup_{\tilde{y}' \in N_1^c} \{\tilde{\varphi}(\tilde{y}') - L|y' - \tilde{y}'|\} \\
&\leq \sup_{\tilde{y} \in N_1^c} \{\tilde{\varphi}(\tilde{y}) - L|y - \tilde{y}| - (\tilde{\varphi}(\tilde{y}) - L|y' - \tilde{y}|)\} \\
&= \sup_{\tilde{y} \in N_1^c} L\left(|y' - \tilde{y}| - |y - \tilde{y}|\right) \leq L|y - y'|
\end{aligned}$$

where the last inequality follows from the reverse triangle inequality. It follows that $\varphi(y_1) = \max\{\bar{\varphi}(y_1), -\|c\|_\infty\}$ is also L-Lipschitz, and thus $\varphi \in \mathcal{F}_c$.

2. Very similar steps show that $(\mathcal{F}_c^c, L_{2,P})$ is complete; the only substantial changes are replacing the lower bounds with $-2\|c\|$ and the upper bounds with 0.

3. Note that since $(\mathcal{F}_c \times \mathcal{F}_c^c, L_2)$ is the product space of $(\mathcal{F}_c, L_{2,P})$ and $(\mathcal{F}_c^c, L_{2,P})$, it follows that $(\mathcal{F}_c \times \mathcal{F}_c^c, L_2)$ is complete.

4. $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ is complete.

   To see that $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ is complete, let $\{(\varphi_n, \psi_n)\}_{n=1}^\infty \subseteq \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ be L_2-Cauchy, and follow the same steps shown above to define $(\varphi, \psi) \in \mathcal{F}_c \times \mathcal{F}_c^c$ such that $L_2((\varphi_n, \psi_n), (\varphi, \psi)) \to 0$. It remains to show that $\varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)$ for all $(y_1, y_0) \in \mathcal{Y} \times \mathcal{Y} \subseteq \mathbb{R}^2$.

Since c is L-Lipschitz,

$$c(y_1, y_0) - c(y_1', y_0) \geq -L\|(y_1, y_0) - (y_1', y_0')\| \geq -L|y_1 - y_1'| - L|y_0 - y_0'|$$

which implies $c(y_1', y_0') - L|y_1 - y_1'| - L|y_0 - y_0'| \leq c(y_1, y_0)$. Thus

$$\begin{aligned}
\bar{\varphi}(y_1) + \bar{\varphi}(y_0) &= \sup_{y_1' \in N_1^c} \{\tilde{\varphi}(y_1') - L|y_1 - y_1'|\} + \sup_{y_0' \in N_0^c} \{\tilde{\psi}(y_0') - L|y_0 - y_0'|\} \\
&= \sup_{(y_1', y_0') \in N_1^c \times N_0^c} \left\{\tilde{\varphi}(y_1') + \tilde{\psi}(y_0') - L|y_1 - y_1'| - L|y_0 - y_0'|\right\} \\
&\leq \sup_{(y_1', y_0') \in N_1^c \times N_0^c} \{c(y_1', y_0') - L|y_1 - y_1'| - L|y_0 - y_0'|\} \\
&\leq \sup_{(y_1', y_0') \in N_1^c \times N_0^c} \{c(y_1, y_0)\} = c(y_1, y_0)
\end{aligned}$$

Finally,

$$\begin{aligned}
\varphi(y_1) + \psi(y_0) &= \max\{\bar{\varphi}(y_1), -\|c\|_\infty\} + \max\{\bar{\psi}(y_0), -2\|c\|\} \\
&= \max\{\bar{\varphi}(y_1) + \bar{\varphi}(y_0), \bar{\varphi}(y_1) - 2\|c\|_\infty, -\|c\|_\infty + \bar{\psi}(y_0), -\|c\|_\infty - 2\|c\|\} \\
&\leq \max\{c(y_1, y_0), -\|c\|_\infty, -\|c\|_\infty, -3\|c\|_\infty\} \\
&\leq c(y_1, y_0)
\end{aligned}$$

where the first inequality follows from $\bar{\varphi}(y_1) \leq \|c\|_\infty$ and $\bar{\psi}(y_0) \leq 0$.

5. $(\mathcal{F}_{1,x} \times \mathcal{F}_{0,x}, L_2)$ is complete.

   As this is the product space of $(\mathcal{F}_{1,x}, L_{2,P})$ and $(\mathcal{F}_{0,x}, L_{2,P})$, it suffices to show these individual spaces are complete.

Now recall that $\mathcal{F}_{d,x}$ is defined by (1.57):

$$\tilde{\mathcal{F}}_1 = \left\{ f : \mathcal{Y} \to \mathbb{R} \; ; \; f = \varphi \text{ for some } \varphi \in \mathcal{F}_c, \text{ or } f = \eta_1^{(k)} \text{ for some } k = 1, \ldots, K_1 \right\}$$
$$\tilde{\mathcal{F}}_0 = \left\{ f : \mathcal{Y} \to \mathbb{R} \; ; \; f = \psi \text{ for some } \psi \in \mathcal{F}_c^c, \text{ or } f = \eta_0^{(k)} \text{ for some } k = 1, \ldots, K_0 \right\}$$
$$\mathcal{F}_{d,x} = \left\{ f : \mathcal{Y} \to \mathbb{R} \; ; \; f = g \text{ or } \mathbb{1}_{\mathcal{Y}_{d,x}} \times g \text{ for some } g \in \tilde{\mathcal{F}}_d \right\}$$

Recall that the union of a finite number of complete sets is complete. Since $(\mathcal{F}_c, L_{2,P})$ and $\mathcal{F}_c^c, L_{2,P})$ are complete and any finite set is complete, $\tilde{\mathcal{F}}_d$ is complete. Next recognize that $\mathcal{F}_{d,x} = \tilde{\mathcal{F}}_d \cup \left\{ \mathbb{1}_{\mathcal{Y}_{d,x}} \times g \; ; \; g \in \tilde{\mathcal{F}}_d \right\}$ is the union of a finite number of sets, and thus it suffices to show $\left\{ \mathbb{1}_{\mathcal{Y}_{d,x}} \times g \; ; \; g \in \tilde{\mathcal{F}}_d \right\}$ is complete.

Let $\{\mathbb{1}_{\mathcal{Y}_{d,x}} \times g_n\}_{n=1}^{\infty} \subseteq \left\{ \mathbb{1}_{\mathcal{Y}_{d,x}} \times g \; ; \; g \in \tilde{\mathcal{F}}_d \right\}$ be $L_{2,P}$-Cauchy. Lemma 1.9.14 shows that $\mathcal{F}_{d,x}$ is Donsker and $\sup_{f \in \mathcal{F}_{d,x}} |P(f)| < \infty$, which implies $(\mathcal{F}_{d,x}, L_{2,P})$ is totally bounded (see van der Vaart & Wellner (1997) problem 2.1.2.). Since $\tilde{\mathcal{F}}_d$ is a complete subset of a totally bounded set, it is compact. Thus $\{g_n\}_{n=1}^{\infty} \subseteq \tilde{\mathcal{F}}_d$ is a sequence in a compact semimetric space, and therefore has a convergent subsequence $\{g_{n_k}\}_{k=1}^{\infty}$. Let $g \in \tilde{\mathcal{F}}_d$ be its limit, and notice that

$$\begin{aligned} 0 \leq L_{2,P}(\mathbb{1}_{\mathcal{Y}_{d,x}} \times g_{n_k}, \mathbb{1}_{\mathcal{Y}_{d,x}} \times g) &= \sqrt{P((\mathbb{1}_{\mathcal{Y}_{d,x}} \times g_{n_k} - \mathbb{1}_{\mathcal{Y}_{d,x}} \times g)^2)} \\ &\leq \sqrt{P((g_{n_k} - g)^2)} \\ &= L_{2,P}(g_{n_k}, g) \to 0 \end{aligned}$$

and thus $\mathbb{1}_{\mathcal{Y}_{d,x}} \times \varphi_{n_k} \to \mathbb{1}_{\mathcal{Y}_{d,x}} g$. It follows that $\mathbb{1}_{\mathcal{Y}_{d,x}} \times \varphi_n \to \mathbb{1}_{\mathcal{Y}_{d,x}} g$, and thus $\left\{ \mathbb{1}_{\mathcal{Y}_{d,x}} \times g \; ; \; g \in \tilde{\mathcal{F}}_d \right\}$ is complete.

This completes the proof. □

Lemma 1.9.21 (Completeness of dual problem feasible set in L_2 for indicator cost functions)**.** *Let $\mathcal{Y} \subseteq \mathbb{R}$, $C \subseteq \mathcal{Y} \times \mathcal{Y}$ be nonempty, open, and convex, and let $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be*

given by $c(y_1, y_0) = \mathbb{1}_C(y_1, y_0) = \mathbb{1}\{(y_1, y_0) \in C\}$. *Let* $\mathcal{F}_c$, $\mathcal{F}_c^c$ *be given by* (1.16) *and* (1.17) *respectively:*

$$\mathcal{F}_c = \{\varphi : \mathcal{Y} \to \mathbb{R} \,;\, \varphi(y_1) = \mathbb{1}_I(y_1) \text{ for some interval } I\},$$
$$\mathcal{F}_c^c = \{\psi : \mathcal{Y} \to \mathbb{R} \,;\, \psi(y_0) = -\mathbb{1}_{I^c}(y_0) \text{ for some interval } I\},$$

Further let Φ_c *be defined by* (1.80)*, and* $\mathcal{F}_{d,x}$ *defined by* (1.57)*. Let* $L_{2,P}$ *be given by* (1.51)*, and* L_2 *be given by* (1.52)*. Then* $(\mathcal{F}_{1,x} \times \mathcal{F}_{0,x}, L_2)$ *and its subset* $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ *are complete.*

Proof. The proof is similar in structure to that of lemma 1.9.20.

1. $(\mathcal{F}_c, L_{2,P})$ is complete.

   Let $\{\varphi_n\}_{n=1}^\infty \subseteq \mathcal{F}_c$ be $L_{2,P}$-Cauchy. Note that $\varphi_n(y) = \mathbb{1}_{I_n}(y)$ for some interval I_n. Just as in the proof of lemma 1.9.20, there exists $\tilde{\varphi}$ such that $L_{2,P}(\varphi_n, \tilde{\varphi}) \to 0$, and a subsequence $\{\varphi_{n_k}\}_{k=1}^\infty$ such that $\lim_{k\to\infty} \varphi_{n_k}(y) = \tilde{\varphi}(y)$ for P-almost every y. Let $N \subset \mathcal{Y}$ be the P-negligible set where this convergence fails.

   Let $y \in N^c$, and notice that $\varphi_{n_k}(y) = \mathbb{1}_{I_{n_k}}(y) \in \{0, 1\}$ for all k and $\{\varphi_{n_k}(y)\}_{k=1}^\infty$ converging in $\mathbb{R}$ implies that $\varphi_{n_k}(y)$ is eventually constant as k grows. This implies $\tilde{\varphi}(y) \in \{0, 1\}$, and hence for some set $A \subset \mathcal{Y}$,

   $$\tilde{\varphi}(y) = \mathbb{1}_A(y) \qquad \text{for all } y \in N^c$$

   We will show that for some interval I, $A \cap N^c = I \cap N^c$. Let $y_1, y_2, y_3 \in N^c$ satisfy $y_1 < y_2 < y_3$ and $y_1, y_3 \in A$, but be otherwise arbitrary. It suffices to show that $y_2 \in A$; we can then define I to be the interval with endpoints $\inf A$ and $\sup A$ (including the lower endpoint if $\inf A = \min A > -\infty$, and including the upper endpoint if $\sup A = \max A < \infty$), and define the function $\varphi : \mathcal{Y}_1 \to \mathbb{R}$ with $\varphi(y_1) = \mathbb{1}_I(y_1)$.[12]

[12]Explicitly, I is defined as follows: (a) $I = (\ell, u)$ if neither $\ell = \inf A$ nor $u = \sup A$ is attained in $\mathbb{R}$ (b) $I = [\ell, u)$ if $\ell = \inf A = \min A$, but $u = \sup A$ is not attained in $\mathbb{R}$ (c) $I = (\ell, u]$ if $\ell = \inf A$ is not

Notice that $\lim_{k\to\infty} \mathbb{1}_{I_{n_k}}(y_3) = \mathbb{1}_A(y_3) = 1$ and $\lim_{k\to\infty} \mathbb{1}_{I_{n_k}}(y_3) = \mathbb{1}_A(y_3) = 1$ implies that $\mathbb{1}_{I_{n_k}}(y_1)$ and $\mathbb{1}_{I_{n_k}}(y_3)$ are eventually constant and equal to 1, i.e. there exists $K_1, K_3 \in \mathbb{N}$ such that

$$y_1 \in I_{n_k} \text{ for all } k \geq K_1, \text{ and} \qquad\qquad y_3 \in I_{n_k} \text{ for all } k \geq K_3$$

Since I_{n_k} is an interval, this implies

$$y_2 \in I_{n_k} \text{ for all } k \geq \max\{K_1, K_3\}$$

i.e. $\mathbb{1}_{I_{n_k}}(y_2) = 1$ for all such k, and therefore $\mathbb{1}_A(y_2) = \lim_{k\to\infty} \mathbb{1}_{A_n}(y_2) = 1$. Thus $y_2 \in A$.

It follows that $\tilde{\varphi}(y) = \varphi(y) = \mathbb{1}_I(y)$ for all $y \in N^c$. Thus $L_{2,P}(\tilde{\varphi}, \varphi) = 0$, and $L_{2,P}(\varphi_n, \varphi) \to 0$. Since $\varphi \in \mathcal{F}_c$, this completes the proof that $(\mathcal{F}_c, L_{2,P})$ is complete.

2. $(\mathcal{F}_c^c, L_{2,P})$ is complete.

   The argument is similar. Let $\{\psi_n\}_{n=1}^{\infty} \subseteq \mathcal{F}_c$ be $L_{2,P}$-Cauchy. Note that $\psi_n(y) = \mathbb{1}_{I_n^c}(y)$ for some interval I_n. There exists $\tilde{\psi}$ such that $L_{2,P}(\psi_n, \tilde{\psi}) \to 0$, and a subsequence $\{\psi_{n_k}\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty} \psi_{n_k}(y) = \tilde{\psi}(y)$ for P-almost every y. Let $N \subset \mathcal{Y}$ be the P-negligible set where this convergence fails.

   Since $\psi_{n_k}(y) = \mathbb{1}_{I_{n_k}^c}(y) \in \{0, 1\}$ for all k and y, and $\lim_{k\to\infty} \psi_{n_k}(y) = \tilde{\psi}(y)$ for all $y \in N^c$, we have $\tilde{\psi}(y) \in \{0, 1\}$ for all such y and thus for some set $A \subseteq \mathcal{Y}$,

$$\tilde{\psi}(y) = \mathbb{1}_{A^c}(y) \qquad\qquad \text{for all } y \in N^c$$

   Once again, it suffices to show $A \cap N^c = I \cap N^c$ for some interval I. Consider $y_1, y_2, y_3 \in N^c$, $y_1 < y_2 < y_3$, with $y_1, y_3 \in A$. $\lim_{k\to\infty} \psi_{n_k}(y_1) = \tilde{\psi}(y_1) = 0$ and $\lim_{k\to\infty} \psi_{n_k}(y_3) =$

attained in $\mathbb{R}$, but $u = \sup A = \max A$ (d) $I = [\ell, u]$ if both $\ell = \inf A = \min A$ and $u = \sup A = \max A$.

$\tilde{\psi}(y_3) = 0$ implies that $\psi_{n_k}(y_1) = \mathbb{1}_{I^c_{n_k}}(y_1)$ and $\psi_{n_k}(y_3) = \mathbb{1}_{I^c_{n_k}}(y_3)$ are eventually constant and equal to 0, i.e. for some $K_1, K_3 \in \mathbb{N}$,

$$y_1 \in I_{n_k} \text{ for all } k \geq K_1, \qquad\qquad y_3 \in I_{n_k} \text{ for all } k \geq K_3$$

since I_{n_k} is an interval for every k, this implies

$$y_2 \in I_{n_k} \text{ for all } k \geq \max\{K_1, K_3\}$$

thus $\tilde{\psi}(y_2) = \lim_{k\to\infty} \psi_{n_k}(y_2) = 0$. It follows that $A \cap N^c = I \cap N^c$, where I is the interval defined by endpoints $\inf A$ and $\sup A$, which are included if attained and finite. Define $\psi(y) = \mathbb{1}_{I^c}(y)$ and notice $\psi \in \mathcal{F}^c_c$. We have $\psi(y) = \tilde{\psi}(y)$ for all $y \in N^c$ and hence $L_{2,P}(\tilde{\psi}, \psi) = 0$. Thus $L_{2,P}(\psi_n, \psi) \to 0$, showing $(\mathcal{F}^c_c, L_{2,P})$ is complete.

3. Note that $(\mathcal{F}_c \times \mathcal{F}^c_c, L_2)$ is the product space of the complete spaces $(\mathcal{F}_c, L_{2,P})$ and $(\mathcal{F}^c_c, L_{2,P})$, and so is complete.

4. We next show $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}^c_c) = \{(\varphi, \psi) \in \mathcal{F}_c \times \mathcal{F}^c_c \,;\, \varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)\}$ is complete.

   Let $\{(\varphi_n, \psi_n)\}_{n=1}^{\infty} \subseteq \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}^c_c)$ be L_2-Cauchy, and let $(\tilde{\varphi}, \tilde{\psi})$ be a limit in $\mathcal{F}_c \times \mathcal{F}^c_c$. Since $L_{2,P}(\varphi_n, \tilde{\varphi}) \to 0$ there exists a subsequence $\{(\varphi_{n_k}, \psi_{n_k})\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty} \varphi_{n_k}(y_1) = \tilde{\varphi}(y_1)$ for P-almost all y_1. Let N_1 be the negligible set where this fails. Furthermore, $L_{2,P}(\psi_{n_k}, \tilde{\psi}) \to 0$ as $k \to \infty$ and so there is a further subsequence $\{(\varphi_{n_{k_j}}, \psi_{n_{k_j}})\}_{j=1}^{\infty}$ such that $\lim_{j\to\infty} \psi_{n_{k_j}}(y_0) = \tilde{\psi}(y_0)$ for P-almost all y_0. Let N_0 be the negligible set where this fails. It is then clear that if $(y_1, y_0) \in N_1^c \times N_0^c$, then

$$\tilde{\varphi}(y_1) + \tilde{\psi}(y_0) = \lim_{j\to\infty} \{\varphi_{n_{k_j}}(y_1) + \psi_{n_{k_j}}(y_0)\} \leq \lim_{j\to\infty} c(y_1, y_0) = \mathbb{1}_C(y_1, y_0) \tag{1.68}$$

Note that $\tilde{\varphi} = \mathbb{1}_{I_{\tilde{\varphi}}}$, and $\tilde{\psi} = -\mathbb{1}_{I^c_{\tilde{\psi}}}$ for some intervals $I_{\tilde{\varphi}}$ and $I_{\tilde{\psi}}$. Let

$$\ell_1 = \inf I_{\tilde{\varphi}} \cap N_1^c, \qquad u_1 = \sup I_{\tilde{\varphi}} \cap N_1^c, \qquad \ell_0 = \inf I_{\tilde{\psi}} \cap N_0^c, \qquad u_0 = \sup I_{\tilde{\psi}} \cap N_0^c$$

and define $\varphi = \mathbb{1}_{I_\varphi}$ where I_φ is the interval with endpoints ℓ_1, u_1 (included if the inf/sup are finite and attained), and $\psi = -\mathbb{1}_{I^c_\psi}$ where I^c_ψ is the interval with endpoints ℓ_0, u_0 (included if the inf/sup are finite and attained). Notice that $I_\varphi = I_{\tilde{\varphi}}$, P-almost surely and $I_\psi = I_{\tilde{\psi}}$, P-almost surely.

Notice that for $(y_1, y_0) \in (N_1^c \times N_0^c)^c$ to satisfy $\varphi(y_1) + \psi(y_0) = \mathbb{1}_{I_\varphi}(y_1) - \mathbb{1}_{I^c_\psi}(y_0) > \mathbb{1}_C(y_1, y_0)$, it would have to be the case that $(y_1, y_0) \in (I_{\tilde{\varphi}} \times I_{\tilde{\psi}}) \cap (N_1^c \times N_0^c)^c \setminus C$. Let $(y_1, y_0) \in (I_\varphi \times I_\psi) \cap (N_1^c \times N_0^c)^c$, and note that there exists $y_1^\ell, y_1^u \in I_\varphi \cap N_1^c$ with $y_1^\ell \leq y_1 \leq y_1^u$ and $y_0^\ell, y_0^u \in I_\psi \cap N_0^c$ with $y_0^\ell \leq y_0 \leq y_0^u$. Notice that $[y_1^\ell, y_1^u] \times [y_0^\ell, y_0^u] \subseteq C$, because C is convex and (1.68) holds for the "corners": $(\ell_1, \ell_0), (\ell_1, u_0), (u_1, \ell_0), (u_1, u_0) \in (I_\varphi \times I_\psi) \cap (N_1^c \times N_0^c)$. Thus $(I_{\tilde{\varphi}} \times I_{\tilde{\psi}}) \cap (N_1^c \times N_0^c)^c \setminus C = \varnothing$, showing that $\varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)$ holds for all $(y_1, y_0) \in \mathcal{Y}_1 \times \mathcal{Y}_0$. This shows $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ is complete.

5. The argument thet $(\mathcal{F}_{1,x} \times \mathcal{F}_{0,x}, L_2)$ is complete is identical to the argument given in step 5 of the proof of lemma 1.9.20.

This completes the proof. □

1.9.3.4 Differentiability of T_2

We first apply lemma 1.9.7 to show show that $\theta^L(\cdot, \cdot)$ and $\theta^H(\cdot, \cdot)$, given by either (1.19) or (1.20) depending on the function c, are Hadamard differentiable.

Lemma 1.9.22. *Suppose assumptions 1, 2, and 3 hold. Then $\theta^L(\cdot, \cdot)$ and $\theta^H(\cdot, \cdot)$ given by* (1.19) *or* (1.20) *are Hadamard directionally differentiable at $(P_{1|x}, P_{0|x})$ tangentially to*

$\mathcal{C}(\mathcal{F}_{1,x}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x}, L_{2,P})$. *The argmax sets*

$$\Psi_{c_L}(P_{1|x}, P_{0|x}) = \underset{(\varphi,\psi)\in\Phi_{c_L}\cap(\mathcal{F}_c\times\mathcal{F}_c^c)}{\arg\max} P_{1|x}(\varphi) + P_{0|x}(\psi)$$

$$\Psi_{c_H}(P_{1|x}, P_{0|x}) = \underset{(\varphi,\psi)\in\Phi_{c_H}\cap(\mathcal{F}_c\times\mathcal{F}_c^c)}{\arg\max} P_{1|x}(\varphi) + P_{0|x}(\psi)$$

are nonempty, and the derivatives $\theta^{L\prime}_{(P_{1|x},P_{0|x})}, \theta^{H\prime}_{(P_{1|x},P_{0|x})} : \mathcal{C}(\mathcal{F}_{1,x}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x}, L_{2,P}) \to \mathbb{R}$ *are given by*

$$\theta^{L\prime}_{(P_{1|x},P_{0|x})}(H_1, H_0) = \sup_{(\varphi,\psi)\in\Psi_{c_L}(P_{1|x},P_{0|x})} H_1(\varphi) + H_0(\psi) \tag{1.69}$$

$$\theta^{H\prime}_{(P_{1|x},P_{0|x})}(H_1, H_0) = -\left[\sup_{(\varphi,\psi)\in\Psi_{c_H}(P_{1|x},P_{0|x})} H_1(\varphi) + H_0(\psi)\right] \tag{1.70}$$

If assumption 4 also holds, then θ^L *and* θ^H *are fully Hadamard differentiable at* $(P_{1|x}, P_{0|x})$ *tangentially to*

$$\mathbb{D}_{Tan,Full,x} = \left(\ell^\infty_{\mathcal{Y}_{1,x}}(\mathcal{F}_{1,x}) \times \ell^\infty_{\mathcal{Y}_{0,x}}(\mathcal{F}_{0,x})\right) \cap \left(\mathcal{C}(\mathcal{F}_{1,x}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x}, L_{2,P})\right)$$

with the derivatives $\theta^{L\prime}_{(P_{1|x},P_{0|x})}, \theta^{H\prime}_{(P_{1|x},P_{0|x})} : \mathbb{D}_{Tan,Full,x} \to \mathbb{R}$ *also given by* (1.69) *and* (1.70).

Proof. We apply lemma 1.9.7. It is clear from inspection that the cost functions c_L and c_H are lower semicontinuous, the sets $\mathcal{F}_{d,x}$ defined by (1.57) consists of measurable functions mapping $\mathcal{Y}$ to $\mathbb{R}$, and that the subsets $\mathcal{F}_c$ and $\mathcal{F}_c^c$ given by (1.14) and (1.15), or by (1.16) and (1.17), are universally bounded. Moreover,

1. Strong duality holds.

   (i) If assumption 2 (i) holds, then lemma 1.9.38 shows that strong duality holds.

   (ii) If assumption 2 (ii) holds, then lemma 1.9.42 shows that strong duality holds.

2. Assumption 1 implies P dominates $P_{d|x}$ with bounded densities $\frac{dP_{d|x}}{dP}$. Indeed,

$$
\begin{aligned}
&E_{P_{d|x}}[f(Y_d)] \\
&= \frac{E_P[f(Y)\mathbb{1}\{D=d\} \mid X=x, Z=d] - E_P[f(Y)\mathbb{1}\{D=d\} \mid X=x, Z=1-d]}{P(D=d \mid X=x, Z=d) - P(D=d \mid X=x, Z=1-d)} \\
&= E_P\left[f(Y)\frac{\mathbb{1}_{d,x,d}(D,X,Z)/p_{x,d} - \mathbb{1}_{d,x,1-d}(D,X,Z)/p_{x,1-d}}{p_{d,x,d}/p_{x,d} - p_{d,x,1-d}/p_{x,1-d}}\right] \\
&= E_P\left[f(Y)E\left[\frac{\mathbb{1}_{d,x,d}(D,X,Z)/p_{x,d} - \mathbb{1}_{d,x,1-d}(D,X,Z)/p_{x,1-d}}{p_{d,x,d}/p_{x,d} - p_{d,x,1-d}/p_{x,1-d}} \mid Y\right]\right]
\end{aligned}
$$

   Notice that $\frac{dP_{d|x}}{dP}(Y) = E_P\left[\frac{\mathbb{1}_{d,x,d}(D,X,Z)/p_{x,d} - \mathbb{1}_{d,x,1-d}(D,X,Z)/p_{x,1-d}}{p_{d,x,d}/p_{x,d} - p_{d,x,1-d}/p_{x,1-d}} \mid Y\right]$ must be nonnegative P-almost surely; if the set $A = \left\{y\,;\, \frac{dP_{d|x}}{dP}(y) < 0\right\}$ was P-non-negligible, the displays above would imply the contradiction $P(Y_d \in A \mid D_1 > D_0, X = x) < 0$. Moreover, $\frac{dP_{d|x}}{dP}$ is bounded because the integrand of this conditional mean is bounded.

3. Lemma 1.9.14 shows that under assumptions 1, 2, and 3, $\mathcal{F}_{d,x}$ is P-Donsker and $\sup_{f\in\mathcal{F}_{d,x}}|P(f)| < \infty$ for $d = 1, 0$, and

4. The set $(\mathcal{F}_1 \times \mathcal{F}_0, L_2)$ and its subset $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ are complete.

   (i) If assumption 2 (i) holds, then lemma 1.9.20 shows these sets are complete.

   (ii) If assumption 2 (ii) holds, then lemma 1.9.21 shows these sets are complete.

It follows from the chain rule that θ^L and θ^H are Hadamard directionally differentiable with the claimed directional derivatives.

Now suppose assumptions 1, 2, 3, and 4 hold. Lemma 1.9.10 implies θ^L and θ^H are fully Hadamard differentiable at $(P_{1|x}, P_{0|x})$ tangentially to

$$
\mathbb{D}_{T,Full,x} = \left(\ell^\infty_{\mathcal{Y}_{1,x}}(\mathcal{F}_{1,x}) \times \ell^\infty_{\mathcal{Y}_{0,x}}(\mathcal{F}_{0,x}\right) \cap \left(\mathcal{C}(\mathcal{F}_{1,x}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x}, L_{2,P})\right)
$$

with derivatives given by the same expressions. □

We can now show the differentiability properties of T_2.

Lemma 1.9.23 (T_2 is Hadamard differentiable)**.** *Let $\mathbb{D}_{Tan}$ and $\mathbb{D}_{Tan,Full}$ be given by*

$$\mathbb{D}_{Tan} = \prod_{m=1}^{M} \mathcal{C}(\mathcal{F}_{1,x_m}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x_m}, L_{2,P}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$$

$$\mathbb{D}_{Tan,Full} = \prod_{m=1}^{M} \left(\ell^{\infty}_{\mathcal{Y}_{1,x_m}}(\mathcal{F}_{1,x_m}) \times \ell^{\infty}_{\mathcal{Y}_{0,x_m}}(\mathcal{F}_{0,x_m})\right) \cap \left(\mathcal{C}(\mathcal{F}_{1,x_m}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x_m}, L_{2,P})\right) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$$

and define

$$T_2 : \prod_{m=1}^{M} \ell^{\infty}(\mathcal{F}_{1,x}) \times \ell^{\infty}(\mathcal{F}_{0,x}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R} \to \prod_{m=1}^{M} \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R},$$

$$T_2(\{P_{1|x}, P_{0|x}, \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}}) = \left(\{\theta^L(P_{1|x}, P_{0|x}), \theta^H(P_{1|x}, P_{0|x}), \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}}\right)$$

Under assumptions 1, 2, and 3, T_2 is Hadamard directionally differentiable at $T_1(P) = (\{P_{1|x}, P_{0|x}, s_x, \eta_{1,x}, \eta_{0,x}\}_{x\in\mathcal{X}})$ tangentially to $\mathbb{D}_{Tan}$, with derivative

$$T'_{2,T_1(P)} : \mathbb{D}_{Tan} \to \prod_{m=1}^{M} \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$$

$$T'_{2,T_1(P)}\left(\{H_{1,x}, H_{0,x}, h_{\eta_1,x}, h_{\eta_0,x}, h_{s,x}\}_{x\in\mathcal{X}}\right) = \left(\left\{\theta^{L\prime}_{(P_{1|x},P_{0|x})}(H_{1,x}, H_{0,x}), \theta^{H\prime}_{(P_{1|x},P_{0|x})}(H_{1,x}, H_{0,x}), h_{\eta_1,x}, h_{\eta_0,x}, h_{s,x}\right\}_{x\in\mathcal{X}}\right)$$

If assumption 4 also holds, then T_2 is fully Hadamard differentiable at $T_1(P)$ tangentially to $\mathbb{D}_{Tan,Full}$, with derivative $T_{2,T_1(P)} : \mathbb{D}_{Tan,Full} \to \prod_{m=1}^{M} \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$ given by the same expression.

Proof. Lemma 1.9.22 shows that under assumptions 1, 2, and 3, $\theta^L(\cdot)$ and $\theta^H(\cdot)$ are Hadamard directionally differentiable at $(P_{1|x}, P_{0|x})$ tangentially to $\mathcal{C}(\mathcal{F}_{1,x}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x}, L_{2,P})$ for each

$x \in \mathcal{X}$. If assumption 4 also holds, lemma 1.9.22 shows these derivatives are linear on the subspace $\mathbb{D}_{Tan,Full}$, and hence $\theta^L(\cdot)$ and $\theta^H(\cdot)$ are fully Hadamard differentiable tangentially to $\mathbb{D}_{Tan,Full}$. The other coordinates are the identity mapping, which is fully Hadamard differentiable. Apply lemma 1.9.51 to obtain the result. □

1.9.3.5 Expectations, $T_3(\{\theta_x^L, \theta_x^H, \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}}) = (\theta^L, \theta^H, \eta)$

Lemma 1.9.24. *Define*

$$T_3 : \prod_{m=1}^{M} \mathbb{R}\times\mathbb{R}\times\mathbb{R}^{K_1}\times\mathbb{R}^{K_0}\times\mathbb{R} \to \mathbb{R}\times\mathbb{R}\times\mathbb{R}^{K_1}\times\mathbb{R}^{K_0}$$

$$T_3(\{\theta_x^L, \theta_x^H, \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}}) = \left(\sum_{x\in\mathcal{X}} s_x\theta_x^L, \sum_{x\in\mathcal{X}} s_x\theta_x^H, \sum_{x\in\mathcal{X}} s_x\eta_{1,x}, \sum_{x\in\mathcal{X}} s_x\eta_{0,x}\right)$$

T_3 is fully (Hadamard) differentiable at any $V = (\{\theta_x^L, \theta_x^H, \eta_{1,x}, \eta_{0,x}, s_x\}_{x\in\mathcal{X}}) \in \prod_{m=1}^{M} \mathbb{R}\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}^{K_1}\times\mathbb{R}^{K_0}$ tangentially to $\prod_{m=1}^{M} \mathbb{R}\times\mathbb{R}\times\mathbb{R}^{K_1}\times\mathbb{R}^{K_0}\times\mathbb{R}$ with derivative

$$T'_{3,V} : \prod_{m=1}^{M} \mathbb{R}\times\mathbb{R}\times\mathbb{R}^{K_1}\times\mathbb{R}^{K_0}\times\mathbb{R} \to \mathbb{R}\times\mathbb{R}\times\mathbb{R}^{K_1}\times\mathbb{R}^{K_0}$$

$$T'_{3,V}(\{h_x^L, h_x^H, h_{\eta_1,x}, h_{\eta_0,x}, h_{s,x}\}_{x\in\mathcal{X}})$$
$$= \left(\sum_{x\in\mathcal{X}} s_x h_x^L + h_{s,x}\theta^L(x), \sum_{x\in\mathcal{X}} s_x h_x^H + h_{s,x}\theta^H(x), \sum_{x\in\mathcal{X}} s_x h_{\eta_1,x} + h_{s,x}\eta_{1,x}, \sum_{x\in\mathcal{X}} s_x h_{\eta_0,x} + h_{s,x}\eta_{0,x}\right)$$

Proof. The inner product

$$IP : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}, \qquad IP(r_1, r_2) = \langle r_1, r_2\rangle = \sum_{m=1}^{M} r_1^{(m)} r_2^{(m)}$$

is fully Hadamard differentiable at any $(r_1, r_2) \in \mathbb{R}^M \times \mathbb{R}^M$ tangentially to $\mathbb{R}^M \times \mathbb{R}^M$ with

derivative

$$IP'_{(r_1,r_2)} : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R},$$
$$IP'_{(r_1,r_2)}(h_1, h_2) = \langle r_1, h_2 \rangle + \langle h_1, r_2 \rangle = \sum_{m=1}^{M} r_1^{(m)} h_2^{(m)} + h_1^{(m)} r_2^{(m)}$$

Apply lemma 1.9.51 to obtain the result. □

1.9.3.6 Optimization over $t \in [\theta^L, \theta^H]$: $T_4(\theta^L, \theta^H, \eta) = (\gamma^L, \gamma^H)$

Lemma 1.9.25. *Let $g^L, g^H : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \to \mathbb{R}$ be as defined in assumption 3:*

$$g^L(\theta^L, \theta^H, \eta_1, \eta_0) = \inf_{t \in [\theta^L, \theta^H]} g(t, \eta_1, \eta_0), \qquad g^H(\theta^L, \theta^H, \eta_1, \eta_0) = \sup_{t \in [\theta^L, \theta^H]} g(t, \eta_1, \eta_0)$$

Define

$$T_4 : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \to \mathbb{R} \times \mathbb{R}$$
$$T_4(\theta^L, \theta^H, \eta_1, \eta_0) = \big(g^L(\theta^L, \theta^H, \eta_1, \eta_0), g^H(\theta^L, \theta^H, \eta_1, \eta_0)\big)$$

Under assumption 3, g^L and g^H are continuously differentiable at $(\theta^L, \theta^H, \eta_1, \eta_0) = T_3(T_2(T_1(P)))$ with gradients

$$\nabla g^L = \nabla g^L(\theta^L, \theta^H, \eta_1, \eta_0) \in \mathbb{R}^{2+K_1+K_0}, \qquad \nabla g^H = \nabla g^H(\theta^L, \theta^H, \eta_1, \eta_0) \in \mathbb{R}^{2+K_1+K_0}$$

Therefore T_4 is fully Hadamard differentiable at $(\theta^L, \theta^H, \eta_1, \eta_0)$ tangentially to $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^{K_1} \times$

$\mathbb{R}^{K_0}$, *with derivative*

$$
\begin{aligned}
&T'_{4,T_3(T_2(T_1(P)))} : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \to \mathbb{R} \times \mathbb{R} \\
&T'_{4,T_3(T_2(T_1(P)))}(h^L, h^H, h_{\eta_1}, h_{\eta_0}) \\
&\qquad = \left(\left\langle \nabla g^L, (h^L, h^H, h_{\eta_1}, h_{\eta_0}) \right\rangle, \left\langle \nabla g^H, (h^L, h^H, h_{\eta_1}, h_{\eta_0}) \right\rangle\right)
\end{aligned}
$$

Proof. Assumption 3 (iii) is that g^L and g^H are continuously differentiable. The result follows. □

1.9.3.7 The map $T(P) = (\gamma^L, \gamma^H)$, consistency, and weak convergence

Lemma 1.9.26. *Let T_1, T_2, T_3, and T_4 be as defined in lemmas 1.9.17, 1.9.23, 1.9.24, and 1.9.25 respectively. Let*

$$
\begin{aligned}
\left(\left\{\hat{P}_{1|x}, \hat{P}_{0|x}, \hat{\eta}_{1,x}, \hat{\eta}_{0,x}, \hat{s}_x\right\}_{x\in\mathcal{X}}\right) &= T_1(\mathbb{P}_n) \\
\left(\{\hat{\theta}^L_x, \hat{\theta}^H_x, \hat{\eta}_{1,x}, \hat{\eta}_{0,x}, \hat{s}_x\}_{x\in\mathcal{X}}\right) &= T_2(T_1(\mathbb{P}_n)) \\
(\hat{\theta}^L, \hat{\theta}^H, \hat{\eta}) &= T_3(T_2(T_1(\mathbb{P}_n))), \\
(\hat{\gamma}^L, \hat{\gamma}^H) &= T_4(T_3(T_2(T_1(\mathbb{P}_n))))
\end{aligned}
$$

be the empirical analogue estimators. If assumptions 1, 2, and 3 hold, then each of these estimators are consistent.

Proof. Lemmas 1.9.17, 1.9.23, 1.9.24, and 1.9.25 show that T_1, T_2, T_3, and T_4 are Hadamard (directionally) differentiable at P, $T_1(P)$, $T_2(T_1(P))$, and $T_3(T_2(T_1(P)))$ respectively, tangentially to sets that include zero. It follows that these functions are continuous at P, $T_1(P)$, $T_2(T_1(P))$, and $T_3(T_2(T_1(P)))$ respectively. Lemma 1.9.15 implies that $\mathbb{P}_n \xrightarrow{p} P$ in $\ell^\infty(\mathcal{F})$,

so it follows from the continuous mapping theorem that

$$\begin{aligned} T_1(\mathbb{P}_n) &\xrightarrow{p} T_1(P), \\ T_2(T_1(\mathbb{P}_n)) &\xrightarrow{p} T_2(T_1(P)), \\ T_3(T_2(T_1(\mathbb{P}_n))) &\xrightarrow{p} T_3(T_2(T_1(P))), \text{ and} \\ T_4(T_3(T_2(T_1(\mathbb{P}_n)))) &\xrightarrow{p} T_4(T_3(T_2(T_1(P)))). \end{aligned}$$

□

Lemma 1.9.27 (T is Hadamard directionally differentiable)**.** *Let $\mathbb{D}_C$ be defined by* (1.60)*, and*

$$T : \mathbb{D}_C \to \mathbb{R}^2, \qquad T(G) = T_4(T_3(T_2(T_1(G))))$$

If assumptions 1, 2, 3 holds, then T is Hadamard directionally differentiable at P tangentially to $\mathcal{C}(\mathcal{F}, L_{2,P})$ with derivative given by

$$T'_P : \mathcal{C}(\mathcal{F}, L_{2,P}) \to \mathbb{R}^2, \qquad T'_P(G) = T'_{4,T_3(T_2(T_1(P)))}(T'_{3,T_2(T_1(P))}(T'_{2,T_1(P)}(T'_{1,P}(G))))$$

If assumption 4 also holds, then T is fully Hadamard differentiable at P tangentially to the support of $\mathbb{G}$ as defined in lemma 1.9.15.

Proof. Lemma 1.9.17 shows that T_1 is fully Hadamard differentiable at any point in $\mathbb{D}_C$ tangentially to $\ell^\infty(\mathcal{F})$. Lemma 1.9.23 shows that under assumptions 1, 2, and 3, T_2 is Hadamard directionally differentiable at $T_1(P)$ tangentially to

$$\mathbb{D}_{Tan} = \prod_{m=1}^{M} \mathcal{C}(\mathcal{F}_{1,x_m}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x_m}, L_{2,P}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$$

Lemma 1.9.18 implies that if $H \in \mathcal{C}(\mathcal{F}, L_{2,P})$, then $T'_{1,P}(H) \in \mathbb{D}_{Tan}$. It follows from the chain

rule (lemma 1.9.50) that $T_2 \circ T_1$ is Hadamard directionally differentiable at P tangentially to $\mathcal{C}(\mathcal{F}, L_{2,P})$. Lemma 1.9.24 shows T_3 is fully differentiable at any point in its domain tangentially to the entire relevant space, and lemma 1.9.25 shows T_4 is fully differentiable at $T_3(T_2(T_1(P)))$ tangentially to the entire relevant space. The chain rule thus implies the first claim: under assumptions 1, 2, and 3, $T = T_4 \circ T_3 \circ T_2 \circ T_1$ is Hadamard directionally differentiable at P tangentially to $\mathcal{C}(\mathcal{F}, L_{2,P})$ with the claimed derivative.

If assumption 4 also holds, lemma 1.9.23 implies that T_2 is fully differentiable at $T_1(P)$ tangentially to $\mathbb{D}_{Tan,Full}$. Lemma 1.9.19 shows the support of $T'_{1,P}(\mathbb{G})$ is contained within $\mathbb{D}_{Tan,Full}$. It follows that $T'_P(\cdot) = T'_{4,T_3(T_2(T_1(P)))}(T'_{3,T_2(T_1(P))}(T'_{2,T_1(P)}(T'_{1,P}(\cdot))))$ is linear on the support of $\mathbb{G}$, and hence Fang & Santos (2019) proposition 2.1 implies T is fully Hadamard differentiable at P tangentially to the support of $\mathbb{G}$. □

Lemma 1.5.1. *Suppose that*

(i) assumption 2 (i) holds, with cost function $c(y_1, y_0)$ that is continuously differentiable, and

(ii) for each (d, x), the support of $P_{d|x}$ is $\mathcal{Y}_{d,x}$, which is a bounded interval.

Then assumption 4 holds.

Proof. Note that both $c_L(y_1, y_0) = c(y_1, y_0)$ and $c_H(y_1, y_0) = -c(y_1, y_0)$ are continuously differentiable. Moreover, since the support of $P_{d|x}$ is $\mathcal{Y}_{d,x}$ which is a bounded interval, the support can be written as $[y^\ell_{d,x}, y^u_{d,x}]$. So for any $x \in \mathcal{X}$ and either $c \in \{c_L, c_H\}$, lemma 1.9.8 shows that for any $(\varphi_1, \psi_1), (\varphi_2, \psi_2) \in \Psi_c(P_{1|x}, P_{0|x})$, there exists $s \in \mathbb{R}$ such that for all $(y_1, y_0) \in \mathcal{Y}_{1,x} \times \mathcal{Y}_{0,x}$

$$\varphi_1(y_1) - \varphi_2(y_1) = s, \qquad \psi_1(y_0) - \psi_2(y_0) = -s$$

and thus

$$\mathbb{1}_{\mathcal{Y}_{1,x}} \times \varphi_1 = \mathbb{1}_{\mathcal{Y}_{1,x}} \times (\varphi_2 + s),\ P\text{-a.s.} \qquad \text{and} \qquad \mathbb{1}_{\mathcal{Y}_{0,x}} \times \psi_1 = \mathbb{1}_{\mathcal{Y}_{0,x}} \times (\psi_2 - s),\ P\text{-a.s.}.$$

Therefore assumption 4 holds. □

Theorem 1.5.2 (Weak convergence)**.** *Suppose assumptions 1, 2, and 3 hold, and let* $\mathbb{G}$ *be the weak limit of* $\sqrt{n}(\mathbb{P}_n - P)$ *in* $\ell^\infty(\mathcal{F})$*. Then* T *is Hadamard directionally differentiable at* P *tangentially to the support of* $\mathbb{G}$*, and*

$$\sqrt{n}((\hat{\gamma}^L, \hat{\gamma}^H) - (\gamma^L, \gamma^H)) = \sqrt{n}(T(\mathbb{P}_n) - T(P)) \xrightarrow{L} T'_P(\mathbb{G})$$

If assumption 4 also holds, then T'_P *is linear on the support of* $\mathbb{G}$ *and* $T'_P(\mathbb{G})$ *is bivariate normal.*

Proof. The result is an application of the functional delta method (see Fang & Santos (2019) theorem 2.1) and lemma 1.9.27.

Indeed, $\ell^\infty(\mathcal{F})$ and $\mathbb{R}^2$ are Banach spaces, and under assumptions 1, 2, and 3 lemma 1.9.27 shows T is Hadamard directionally differentiable at P tangentially to $\mathcal{C}(\mathcal{F}, L_{2,P})$. Lemma 1.9.15 shows that $\sqrt{n}(\mathbb{P}_n - P) \xrightarrow{L} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$, where $\mathbb{G}$ is tight and supported in $\mathcal{C}(\mathcal{F}, L_{2,P})$. Fang & Santos (2019) theorem 2.1 gives the result that $\sqrt{n}(T(\mathbb{P}_n) - T(P)) \xrightarrow{L} T'_P(\mathbb{G})$.

If assumption 4 holds as well as assumptions 1, 2, and 3, then lemma 1.9.27 shows that T is fully differentiable on the support of $\mathbb{G}$. Since $\mathbb{G}$ is Gaussian and T'_P is continuous and linear on the support of $\mathbb{G}$, $T'_P(\mathbb{G}) \in \mathbb{R}^2$ is Gaussian. □

1.9.4 Appendix: inference

1.9.4.1 Bootstrap

Lemma 1.9.28. *Suppose assumptions 1, 2, and 3 are satisfied. Let $\mathbb{P}_n^*$ be given by definition 1.5.1 or 1.5.2. Then Fang & Santos (2019) assumption 3 is satisfied:*

(i) $\mathbb{P}_n^*$ *is a function of* $\{Y_i, D_i, Z_i, X_i, W_i\}_{i=1}^n$, *with* $\{W_i\}_{i=1}^n$ *independent of* $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$.

(ii) $\mathbb{P}_n^*$ *satisfies* $\sup_{f \in BL_1} \left| E\left[f(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)) \mid \{Y_i, D_i, Z_i, X_i\}_{i=1}^n\right] - E[f(\mathbb{G})]\right| = o_p(1)$.

(iii) $\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$ *is asymptotically measurable (jointly in* $\{Y_i, D_i, Z_i, X_i, W_i\}_{i=1}^n$*).*

(iv) $f(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n))$ *is a measurable function of* $\{W_i\}_{i=1}^n$ *outer almost surely in* $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$ *for any continuous and bounded real-valued* f.

Proof. Note that assumption 3(i) is satisfied by construction. van der Vaart & Wellner (1997) example 3.6.9, 3.6.10, and theorem 3.6.13 implies assumpion 3(ii) holds:

$$\sup_{f \in \mathrm{BL}_1} \left| E\left[f(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)) \mid \{Y_i, D_i, Z_i, X_i\}_{i=1}^n\right] - E[f(\mathbb{G})]\right| \xrightarrow{P^*} 0$$

and further that

$$E\left[f(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n))^*\right] - E\left[f(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n))_*\right] = o_p(1)$$

for any $f \in \mathrm{BL}_1$, where $f(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n))^*$ and $f(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n))_*$ denote the minimal measurable majorant and maximal measurable minorant of $f(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n))$, respectively. Note that for any continuous and bounded f, $f(\sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n))$ is continuous in $\{W_i\}_{i=1}^n$, and is hence measurable satisfying Fang & Santos (2019) assumption 3(iv). Fang & Santos (2019) lemma S.3.9 then implies assumption 3(iii) is satisfied as well. □

Theorem 1.5.3. *Suppose assumptions 1, 2, 3, and 4 hold, and let $\mathbb{P}_n^*$ be given by definition 1.5.1 or 1.5.2. Then conditional on $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$,*

$$\sqrt{n}(T(\mathbb{P}_n^*) - T(\mathbb{P}_n)) \xrightarrow{L} T'_P(\mathbb{G})$$

in outer probability.

Proof. By application of Fang & Santos (2019) theorem 3.1. There are three numbered assumptions:

1. Fang & Santos (2019) assumption 1 is satisfied; $\ell^\infty(\mathcal{F})$ and $\mathbb{R}^2$ are indeed Banach spaces, and lemma 1.9.27 shows that under this paper's assumptions 1, 2, and 3, the map T is Hadamard directionally differentiable at P tangentially to $\mathcal{C}(\mathcal{F}, L_{2,P})$.
2. Fang & Santos (2019) assumption 2 is satisfied; lemma 1.9.15 shows that $\sqrt{n}(\mathbb{P}_n - P) \xrightarrow{L} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$, where $\mathbb{G}$ is tight and supported in $\mathcal{C}(\mathcal{F}, L_{2,P})$.
3. Lemma 1.9.28 shows that Fang & Santos (2019) assumption 3 is satisfied.

Finally, note that $\mathbb{G}$ is Gaussian and mean zero; it follows that its support is a vector subspace of $\ell^\infty(\mathcal{F})$. Thus Fang & Santos (2019) theorem 3.1 implies T is (fully) Hadamard differentiable tangentially to the support of $\mathbb{G}$ if and only if

$$\sup_{f \in \mathrm{BL}_1} \left| E\left[f\left(\sqrt{n}(T(\mathbb{P}_n^*) - T(\mathbb{P}_n)) \right) \mid \{Y_i, D_i, Z_i, X_i\}_{i=1}^n \right] - E\left[f(T'_P(\mathbb{G})) \right] \right| = o_p(1)$$

Since lemma 1.9.27 shows that under assumptions 1, 2, 3, and 4, T is fully Hadamard differentiable tangentially to the support of $\mathbb{G}$, this completes the proof. □

1.9.4.2 Alternative for directional differentiability

Lemma 1.9.29. *Let assumptions 1, 2, and 3 hold, and* $\{\kappa_n\}_{n=1}^{\infty} \subseteq \mathbb{R}$ *satisfy* $\kappa_n \to \infty$ *and* $\kappa_n/\sqrt{n} \to 0$. *For* $c \in \{c_L, c_H\}$, *let*

$$\Psi_c(P_{1|x}, P_{0|x}) = \underset{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)}{\arg\max} P_{1|x}(\varphi) + P_{0|x}(\psi)$$
$$\widehat{\Psi}_{c,x} = \left\{(\varphi,\psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)\ ;\ OT_c(\hat{P}_{1|x}, \hat{P}_{0|x}) \leq \hat{P}_{1|x}(\varphi) + \hat{P}_{0|x}(\psi) + \frac{\kappa_n}{\sqrt{n}}\right\}$$

and $OT'_{c,(P_{1|x},P_{0|x})}, \widehat{OT}'_{c,x} : \mathcal{C}(\mathcal{F}_{1,x}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x}, L_{2,P}) \to \mathbb{R}$, *be given by*

$$OT'_{c,(P_{1,|x},P_{0|x})}(H_1, H_0) = \sup_{(\varphi,\psi)\in\Psi_c(P_{1|x},P_{0|x})} H_1(\varphi) + H_0(\psi)$$
$$\widehat{OT}'_{c,x}(H_1, H_0) = \sup_{(\varphi,\psi)\in\widehat{\Psi}_{c,x}} H_1(\varphi) + H_0(\psi)$$

Then for any $(H_1, H_0) \in \mathcal{C}(\mathcal{F}_{1,x}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x}, L_{2,P})$,

$$\left|\widehat{OT}'_{c,x}(H_1, H_0) - OT'_{c,(P_{1,|x},P_{0|x})}(H_1, H_0)\right| \xrightarrow{p} 0$$

Proof. The proof is similar that of Fang & Santos (2019) lemma S.4.8. As the subscript x plays no role, we drop it from the notation.

In steps:

1. We first esteablish an inequality used several times below. Note that for any $(\tilde{\varphi}, \tilde{\psi}), (\varphi, \psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$,

$$\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0} \geq \hat{P}_1(\varphi) - P_1(\varphi) + \hat{P}_0(\psi) - P_0(\psi)$$
$$\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0} \geq P_1(\tilde{\varphi}) - \hat{P}_1(\tilde{\varphi}) + P_0(\tilde{\psi}) - \hat{P}_0(\tilde{\psi})$$

Add these to obtain

$$2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right)$$
$$\geq \hat{P}_1(\varphi) - P_1(\varphi) + \hat{P}_0(\psi) - P_0(\psi) + P_1(\tilde{\varphi}) - \hat{P}_1(\tilde{\varphi}) + P_0(\tilde{\psi}) - \hat{P}_0(\tilde{\psi}), \quad (1.71)$$

2. We next show

$$\lim_{n\to\infty} P\left(\Psi(P_1, P_0) \subseteq \widehat{\Psi}_c\right) = 1 \quad (1.72)$$

Let $(\tilde{\varphi}, \tilde{\psi}) \in \Psi(P_1, P_0)$, and rearrange (1.71) to find

$$2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right)$$
$$\geq \hat{P}_1(\varphi) + \hat{P}_0(\psi) - \hat{P}_1(\tilde{\varphi}) - \hat{P}(\tilde{\psi}) + \underbrace{P_1(\tilde{\varphi}) + P_0(\tilde{\psi}) - P_1(\varphi) - P_0(\psi)}_{\geq 0}$$
$$\geq \hat{P}_1(\varphi) + \hat{P}_0(\psi) - \hat{P}_1(\tilde{\varphi}) - \hat{P}(\tilde{\psi})$$

and therefore

$$\sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \hat{P}_1(\varphi) + \hat{P}_0(\psi) \leq \hat{P}_1(\tilde{\varphi}) + \hat{P}(\tilde{\psi}) + 2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right)$$

holds for any $(\tilde{\varphi}, \tilde{\psi}) \in \Psi_c(P_1, P_0)$. It follows that $2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) < \frac{\kappa_n}{\sqrt{n}}$ implies $(\tilde{\varphi}, \tilde{\psi}) \in \widehat{\Psi}_c$, and hence

$$P\left(2\frac{\sqrt{n}}{\kappa_n}\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) < 1\right) \leq P\left(\Psi(P_1, P_0) \subseteq \widehat{\Psi}_c\right)$$

Lemma 1.9.26 implies $\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0} \xrightarrow{p} 0$. Since $\frac{\sqrt{n}}{\kappa_n} \to 0$, this implies that $2\frac{\sqrt{n}}{\kappa_n}\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) = o_p(1)$ and therefore

$$\lim_{n\to\infty} P\left(\Psi(P_1, P_0) \subseteq \widehat{\Psi}_c\right) \geq \lim_{n\to\infty} P\left(2\frac{\sqrt{n}}{\kappa_n}\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) < 1\right) = 1$$

as was to be shown.

3. We next show that for any $\delta > 0$,

$$\lim_{n\to\infty} P\left(\widehat{\Psi}_c \subseteq (\Psi(P_1, P_0))^{\delta}\right) = 1 \tag{1.73}$$

where $(\Psi(P_1, P_0))^{\delta}$ is an open δ-enlargement of $\Psi(P_1, P_0)$ under L_2; i.e.

$$(\Psi(P_1, P_0))^{\delta} = \left\{(f, g)\ ;\ \inf_{(\varphi,\psi)\in\Psi(P_1,P_0)} L_2((\varphi, \psi), (f, g)) < \delta\right\}$$

Toward this end, note that

$$\eta \equiv \left[\sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \{P_1(\varphi) + P_0(\psi)\} - \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)\setminus(\Psi(P_1,P_0))^{\delta}} \{P_1(\varphi) + P_0(\psi)\}\right] > 0$$

$\eta > 0$ follows from compactness of $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ and continuity of $P_1 + P_0$ with respect to L_2 (see the proof of lemma 1.9.7).

Rearrange (1.71) to find

$$\begin{aligned} &P_1(\tilde{\varphi}) + P_0(\tilde{\psi}) - P_1(\varphi) - P_0(\psi) \\ &\quad \leq 2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) + \hat{P}_1(\tilde{\varphi}) + \hat{P}_0(\tilde{\psi}) - \hat{P}_1(\varphi) - \hat{P}_0(\psi) \end{aligned}$$

Take suprema over $(\tilde{\varphi}, \tilde{\psi}) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ to find

$$\begin{aligned} &\sup_{(\tilde{\varphi},\tilde{\psi})\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} P_1(\tilde{\varphi}) + P_0(\tilde{\psi}) - P_1(\varphi) - P_0(\psi) \\ &\quad \leq 2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) \\ &\qquad + \sup_{(\tilde{\varphi},\tilde{\psi})\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \hat{P}_1(\tilde{\varphi}) + \hat{P}_0(\tilde{\psi}) - \hat{P}_1(\varphi) - \hat{P}_0(\psi) \end{aligned} \tag{1.74}$$

Suppose there exists $(\varphi, \psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c) \setminus (\Psi(P_1, P_0))^\delta$ such that

$$\sup_{(\tilde{\varphi},\tilde{\psi})\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \hat{P}_1(\tilde{\varphi}) + \hat{P}_0(\tilde{\psi}) \leq \hat{P}_1(\varphi) + \hat{P}_0(\psi) + \frac{\kappa}{\sqrt{n}}.$$

For any such (φ, ψ), (1.74) implies

$$\begin{aligned}\sup_{(\tilde{\varphi},\tilde{\psi})\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} & P_1(\tilde{\varphi}) + P_0(\tilde{\psi}) - P_1(\varphi) - P_0(\psi) \\ & \leq 2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) + \frac{\kappa_n}{\sqrt{n}}\end{aligned}$$

from which it follows that

$$\begin{aligned}& 2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) + \frac{\kappa_n}{\sqrt{n}} \\ & \quad \geq \sup_{(\tilde{\varphi},\tilde{\psi})\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} P_1(\tilde{\varphi}) + P_0(\tilde{\psi}) - \sup_{(\varphi,\psi)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)\setminus(\Psi(P_1,P_0))^\delta} \{P_1(\varphi) + P_0(\psi)\} \\ & \quad = \eta\end{aligned}$$

To summarize: if there exists $(\varphi, \psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c) \setminus (\Psi(P_1, P_0))^\delta$ such that

$$\sup_{(\tilde{\varphi},\tilde{\psi})\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \hat{P}_1(\tilde{\varphi}) + \hat{P}_0(\tilde{\psi}) \leq \hat{P}_1(\varphi) + \hat{P}_0(\psi) + \frac{\kappa}{\sqrt{n}},$$

then $2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) + \frac{\kappa_n}{\sqrt{n}} \geq \eta$, from which it follows that

$$
\begin{aligned}
&P\left(\widehat{\Psi}_c \not\subseteq (\Psi(P_1, P_0))^{\delta}\right) \\
&\quad = P\Bigg(\sup_{(\tilde{\varphi},\tilde{\psi})\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \hat{P}_1(\tilde{\varphi}) + \hat{P}_0(\tilde{\psi}) \leq \hat{P}_1(\varphi) + \hat{P}_0(\psi) + \frac{\kappa}{\sqrt{n}} \\
&\qquad\qquad \text{for some } (\varphi,\psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c) \setminus (\Psi(P_1, P_0))^{\delta} \Bigg) \\
&\quad \leq P\left(2\left(\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0}\right) + \frac{\kappa_n}{\sqrt{n}} \geq \eta\right) \to 0
\end{aligned}
$$

where the final limit claim follows from $\eta > 0$, $\kappa_n/\sqrt{n} \to 0$, and $\|\hat{P}_1 - P_1\|_{\mathcal{F}_1} + \|\hat{P}_0 - P_0\|_{\mathcal{F}_0} = o_p(1)$.

4. (1.72) and (1.73) imply that for any $\delta > 0$, $P\left(\Psi_c(P_1, P_0) \subseteq \widehat{\Psi}_c \subseteq \Psi_c(P_1, P_0)^{\delta}\right) \to 1$. It follows that there exists a sequence $\{\delta_n\}_{n=1}^{\infty} \subseteq \mathbb{R}_+$ with $\delta_n \downarrow 0$ such that $P\left(\Psi(P_1, P_0) \subseteq \widehat{\Psi}_c \subseteq \Psi(P_1, P_0)^{\delta_n}\right) \to 1$. When $\Psi(P_1, P_0) \subseteq \widehat{\Psi}_c \subseteq \Psi(P_1, P_0)^{\delta_n}$ holds,

$$
\begin{aligned}
&\left|\widehat{OT}'_{c,x}(H_1, H_0) - OT'_{c,(P_1,P_0)}(H_1, H_0)\right| \\
&\leq \sup_{(\varphi,\psi)\in\Psi_c(P_1,P_0)^{\delta_n}\cap\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c)} \{H_1(\varphi) + H_0(\psi)\} - \sup_{(\varphi,\psi)\in\Psi_c(P_1,P_0)} \{H_1(\varphi) + H_0(\psi)\} \\
&\leq \sup_{(\varphi_1,\psi_1),(\varphi_2,\psi_2)\in\Phi_c\cap(\mathcal{F}_c\times\mathcal{F}_c^c);\, L_2((\varphi_1,\psi_1),(\varphi_2,\psi_2))<\delta_n} \{H_1(\varphi_1) + H_0(\psi_1) - H_1(\varphi_2) - H_0(\psi_0)\} \\
&= o_p(1)
\end{aligned}
$$

where the $o_p(1)$ claim follows from $H_1 + H_0$ being continuous and $\Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)$ being compact, implying $H_1 + H_0$ is in fact uniformly continuous.

This concludes the proof. □

Theorem 1.5.4. *Suppose assumptions 1, 2, and 3 hold, let $\mathbb{P}_n^*$ be given by definition 1.5.1 or 1.5.2, and $\{\kappa_n\}_{n=1}^{\infty} \subseteq \mathbb{R}$ satisfy $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$. Then conditional on*

$\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$,

$$\hat{D}_4 \hat{D}_3 \widehat{T}_{2,T_1(P)}(\sqrt{n}(T_1(\mathbb{P}_n^*) - T_1(\mathbb{P}_n))) \xrightarrow{L} T'_P(\mathbb{G})$$

in outer probability.

Proof. The overall strategy is to apply Fang & Santos (2019) theorem 3.2, viewing $T_1(\mathbb{P}_n)$ as the estimator for $T_1(P)$, $T_1(\mathbb{P}_n^*)$ as the bootstrap, and $T_{-1} \equiv T_4 \circ T_3 \circ T_2$ as the directionally differentiable function. There are four assumption to verify.

1. To see that Fang & Santos (2019) assumption 1 holds,

   (i) the map

   $$T_4 \circ T_3 \circ T_2 : \prod_{m=1}^{M} \ell^\infty(\mathcal{F}_{1,x}) \times \ell^\infty(\mathcal{F}_{0,x}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R} \to \mathbb{R}^2$$

   is a map between Banach spaces.

   (ii) by lemmas 1.9.23, 1.9.24, 1.9.25 and the chain rule (lemma 1.9.50), $T_{-1} \equiv T_4 \circ T_3 \circ T_2$ is Hadamard directionally differentiable at $T_1(P)$ tangentially to

   $$\mathbb{D}_{Tan} = \prod_{m=1}^{M} \mathcal{C}(\mathcal{F}_{1,x_m}, L_{2,P}) \times \mathcal{C}(\mathcal{F}_{0,x_m}, L_{2,P}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}.$$

2. To see that the estimator $T_1(\mathbb{P}_n)$ satisfies Fang & Santos (2019) assumption 2, note that

   (i) $T_1(P) \in \prod_{m=1}^{M} \ell^\infty(\mathcal{F}_{1,x}) \times \ell^\infty(\mathcal{F}_{0,x}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$ and lemma 1.9.19 shows

   $$T_1(\mathbb{P}_n) : \{Y_i, D_i, Z_i, X_i\}_{i=1}^n \to \prod_{m=1}^{M} \ell^\infty(\mathcal{F}_{1,x}) \times \ell^\infty(\mathcal{F}_{0,x}) \times \mathbb{R}^{K_1} \times \mathbb{R}^{K_0} \times \mathbb{R}$$

   satisfies $\sqrt{n}(T_1(\mathbb{P}_n) - T_1(P)) \xrightarrow{L} T'_{1,P}(\mathbb{G})$.

(ii) $T'_{1,P}(\mathbb{G})$ is tight because $\mathbb{G}$ is tight and $T'_{1,P}$ is continuous. Lemma 1.9.19 also shows the support of $T'_{1,P}(\mathbb{G})$ is included in $\mathbb{D}_{Tan}$.

3. The bootstrap $T_1(\mathbb{P}_n^*)$ satisfies Fang & Santos (2019) assumption 3:

   (i) $T_1(\mathbb{P}_n^*)$ is a function of $\{Y_i, D_i, Z_i, X_i, W_i\}_{i=1}^n$ with $\{W_i\}_{i=1}^n$ independent of $\{Y_i, D_i, Z_i, X_i\}_{i=1}^n$.

   (ii) T_1 is fully Hadamard differentiable at P tangentially to $\ell^\infty(\mathcal{F})$, and hence the functional delta method implies $\sqrt{n}(T_1(\mathbb{P}_n) - T_1(P)) \xrightarrow{L} T'_{1,P}(\mathbb{G})$. Lemma 1.9.28 shows that $\mathbb{P}_n^*$ satisfies Fang & Santos (2019) assumption 3, and thus Fang & Santos (2019) theorem 3.1 implies

$$\sup_{f\in \mathrm{BL}_1} \left| E\left[f(\sqrt{n}(T_1(\mathbb{P}_n^*) - T_1(\mathbb{P}_n))) \mid \{Y_i, D_i, Z_i, X_i\}_{i=1}^n\right] - E[f(T'_{1,P}(\mathbb{G}))]\right| = o_p(1)$$

   (iii) Condition (iv) below holds, and hence Fang & Santos (2019) lemma S.3.9 implies $\sqrt{n}(T_1(\mathbb{P}_n^*) - T_1(\mathbb{P}_n))$ is asymptotically measurable.

   (iv) Note that for any continuous and bounded function f, $f(\sqrt{n}(T_1(\mathbb{P}_n^*) - T_1(\mathbb{P}_n)))$ is continuous in $\{W_i\}_{i=1}^n$ and hence is a measurable function of $\{W_i\}_{i=1}^n$.

4. Fang & Santos (2019) assumption 4 is about the estimator of the derivative.

   Notice that $T'_{-1,T_1(P)} = T'_{4,T_3(T_2(T_1(P)))} \circ T'_{3,T_2(T_1(P))} \circ T'_{2,T_1(P)}$ is given by

$$T'_{-1,T_1(P)} : \mathbb{D}_{Tan} \to \mathbb{R}^2, \qquad T'_{-1,T_1(P)}(h) = D_4 D_3 T'_{2,T_1(P)}(h)$$

   Estimate this derivative with

$$\widehat{T}'_{-1,T_1(P)} : \mathbb{D}_{Tan} \to \mathbb{R}^2, \qquad \hat{D}_4 \hat{D}_3 \widehat{T}'_{2,T_1(P)}(h)$$

   The estimator $\widehat{T}'_{-1,T_1(P)}$ satisfies the conditions of Fang & Santos (2019) lemma S.3.6, and therefore Fang & Santos (2019) assumption 4. These conditions are

(a) Modulus of continuity: $\|\widehat{T}'_{-1,T_1(P)}(h_1) - \widehat{T}'_{-1,T_1(P)}(h_2)\| \leq C_n\|h_1 - h_2\|$ for some $C_n = O_p(1)$.

(b) Pointwise consistency: for any h, $\|\widehat{T}_{-1,T_1(P)}(h) - T_{-1,T_1(P)}(h)\| = o_p(1)$.

To see these claims in detail:

(a) For any matrix A, let $\|A\|_o = \sup_{x;\|x\|_2=1}\|Ax\|_2$ be the operator norm.

$$\begin{aligned}\|\widehat{T}'_{-1,T_1(P)}(h_1) - \widehat{T}'_{-1,T_1(P)}(h_2)\| &= \|\hat{D}_4\hat{D}_3\widehat{T}'_{2,T_1(P)}(h_1) - \hat{D}_4\hat{D}_3\widehat{T}'_{2,T_1(P)}(h_2)\| \\ &\leq \|\hat{D}_4\hat{D}_3\|_o \times \|\widehat{T}'_{2,T_1(P)}(h_1) - \widehat{T}'_{2,T_1(P)}(h_2)\| \\ &\leq \|\hat{D}_4\hat{D}_3\|_o \times \|h_1 - h_2\|\end{aligned}$$

where the last claim follows because $\widehat{T}'_{2,T_1(P)}$ is 1-Lipschitz (shown below). Next notice $\hat{D}_4 \xrightarrow{p} D_4$ and $\hat{D}_3 \xrightarrow{p} D_3$ by the CMT, which implies $\|\hat{D}_4\hat{D}_3\|_o = O_p(1)$ as required.

To see that $\widehat{T}'_{2,T_1(P)}$ is 1-Lipschitz, recall from appendix 1.9.3.3 that

$$\begin{aligned}&\widehat{T}'_{2,T_1(P)}\left(\{H_{1,x}, H_{0,x}, h_{\eta_1,x}, h_{\eta_0,x}, h_{s,x}\}_{x\in\mathcal{X}}\right) \\ &\quad= \left(\left\{\widehat{OT}'_{c_L,x}(H_{1,x}, H_{0,x}), -\widehat{OT}'_{c_H,x}(H_{1,x}, H_{0,x}), h_{\eta_1,x}, h_{\eta_0,x}, h_{s,x}\right\}_{x\in\mathcal{X}}\right)\end{aligned}$$

The maps $\widehat{OT}_{c_L,x}, -\widehat{OT}_{c_H,x}$ are 1-Lipschitz. Specifically, note that

$$\begin{aligned}&|\widehat{OT}'_{c_L,x}(H_{1,x}, H_{0,x}) - \widehat{OT}'_{c_L,x}(G_{1,x}, G_{0,x})| \\ &\quad= \left|\sup_{(\varphi,\psi)\in\widehat{\Psi}_{c,x}}\{H_{1,x}(\varphi) + H_{0,x}(\psi)\} - \sup_{(\varphi,\psi)\in\widehat{\Psi}_{c,x}}\{G_{1,x}(\varphi) + G_{0,x}(\psi)\}\right| \\ &\quad\leq \sup_{\varphi\in\mathcal{F}_{1,x}}|H_{1,x}(\varphi) - G_{1,x}(\varphi)| + \sup_{\psi\in\mathcal{F}_{0,x}}|H_{0,x}(\psi) - G_{0,x}(\psi)| \\ &\quad= \|H_{1,x} - G_{1,x}\|_{\mathcal{F}_{1,x}} + \|H_{0,x} - G_{0,x}\|_{\mathcal{F}_{0,x}}\end{aligned}$$

and similarly, $-\widehat{OT}_{c_H,x}$ is 1-Lipschitz. The other maps in $\widehat{T}_{2,T_1(P)}$ are the identity map, which is also 1-Lipschitz. It follows that $\widehat{T}_{2,T_1(P)}$ is 1-Lipschitz.

(b) To show pointwise consistency, fix $h = (\{H_{1,x}, H_{0,x}, h_{\eta_1,x}, h_{\eta_0,x}, h_{s,x}\}_{x\in\mathcal{X}})$ and note that

$$\begin{aligned}
\|\widehat{T}'_{-1,T_1(P)}(h) - T_{-1,T_1(P)}\| &= \|\hat{D}_4\hat{D}_3\widehat{T}_{2,T_1(P)}(h) - D_4D_3T_{2,T_1(P)}(h)\| \\
&\leq \|(\hat{D}_4\hat{D}_3 - D_4D_3)T'_{2,T_1(P)}(h)\| + \|D_4D_3(\widehat{T}_{2,T_1(P)}(h) - T_{2,T_1(P)}(h))\| \\
&\leq \|\hat{D}_4\hat{D}_3 - D_4D_3\|_o \times \|T'_{2,T_1(P)}(h)\| + \|D_4D_3\|_o \times \|\widehat{T}_{2,T_1(P)}(h) - T_{2,T_1(P)}(h)\|
\end{aligned}$$

Since $\hat{D}_4\hat{D}_3 \xrightarrow{p} D_4D_3$ by the CMT, it suffices to show

$$\|\widehat{T}_{2,T_1(P)}(h) - T_{2,T_1(P)}(h)\| = o_p(1)$$

The only nonzero coordinates correspond to $\widehat{OT}^{L\prime}_{c_L,x}(H_{1,x}, H_{0,x})$ and $-\widehat{OT}^{H\prime}_{c_H,x}(H_{1,x}, H_{0,x})$:

$$\begin{aligned}
&\|\widehat{T}_{2,T_1(P)}(h) - T_{2,T_1(P)}(h)\|^2 \\
&\quad= \left(\widehat{OT}'_{c_L,x}(H_{1,x}, H_{0,x}) - OT'_{c_L,(P_{1|x},P_{0|x})}(H_{1,x}, H_{0,x})\right)^2 \\
&\qquad\qquad + \left(\widehat{OT}'_{c_H,x}(H_{1,x}, H_{0,x}) - OT'_{c_H,(P_{1|x},P_{0|x})}(H_{1,x}, H_{0,x})\right)^2 \\
&\quad= o_p(1) + o_p(1)
\end{aligned}$$

where the last $o_p(1)$ claim follows from lemma 1.9.29.

We conclude through Fang & Santos (2019) lemma S.3.6 that Fang & Santos (2019) assumption 4 is satisfied.

Finally, apply Fang & Santos (2019) theorem 3.2 to find that

$$\sup_{f\in \mathrm{BL}_1} \left| E\left[f(\hat{D}_4 \hat{D}_3 \widehat{T}_{2,T_1(P)}(\sqrt{n}(T_1(\mathbb{P}_n^*) - T_1(\mathbb{P}_n)))) \right] - E\left[f(T'_P(\mathbb{G})) \right] \right| = o_p(1)$$

as was to be shown. □

1.9.5 Appendix: duality in optimal transport

This appendix contains terminology, notation, and results regarding optimal transport used in this paper. Many of these results can be found in the monographs Villani (2003), Villani (2009), or Santambrogio (2015).

1.9.5.1 Primal and dual problems

Let $\mathcal{Y}_1, \mathcal{Y}_0$ be Polish subsets of $\mathbb{R}$, equipped with their Borel sigma algebras. Let $\mathcal{P}(\mathcal{Y}_d)$ be the set of probability distributions defined on $\mathcal{Y}_d$, and $P_d \in \mathcal{P}(\mathcal{Y}_d)$. Let $\mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_0)$ be the set of probability distributions on the product space $\mathcal{Y}_1 \times \mathcal{Y}_0$.

A probability measure $\pi \in \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_0)$ has marginals P_1 and P_0 if

$$\text{For all } A \subset \mathcal{Y}_1 \text{ measurable, } \pi(A \times \mathcal{Y}_0) = P_1(A) = \int \mathbb{1}_A(y_1) dP_1(y_1) \tag{1.75}$$

$$\text{For all } B \subset \mathcal{Y}_0 \text{ measurable, } \pi(\mathcal{Y}_1 \times B) = P_0(B) = \int \mathbb{1}_B(y_0) dP_0(y_0) \tag{1.76}$$

The collection of such joint distributions with marginals P_1 and P_0 is denoted

$$\Pi(P_1, P_0) = \{\pi \in \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_0) \; ; \; \pi \text{ satisfies (1.75) and (1.76)}\} \tag{1.77}$$

The *cost function* is a measurable function $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$. The functional $I : \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_0) \to \mathbb{R} \cup \{+\infty\}$ is defined as

$$I_c[\pi] = \int c(y_1, y_0) d\pi(y_1, y_0) \tag{1.78}$$

The *optimal cost* $OT_c(P_1, P_0)$ is the infimum of $I_c[\pi]$ over $\Pi(P_1, P_0)$:

$$OT_c(P_1, P_0) = \inf_{\pi \in \Pi(P_1, P_0)} I_c[\pi] = \inf_{\pi \in \Pi(P_1, P_0)} \int c(y_1, y_0) d\pi(y_1, y_0) \tag{1.79}$$

This minimization problem in (1.79) is known as *optimal transport*. When attained, a solution to (1.79) is called an *optimal transference plan* or *optimal coupling*. Attainment is common; Villani (2009) theorem 4.1 implies:

Lemma 1.9.30 (Optimal transport is attained)**.** *Let $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ be lower semicontinuous and bounded from below. Then there exists $\pi^* \in \Pi(P_1, P_0)$ such that*

$$E_{\pi^*}[c(Y_1, Y_0)] = \inf_{\pi \in \Pi(P_1, P_0)} \int c(y_1, y_0) d\pi(y_1, y_0)$$

The dual problem will require some additional notation. For any probability measure P let $L^1(P)$ denote the P-integrable functions. Define

$$\Phi_c = \left\{ (\varphi, \psi) \in L^1(P_1) \times L^1(P_0) \; ; \; \varphi(y_1) + \psi(y_0) \leq c(y_1, y_0) \right\}, \tag{1.80}$$

and $J : L^1(P_1) \times L^1(P_0) \to \mathbb{R}$ by

$$J(\varphi, \psi) = \int_{\mathcal{Y}_1} \varphi(y_1) dP_1(y_1) + \int_{\mathcal{Y}_0} \psi(y_0) dP_0(y_0) \tag{1.81}$$

The *dual problem* of optimal transport is

$$\sup_{(\varphi,\psi) \in \Phi_c} J(\varphi, \psi) = \sup_{(\varphi,\psi) \in \Phi_c} \int \varphi(y_1) dP_1(y_1) + \int \psi(y_0) dP_0(y_0) \tag{1.82}$$

1.9.5.2 Duality

For any topological space $\mathcal{Z}$, let $\mathcal{C}_b(\mathcal{Z})$ denotes the set of functions $f : \mathcal{Z} \to \mathbb{R}$ that are continuous and bounded, and

$$\Phi_c \cap \mathcal{C}_b = \{(\varphi, \psi) \in \mathcal{C}_b(\mathcal{Y}_1) \times \mathcal{C}_b(\mathcal{Y}_0) \; ; \; \varphi(y_1) + \psi(y_0) \leq c(y_1, y_0)\} \tag{1.83}$$

The following weak duality statement is Villani (2003) proposition 1.5.

Lemma 1.9.31 (Weak duality)**.**

$$\sup_{(\varphi,\psi)\in\Phi_c\cap\mathcal{C}_b} J(\varphi, \psi) \leq \sup_{(\varphi,\psi)\in\Phi_c} J(\varphi, \psi) \leq \inf_{\pi\in\Pi(P_1,P_0)} I_c[\pi]$$

The following strong duality statement can be directly inferred from Villani (2009) theorem 5.10, or Santambrogio (2015) theorem 1.42, and so is presented without proof.

Theorem 1.9.32 (Strong duality)**.** *Let $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ be lower semi-continuous and bounded from below. Then*

$$\inf_{\pi\in\Pi(P_1,P_0)} I_c[\pi] = \sup_{\varphi,\psi\in\Phi_c} J(\varphi, \psi) = \sup_{(\varphi,\psi)\in\Phi_c\cap\mathcal{C}_b} J(\varphi, \psi) \tag{1.84}$$

Moreover, the infimum of the left-hand side of (1.84) *is attained.*

1.9.5.3 c-concave functions

For any function $\varphi : \mathcal{Y}_1 \to \mathbb{R}$ and cost function $c(y_1, y_0)$, define the *c-transform* of φ as the function $\varphi^c : \mathcal{Y}_0 \to \mathbb{R}$ given by

$$\varphi^c(y_0) = \inf_{y_1\in\mathcal{Y}_1} \{c(y_1, y_0) - \varphi(y_1)\}.$$

Similarly, $\psi^c(y_1) = \inf_{y_0 \in \mathcal{Y}_0}\{c(y_1, y_0) - \psi(y_0)\}$ is the c-transform of ψ. φ is called *c-concave* if $\varphi^{cc} = (\varphi^c)^c = \varphi$. If φ is c-concave, then (φ, φ^c) is called a *c-concave conjugate pair*.

The following lemma 1.9.33 is exercise 2.35 found in Villani (2003) and presented without proof.

Lemma 1.9.33 (Villani (2003) exercise 2.35)**.** *Let $\mathcal{Y}_1$ and $\mathcal{Y}_0$ be nonempty sets and $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ be an arbitrary function. Let $\varphi : \mathcal{Y}_1 \to \mathbb{R}$. Then*

(i) $\varphi(y_1) + \varphi^c(y_0) \leq c(y_1, y_0)$ *for all* $(y_1, y_0) \in \mathcal{Y}_1 \times \mathcal{Y}_0$

(ii) $\varphi^{cc}(y_1) \geq \varphi(y_1)$ *for all* $y_1 \in \mathcal{Y}_1$*, and*

(iii) $\varphi^{ccc}(y_0) = \varphi^c(y_0)$ *for all* $y_0 \in \mathcal{Y}_0$

It follows that $\varphi^{cc} = \varphi$ if and only if φ is c-concave.

For $H \subseteq \{(f, g)\ ;\ f : \mathcal{Y}_1 \to \mathbb{R},\ \text{and}\ g : \mathcal{Y}_0 \to \mathbb{R}\}$, let

$$\mathcal{F}_c^c(H) = \left\{\varphi^c : \mathcal{Y}_0 \to \mathbb{R}\ ;\ \exists (f, g) \in H \text{ s.t. } \varphi^c(y_0) = \inf_{y_1 \in \mathcal{Y}_1}\{c(y_1, y_0) - f(y_1)\}\right\} \tag{1.85}$$
$$\mathcal{F}_c(H) = \left\{\varphi : \mathcal{Y}_1 \to \mathbb{R}\ ;\ \exists \varphi^c \in F_c^c(H) \text{ s.t. } \varphi(y_1) = \inf_{y_0 \in \mathcal{Y}_0}\{c(y_1, y_0) - \varphi^c(y_0)\}\right\}$$

$\mathcal{F}_c(H)$ is called the *c-concave functions generated by* H, and $\mathcal{F}_c^c(H)$ the *c-conjugates generated by* H.[13] Notice that not every $(\varphi, \psi) \in \mathcal{F}_c(H) \times \mathcal{F}_c^c(H)$ is a c-concave conjugate pair.

Lemma 1.9.34 (Restricting the dual to c-concave functions)**.** *Let $\Phi_{cs} \subseteq \Phi_c$ be such that*

1. *strong duality holds:* $\inf_{\pi \in \Pi(P_1, P_0)} I_c[\pi] = \sup_{(\varphi, \psi) \in \Phi_{cs}} J(\varphi, \psi)$*, and*

[13] H is a typically a subset of $L^1(P_1) \times L^1(P_0)$. As defined the sets $\mathcal{F}_c(H)$ and $\mathcal{F}_c^c(H)$ only depend on the functions in H that map $\mathcal{Y}_0$ to $\mathbb{R}$. This notational choice is more natural with the reasoning of lemma 1.9.34 below.

2. *the c-concave functions generated by* Φ_{cs} *are integrable:* $\mathcal{F}_c(\Phi_{cs}) \times \mathcal{F}_c^c(\Phi_{cs}) \subset L^1(P_1) \times L^1(P_0)$.

Then

$$\inf_{\pi \in \Pi(P_1,P_0)} I_c[\pi] = \sup_{\varphi \in \mathcal{F}_c(\Phi_{cs})} J(\varphi, \varphi^c) = \sup_{(\varphi,\psi) \in \Phi_c \cap \left(\mathcal{F}_c(\Phi_{cs}) \times \mathcal{F}_c^c(\Phi_{cs})\right)} J(\varphi, \psi).$$

Proof. Let $(\varphi, \psi) \in \Phi_{cs}$. $\psi(y_0) \leq c(y_1, y_0) - \varphi(y_1)$ implies $\psi(y_0) \leq \varphi^c(y_0)$, and lemma 1.9.33 shows both that $\varphi(y_1) \leq \varphi^{cc}(y_1)$ and the pair $(\varphi^{cc}, \varphi^c)$ is a c-concave conjugate pair; thus $(\varphi^{cc}, \varphi^c) \in \Phi_c \cap \left(\mathcal{F}_c(\Phi_{cs}) \times \mathcal{F}_c^c(\Phi_{cs})\right)$.

Since φ^{cc} and φ^c are integrable by assumption, $J(\varphi, \psi) \leq J(\varphi^{cc}, \varphi^c)$ and hence

$$\inf_{\pi \in \Pi(P_1,P_0)} I_c[\pi] = \sup_{(\varphi,\psi) \in \Phi_{cs}} J(\varphi, \psi) \leq \sup_{\varphi^{cc} \in \mathcal{F}_c(\Phi_{cs})} J(\varphi^{cc}, \varphi^c) \leq \sup_{(\varphi,\psi) \in \Phi_c \cap (\mathcal{F}_c(\Phi_{cs}) \times \mathcal{F}_c^c(\Phi_{cs}))} J(\varphi, \psi)$$

Finally, since $\Phi_c \cap (\mathcal{F}_c(\Phi_{cs}) \times \mathcal{F}_c^c(\Phi_{cs})) \subset \Phi_c$, it follows that

$$\sup_{\varphi \in \mathcal{F}_c(\Phi_{cs})} J(\varphi, \varphi^c) \leq \sup_{(\varphi,\psi) \in \Phi_c} J(\varphi, \psi) = \inf_{\pi \in \Pi(P_1,P_0)} I_c[\pi]$$

with the final equality following from strong duality. □

Lemma 1.9.35 (Continuous cost function implies measurability of c-concave functions)**.** *If* $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ *is continuous, then for any* $\psi : \mathcal{Y}_0 \to \mathbb{R}$, $\varphi(y_1) = \inf_{y_0 \in \mathcal{Y}_0}\{c(y_1, y_0) - \psi(y_0)\}$ *and* $\varphi^c(y_0) = \inf_{y_1 \in \mathcal{Y}_1}\{c(y_1, y_0) - \varphi(y_1)\}$ *are upper semicontinuous and hence measurable.*

Proof. The pointwise infimum of a family of upper semicontinuous functions is upper semicontinuous (Aliprantis & Border (2006) Lemma 2.41). Since $c(y_1, y_0)$ is continuous, for any fixed $y_0 \in \mathcal{Y}_0$ the function $y_1 \mapsto c(y_1, y_0) - \psi(y_0)$ is continuous and hence

$$\varphi(y_1) = \inf_{y_0 \in \mathcal{Y}_0}\{c(y_1, y_0) - \psi(y_0)\}$$

is upper semicontinuous. Similarly, $\varphi^c(y_0) = \inf_{y_1 \in \mathcal{Y}_1}\{c(y_1, y_0) - \varphi(y_1)\}$ is upper semicontinuous. Being upper semicontinuous, φ and φ^c are measurable. □

Remark 1.9.1. Compare lemma 1.9.35 with Villani (2009) Remark 5.5 discussing measurability of c-concave functions. Note that continuity of c is sufficient but not necessary for measurability of c-concave functions; see section 1.9.5.5 for counterexamples.

Lemma 1.9.36 (Universal bound on the the dual problem feasible set)**.** *Suppose* $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ *is bounded. Let* $c_L \equiv \inf_{(y_1,y_0)\in\mathcal{Y}_1\times\mathcal{Y}_0} c(y_1, y_0)$ *and* $c_H \equiv \sup_{(y_1,y_0)\in\mathcal{Y}_1\times\mathcal{Y}_0} c(y_1, y_0)$.

1. *For any bounded functions* $\varphi : \mathcal{Y}_1 \to \mathbb{R}$ *and* $\psi : \mathcal{Y}_0 \to \mathbb{R}$, φ^c *and* ψ^c *are bounded.*
2. *For any bounded, measurable c-conjugate pair* (φ, φ^c) *there exists* $\bar{\varphi}$ *such that*

   (i) $\bar{\varphi}$ *and* $\bar{\varphi}^c$ *satisfy the bounds:*

   $$c_L \leq \bar{\varphi}(y_1) \leq c_H \qquad\qquad c_L - c_H \leq \bar{\varphi}^c(y_0) \leq 0$$

   for all $(y_1, y_0) \in \mathcal{Y}_1 \times \mathcal{Y}_0$.

   (ii) $J(\varphi, \varphi^c) = J(\bar{\varphi}, \bar{\varphi}^c)$.

Proof. For claim 1, let φ be bounded and note that

$$c_L - \sup\varphi \leq \underbrace{\inf_{y_1\in\mathcal{Y}_1}\{c(y_1, y_0) - \varphi(y_1)\}}_{=\varphi^c(y_0)} \leq c_H - \sup\varphi \tag{1.86}$$

are finite bounds on φ^c. The upper bound on φ^c follows from the existence of a sequence $\{y_{1j}\}_{j=1}^{\infty}$ with $\varphi(y_{1j}) \to \sup_{y_1\in\mathcal{Y}_1}\varphi(y_1)$, because $\varphi^c(y_0) = \inf_{y_1\in\mathcal{Y}_1}\{c(y_1, y_0) - \varphi(y_1)\} \leq c(y_{1j}, y_0) - \varphi(y_{1j}) \leq c_H - \varphi(y_{1j})$ for all j. The same argument shows ψ^c is bounded, specifically,

$$c_L - \sup\psi \leq \underbrace{\inf_{y_0\in\mathcal{Y}_0}\{c(y_1, y_0) - \psi(y_0)\}}_{=\psi^c(y_1)} \leq c_H - \sup\psi \tag{1.87}$$

For claim 2, let (φ, φ^c) be a c-conjugate pair, i.e. $\varphi(y_1) = \inf_{y_0 \in \mathcal{Y}_0}\{c(y_1, y_0) - \varphi^c(y_0)\}$. Notice that for any $s \in \mathbb{R}$,

$$(\varphi + s)^c(y_0) = \inf_{y_1 \in \mathcal{Y}_1}\{c(y_1, y_0) - \varphi(y_1) - s\} = \varphi^c(y_0) - s$$
$$(\varphi + s)^{cc}(y_0) = \inf_{y_0 \in \mathcal{Y}_0}\{c(y_1, y_0) - \varphi^c(y_1) + s\} = \varphi(y_1) + s$$

Define $\bar{\varphi}(y_1) = \varphi(y_1) - \sup\varphi + c_H$, and notice that $\sup\bar{\varphi} = c_H$. Thus (1.86) implies $c_L - c_H \leq \bar{\varphi}^c(y_0) \leq 0$ for all $y_0 \in \mathcal{Y}_0$, and so (1.87) implies $c_L \leq \bar{\varphi}^{cc}(y_1) = \bar{\varphi}(y_1) \leq c_H$. Finally,

$$\begin{aligned} J(\varphi, \varphi^c) &= \int \varphi(y_1) dP_1(y_1) + \int \varphi^c(y_0) dP_0(y_0) \\ &= \int \varphi(y_1) - \sup\varphi + c_H dP_1(y_1) + \int \varphi^c(y_0) + \sup\varphi - c_H dP_0(y_0) \\ &= J(\bar{\varphi}, \bar{\varphi}^c) \end{aligned}$$

which completes the proof. □

Remark 1.9.2. Lemma 1.9.36 shows that it is often without loss of generality to restrict the dual to classes of functions sharing universal bounds. For an example, see lemma 1.9.38 below.

Note that when $c_L = 0$, the bounds simplify to

$$0 \leq \bar{\varphi}(y_1) \leq \|c\|_\infty, \qquad -\|c\|_\infty \leq \bar{\varphi}^c(y_0) \leq 0$$

as in Villani (2003) Remark 1.13. Also note that, when any universal bound suffices, one can take

$$-\|c\|_\infty \leq \bar{\varphi}(y_1) \leq \|c\|_\infty, \qquad -2\|c\|_\infty \leq \bar{\varphi}^c(y_0) \leq 0$$

which depend only on $\|c\|_\infty = \sup_{(y_1,y_0)\in\mathcal{Y}_1\times\mathcal{Y}_0}|c(y_1,y_0)|$.

1.9.5.4 c-concave functions of smooth cost functions

For $\alpha \in (0,1]$ and $L > 0$, $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ is called (α, L)-*Hölder continuous* if

$$|c(y_1,y_0) - c(y_1',y_0')| \le L\|(y_1,y_0) - (y_1',y_0')\|^\alpha$$

for all $(y_1,y_0),(y_1',y_0') \in \mathcal{Y}_1 \times \mathcal{Y}_0$.

Lemma 1.9.37 (Hölder cost implies Hölder c-concave functions)**.** *Let $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ be (α, L)-Hölder continuous. For any $g : \mathcal{Y}_0 \to \mathbb{R}$,*

$$\varphi(y_1) = \inf_{y_0\in\mathcal{Y}_0}\{c(y_1,y_0) - g(y_0)\}, \qquad \varphi^c(y_0) = \inf_{y_1\in\mathcal{Y}_1}\{c(y_1,y_0) - \varphi(y_1)\}$$

are (α, L)-Hölder continuous.

Proof. Hölder continuity implies $c(y_1,y_0) \le c(y_1',y_0) + L|y_1 - y_1'|^\alpha$ holds for any $y_0 \in \mathcal{Y}_0$ and any $y_1, y_1' \in \mathcal{Y}_1$. It follows that

$$\varphi(y_1) = \inf_{\tilde{y}_0\in\mathcal{Y}_0}\{c(y_1,\tilde{y}_0) - g(\tilde{y}_0)\} \le c(y_1,y_0) - g(y_0) \le c(y_1',y_0) - g(y_0) + L|y_1 - y_1'|^\alpha$$

implying $\varphi(y_1) - (c(y_1',y_0) - g(y_0)) \le L|y_1 - y_1'|^\alpha$. Therefore

$$\varphi(y_1) - \varphi(y_1') = \varphi(y_1) - \inf_{y_0\in\mathcal{Y}_0}\{c(y_1',y_0) - g(y_0)\} \le L|y_1 - y_1'|^\alpha$$

holds for any $y_1, y_1' \in \mathcal{Y}_1$. Reverse the role of y_1 and y_1' to find $\varphi(y_1') - \varphi(y_1) \le L|y_1' - y_1|^\alpha$, and hence φ is (α, L)-Hölder. The same argument implies φ^c is (α, L)-Hölder. □

Lemma 1.9.38 is relevant for compact $\mathcal{Y}_1, \mathcal{Y}_0 \subset \mathbb{R}$, and L-Lipscthiz $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$.

Under these conditions, define

$$\mathcal{F}_c = \{\varphi : \mathcal{Y}_1 \to \mathbb{R} \,;\; -\|c\|_\infty \le \varphi(y_1) \le \|c\|_\infty,\; |\varphi(y_1) - \varphi(y_1')| \le L|y_1 - y_1'|\} \tag{1.88}$$

$$\mathcal{F}_c^c = \{\psi : \mathcal{Y}_0 \to \mathbb{R} \,;\; -2\|c\|_\infty \le \psi(y_0) \le 0,\; |\psi(y_0) - \psi(y_0')| \le L|y_0 - y_0'|\} \tag{1.89}$$

Lemma 1.9.38 (Strong duality for smooth cost functions)**.** *Let $\mathcal{Y}_1, \mathcal{Y}_0 \subset \mathbb{R}$ be compact, $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ be L-Lipschitz, and $\mathcal{F}_c$, $\mathcal{F}_c^c$ be given by* (1.88) *and* (1.89) *respectively. Then strong duality holds:*

$$\inf_{\pi \in \Pi(P_1, P_0)} I_c[\pi] = \sup_{(\varphi,\psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)} J(\varphi, \psi)$$

Proof. First notice lemma 1.9.37 implies $\mathcal{F}_c(\Phi_c \cap \mathcal{C}_b)$ and $\mathcal{F}_c^c(\Phi_c \cap \mathcal{C}_b)$, the set of c-concave functions and c-conjugates generated by $\Phi_c \cap \mathcal{C}_b$ respectively, consist of L-Lipschitz functions.[14] Since c is continuous and $\mathcal{Y}_1 \times \mathcal{Y}_0$ is compact, $\|c\|_\infty = \sup_{y_1, y_0 \in \mathcal{Y}_1 \times \mathcal{Y}_0} |c(y_1, y_0)| < \infty$. Continuity implies these c-concave functions are measurable, and lemma 1.9.36 shows they are bounded. Thus $\mathcal{F}_c(\Phi_c \cap \mathcal{C}_b) \times \mathcal{F}_c^c(\Phi_c \cap \mathcal{C}_b) \subseteq L^1(P_1) \times L^1(P_0)$, and so lemma 1.9.34 implies

$$\inf_{\pi \in \Pi(P_1, P_0)} I_c[\pi] = \sup_{\varphi \in \mathcal{F}_c(\Phi_c \cap \mathcal{C}_b)} J(\varphi, \varphi^c)$$

Lemma 1.9.36 and remark 1.9.2 further shows that for every $\varphi \in \mathcal{F}_c(\Phi_c \cap \mathcal{C}_b)$, there exists a shifted function $\bar{\varphi}$ such that $\sup_{y_1 \in \mathcal{Y}_1} |\bar{\varphi}(y_1)| \le \|c\|_\infty$, $-2\|c\| \le \bar{\varphi}^c(y_0) \le 0$, $\bar{\varphi}$ and $\bar{\varphi}^c$ are L-lipschitz, and $J(\varphi, \varphi^c) = J(\bar{\varphi}, \bar{\varphi}^c)$. Thus

$$\sup_{\varphi \in \mathcal{F}_c(\Phi_c \cap \mathcal{C}_b)} J(\varphi, \varphi^c) = \sup_{\varphi \in \mathcal{F}_c} J(\varphi, \varphi^c)$$

Finally,

$$\sup_{\varphi \in \mathcal{F}_c} J(\varphi, \varphi^c) \le \sup_{(\varphi,\psi) \in \Phi_c \cap (\mathcal{F}_c \times \mathcal{F}_c^c)} J(\varphi, \psi) \le \sup_{(\varphi,\psi) \in \Phi_c} J(\varphi, \psi) = \inf_{\pi \in \Pi(P_1, P_0)} I_c[\pi]$$

[14] Note that $\mathcal{F}_c(\Phi_c \cap \mathcal{C}_b)$ and $\mathcal{F}_c^c(\Phi_c \cap \mathcal{C}_b)$ are not necessarily $\mathcal{F}_c$ and $\mathcal{F}_c^c$ defined in the statement of the lemma.

completes the proof. □

Remark 1.9.3. Suppose $\mathcal{Y}_1$ and $\mathcal{Y}_0$ are compact and $c(y_1, y_0)$ is continuously differentiable on an open set containing $\mathcal{Y}_1 \times \mathcal{Y}_0$. Then c restricted to $\mathcal{Y}_1 \times \mathcal{Y}_0$ is bounded and Lipschitz.

That $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ is bounded follows from c being continuous, $\mathcal{Y}_1 \times \mathcal{Y}_0$ being compact, and the extreme value theorem. To see that c restricted to $\mathcal{Y}_1 \times \mathcal{Y}_0$ is L-Lipschitz, let $(y_1, y_0), (y_1', y_0') \in \mathcal{Y}_1 \times \mathcal{Y}_0$ be arbitrary and note that the mean value theorem applied to $g(t) = c(t(y_1, y_0) + (1-t)(y_1', y_0'))$ implies there exists $s \in (0,1)$ such that

$$\begin{aligned}(c(y_1, y_0) - c(y_1', y_0')) &= g(1) - g(0) = g'(s) \\ &= \langle \nabla c(s(y_1, y_0) + (1-s)(y_1', y_0')), (y_1, y_0) - (y_1', y_0') \rangle\end{aligned}$$

Notice that Cauchy-Schwarz then implies

$$\begin{aligned}|c(y_1, y_0) - c(y_1', y_0')| &\leq \|\nabla c(s(y_1, y_0) + (1-s)(y_1', y_0'))\| \|(y_1, y_0) - (y_1', y_0')\| \\ &\leq \sup_{(y_1'', y_0'') \in \mathcal{Y}_1 \times \mathcal{Y}_0} \|\nabla c(y_1'', y_0'')\| \|(y_1, y_0) - (y_1', y_0')\|\end{aligned}$$

Finally, notice $L = \sup_{(y_1'', y_0'') \in \mathcal{Y}_1 \times \mathcal{Y}_0} \|\nabla c(y_1'', y_0'')\|$ is finite because $\mathcal{Y}_1 \times \mathcal{Y}_0$ is compact and $(y_1, y_0) \mapsto \|\nabla c(y_1, y_0)\|$ is continuous.

1.9.5.5 c-concave functions when $c(y_1, y_0) = \mathbb{1}\{(y_1, y_0) \in C\}$

Theorem 1.9.39 (Strong duality with indicator costs)**.** *Let C be a nonempty, open subset of $\mathcal{Y}_1 \times \mathcal{Y}_0$, and $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ given by $c(y_1, y_0) = \mathbb{1}_C(y_1, y_0) = \mathbb{1}\{(y_1, y_0) \in C\}$. Then*

$$\inf_{\pi \in \Pi(P_1, P_0)} \int \mathbb{1}_C(y_1, y_0) d\pi(y_1, y_0) = \sup_{(A,B) \in \Phi_c^I} \int \mathbb{1}_A(y_1) dP_1(y_1) - \int \mathbb{1}_B(y_0) d\nu(y_0)$$

where

$$\Phi_c^I = \Big\{(A,B)\ ;\ A \subset \mathcal{Y}_1 \textit{ is closed and nonempty, } B \subset \mathcal{Y}_0 \textit{ is measurable,} \\ \textit{and } \mathbb{1}_A(y_1) - \mathbb{1}_B(y_0) \leq \mathbb{1}_C(y_1,y_0)\Big\}.$$

Proof. Villani (2003) Theorem 1.27 implies

$$\inf_{\pi \in \Pi(P_1,P_0)} \int \mathbb{1}_C(y_1,y_0)d\pi(y_1,y_0) = \sup_{A \text{ closed}} \int \mathbb{1}_A(y_1)dP_1(y_1) - \int \mathbb{1}_{A^C}(y_0)dP_0(y_0)$$

where $A^C = \{y \in \mathcal{Y}_0\ ;\ \exists y_1 \in A,\ (y_1,y_0) \notin C\}$ is the projection of $(A \times \mathcal{Y}_0) \setminus C$ onto $\mathcal{Y}_0$. It is clear that

$$\sup_{A \text{ closed}} \int \mathbb{1}_A(y_1)dP_1(y_1) - \int \mathbb{1}_{A^C}(y_0)dP_0(y_0) \leq \sup_{A \subseteq \mathcal{Y}_1, B \subseteq \mathcal{Y}_0} \int \mathbb{1}_A(y_1)dP_1(y_1) - \int \mathbb{1}_B(y_0)d\nu(y_0)$$

with A, B measurable. Notice it is without loss to exclude $A = \varnothing$, because $J(\mathbb{1}_\varnothing, -\mathbb{1}_B) \leq 0 = J(\mathbb{1}_{\mathcal{Y}_1}, \mathbb{1}_{\mathcal{Y}_0})$ and $\mathbb{1}_{\mathcal{Y}_1}(y_1) - \mathbb{1}_{\mathcal{Y}_0}(y_0) = 0 \leq \mathbb{1}_C(y_1,y_0)$ for all $(y_1,y_0) \in \mathcal{Y}_1 \times \mathcal{Y}_0$. Thus

$$\sup_{A \subseteq \mathcal{Y}_1, B \subseteq \mathcal{Y}_0} \int \mathbb{1}_A(y_1)dP_1(y_1) - \int \mathbb{1}_B(y_0)d\nu(y_0) = \sup_{(A,B) \in \Phi_c^I} \int \mathbb{1}_A(y_1)dP_1(y_1) - \int \mathbb{1}_B(y_0)d\nu(y_0)$$

Weak duality (lemma 1.9.31) implies

$$\sup_{(A,B) \in \Phi_c^I} \int \mathbb{1}_A(y_1)dP_1(y_1) - \int \mathbb{1}_B(y_0)dP_0(y_0) \leq \inf_{\pi \in \Pi(P_1,P_0)} \int \mathbb{1}_C(y_1,y_0)d\pi(y_1,y_0)$$

and the result follows. □

The strong duality result of theorem 1.9.39 is especially useful when combined with a careful characterization of the corresponding c-concave functions. To describe these, let

$A \subseteq \mathcal{Y}_1$ be nonempty, and define

$$A^C = \{y_0 \in \mathcal{Y}_0 \,;\, \exists y_1 \in A,\ (y_1, y_0) \notin C\}, \qquad A^{CC} = \left\{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0 \setminus A^C,\ (y_1, y_0) \in C\right\},$$

(1.90)

$$C_{0m} = \{y_0 \in \mathcal{Y}_0 \,;\, \forall y_1 \in \mathcal{Y}_1,\ (y_1, y_0) \in C\}, \quad C_{1m} = \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ (y_1, y_0) \in C\}$$

(1.91)

$$C_{0m}^C = \begin{cases} C_{1m} & \text{if } C_{0m} = \varnothing \\ \varnothing & \text{if } C_{0m} \neq \varnothing \end{cases}, \qquad C_{1m}^C = \begin{cases} C_{0m} & \text{if } C_{1m} = \varnothing \\ \varnothing & \text{if } C_{1m} \neq \varnothing \end{cases} \tag{1.92}$$

Note that A^C is well defined whenever $A \neq \varnothing$, and to ensure A^{CC} is well defined we require $A^C \neq \mathcal{Y}_0$. C_{0m} is denoted as such because $\mathbb{1}_{C_{0m}}(y_0) = \inf_{y_1 \in \mathcal{Y}_1} \mathbb{1}_C(y_1, y_0)$ is the subset of $\mathcal{Y}_{\underline{0}}$ found by minimizing $\mathbb{1}_C(y_1, y_0)$ over $y_1 \in \mathcal{Y}_1$.

Lemma 1.9.40 (c-concave functions for indicator costs)**.** *Let C be a nonempty, open subset of $\mathcal{Y}_1 \times \mathcal{Y}_0$, $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ given by $c(y_1, y_0) = \mathbb{1}_C(y_1, y_0)$, $A \subseteq \mathcal{Y}_1$ be closed and nonempty, and $\varphi(y_1) = \mathbb{1}_A(y_1) = \mathbb{1}\{y_1 \in A\}$. Then*

1. *$\varphi^c(y_0) = -\mathbb{1}_{A^C}(y_0)$,*
2. *if $A^C \neq \mathcal{Y}_0$, then $\varphi^{cc}(y_1) = \mathbb{1}_{A^{CC}}(y_1)$, and*
3. *If $A^C = \mathcal{Y}_0$, then $J(\varphi^{cc}, \varphi^c) = J(\mathbb{1}_{C_{1m}}, 0)$*

Proof. 1. Notice $\mathbb{1}_C(y_1, y_0) - \mathbb{1}_A(y_1) \in \{-1, 0, 1\}$, and

$$\varphi^c(y_0) = \inf_{y_1 \in \mathcal{Y}_1} \{\mathbb{1}_C(y_1, y_0) - \mathbb{1}_A(y_1)\}$$

will never take value 1 because any $y_1 \in A$ implies the objective is at most 0. Furthermore, if there exists $y_1 \in A$ such that $(y_1, y_0) \notin C$, then the infimum attains -1. If there does not exist such y_1, then $\varphi^c(y_0) = 0$. Thus $\varphi^c(y_0) = -\mathbb{1}_{A^C}(y_0)$.

2. Suppose $A^C \neq \mathcal{Y}_0$. Notice that $\mathbb{1}_C(y_1, y_0) + \mathbb{1}_{A^C}(y_0)$ takes values in $\{0, 1, 2\}$, and

$$\varphi^{cc}(y_1) = \inf_{y_0 \in \mathcal{Y}_0} \{\mathbb{1}_C(y_1, y_0) + \mathbb{1}_{A^C}(y_0)\}$$

will never equal 2 because $\mathcal{Y}_0 \setminus A^C \neq \varnothing$. Moreover, the infimum will equal 1 if and only if $(y_1, y_0) \in C$ for all $y_0 \in \mathcal{Y}_0 \setminus A^C$; thus $\varphi^{cc}(y_1) = \mathbb{1}_{A^{CC}}(y_1)$.

3. If $A^C = \mathcal{Y}_0$, then $\varphi^{cc}(y_1) = \inf_{y_0 \in \mathcal{Y}_0}\{\mathbb{1}_C(y_1, y_0) + 1\} = \mathbb{1}_{C_{1m}}(y_1) + 1$ and

$$\varphi^{ccc}(y_0) = \inf_{y_1 \in \mathcal{Y}_1} \{\mathbb{1}_C(y_1, y_0) - \mathbb{1}_{C_{1m}}(y_1) - 1\} = \mathbb{1}_{C_{1m}^C}(y_0) - 1$$

To see that $(\mathbb{1}_{C_{1m}})^c = 0$ if $C_{1m} \neq \varnothing$, notice the objective $\mathbb{1}_C(y_1, y_0) - \mathbb{1}_{C_{1m}}(y_0)$ takes values in $\{-1, 0, 1\}$, and because $C_{1m} \neq \varnothing$ will never take value 1. For the objective to take value -1 at a given y_1, it must be the case that $\mathbb{1}_{C_{1m}}(y_1) = 1$ and there exists y_0 such that $\mathbb{1}_C(y_1, y_0) = 0$, but this contradicts the definition $C_{1m} = \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ (y_1, y_0) \in C\}$.

However, recall that $\varphi^{ccc}(y_0) = \varphi^c(y_0)$ as shown in lemma 1.9.33. Since $\varphi^c(y_0) = -\mathbb{1}_{A^C}(y_0) = -\mathbb{1}_{\mathcal{Y}_0}(y_0) = -1$, this implies $(\mathbb{1}_{C_{0m}^C})(y_0) = 0$. Then notice that

$$J(\varphi^{cc}, \varphi^c) = J(\mathbb{1}_{C_{1m}} + 1, -1) = J(\mathbb{1}_{C_{1m}}, 0)$$

□

Remark 1.9.4. Compare theorem 1.9.39 and lemma 1.9.40 with Villani (2003) theorem 1.27.

Lemma 1.9.41 (Convex C implies c-concave functions defined with convex sets)**.** *Let C be a nonempty, open, convex subset of $\mathcal{Y}_1 \times \mathcal{Y}_0$, and $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ given by $c(y_1, y_0) = \mathbb{1}_C(y_1, y_0)$. Let $A \subseteq \mathcal{Y}_1$ be nonempty.*

1. *A^C equals $\mathcal{Y}_0 \setminus B$ for some convex set B.*

2. *If $A^C \neq \mathcal{Y}_0$, then A^{CC} is convex.*

3. *C_{1m} is convex.*

Proof. For claim 1, notice that

$$\begin{aligned} A^C &= \left\{y_0 \in \mathcal{Y}_0 \,;\, \exists y_1 \in A,\ (y_1, y_0) \in \left(\mathcal{Y}_1 \times \mathcal{Y}_0\right) \setminus C\right\} \\ &= \bigcup_{y_1 \in A} \left\{y_0 \in \mathcal{Y}_0 \,;\, (y_1, y_0) \in \left(\mathcal{Y}_1 \times \mathcal{Y}_0\right) \setminus C\right\} \\ &= \bigcup_{y_1 \in A} \mathcal{Y}_0 \setminus \left\{y_0 \in \mathcal{Y}_0 \,;\, (y_1, y_0) \in C\right\} \\ &= \mathcal{Y}_0 \setminus \bigcap_{y_1 \in A} \left\{y_0 \in \mathcal{Y}_0 \,;\, (y_1, y_0) \in C\right\} \end{aligned}$$

Since C is convex, $\{y \in \mathcal{Y}_0 \,;\, (y_1, y_0) \in C\}$ is also convex for any y_1. The intersection of an arbitrary collection of convex sets is convex, so $A^C = \mathcal{Y}_0 \setminus B$ for some convex B.

Consider claim 2 next. Notice that

$$A^{CC} = \left\{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0 \setminus A^C,\ (y_1, y_0) \in C\right\} = \bigcap_{y_0 \in \mathcal{Y}_0 \setminus A^C} \left\{y_1 \in \mathcal{Y}_1 \,;\, (y_1, y_0) \in C\right\}$$

Since C is convex, $\{y_1 \in \mathcal{Y}_1 \,;\, (y_1, y_0) \in C\}$ is convex as well, and thus A^{CC} is convex.

Finally, we show claim 3. Similar to A^{CC}, notice that

$$C_{1m} = \left\{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ (y_1, y_0) \in C\right\} = \bigcap_{y_0 \in \mathcal{Y}_0} \left\{y_1 \in \mathcal{Y}_1 \,;\, (y_1, y_0) \in C\right\}$$

is the intersection of convex sets and therefore convex. □

Refer to the convex subsets of $\mathbb{R}$ as *intervals*; specifically, $I \subset \mathbb{R}$ is called an interval if I

takes the form

$$(\ell, u) \qquad [\ell, u) \qquad (\ell, u] \qquad [\ell, u]$$

where $\ell = -\infty$ is allowed for (ℓ, u) and $(\ell, u]$ and $u = \infty$ is allowed for (ℓ, u) and $[\ell, u)$. I^c is the complement of the interval I.

Lemma 1.9.42 is relevant when the cost function is $c(y_1, y_0) = \mathbb{1}\{(y_1, y_0) \in C\}$ for some nonempty, open, convex $C \subseteq \mathcal{Y}_1 \times \mathcal{Y}_0$. When this is so, define

$$\mathcal{F}_c = \{\varphi : \mathcal{Y}_1 \to \mathbb{R} \; ; \; \varphi(y_1) = \mathbb{1}_I(y_1) \text{ for some interval } I\} \tag{1.93}$$

$$\mathcal{F}_c^c = \{\psi : \mathcal{Y}_0 \to \mathbb{R} \; ; \; \psi(y_0) = -\mathbb{1}_{I^c}(y_0) \text{ for some interval } I\} \tag{1.94}$$

Lemma 1.9.42 (Strong duality for indicator cost functions of a convex set)**.** *Let $\mathcal{Y}_1, \mathcal{Y}_0 \subseteq \mathbb{R}$, $C \subseteq \mathcal{Y}_1 \times \mathcal{Y}_0$ be nonempty, open, and convex, and let $c : \mathcal{Y}_1 \times \mathcal{Y}_0 \to \mathbb{R}$ be given by $c(y_1, y_0) = \mathbb{1}_C(y_1, y_0)$. Let $\mathcal{F}_c$ and $\mathcal{F}_c^c$ be given by* (1.93) *and* (1.94) *respectively. Then strong duality holds:*

$$\inf_{\pi \in \Pi(P_1, P_0)} \int \mathbb{1}_C(y_1, y_0) d\pi(y_1, y_0) = \sup_{(\varphi, \psi) \in \Phi_c \cap \left(\mathcal{F}_c \times \mathcal{F}_c^c\right)} \int \varphi(y_1) dP_1(y_1) + \int \psi(y_0) dP_0(y_0) \tag{1.95}$$

Proof. Recall that theorem 1.9.39 shows

$$\inf_{\pi \in \Pi(P_1, P_0)} \int \mathbb{1}_C(y_1, y_0) d\pi(y_1, y_0) = \sup_{(A,B) \in \Phi_c^I} \int \mathbb{1}_A(y_1) dP_1(y_1) - \int \mathbb{1}_B(y_0) d\nu(y_0)$$

where

$$\Phi_c^I = \Big\{(A, B) \; ; \; A \subset \mathcal{Y}_1 \text{ is closed and nonempty, } B \subset \mathcal{Y}_0 \text{ is measurable,} \\ \text{and } \mathbb{1}_A(y_1) - \mathbb{1}_B(y_0) \leq \mathbb{1}_C(y_1, y_0)\Big\}$$

Next apply lemma 1.9.34. Let $\varphi(y_1) = \mathbb{1}_A(y_1)$ for some closed and nonempty $A \subset \mathcal{Y}_1$. There are two possibilities:

1. $A^C = \mathcal{Y}_0$, in which case $J(\varphi^{cc}, \varphi^c) = J(\mathbb{1}_{C_{1m}}, 0)$, or
2. $A^C \neq \mathcal{Y}_0$, in which case $J(\varphi^{cc}, \varphi^c) = J(\mathbb{1}_{A^{CC}}, -\mathbb{1}_{A^C})$.

Since C is convex, C_{1m}, and A^{CC} are convex subsets of $\mathbb{R}$ (i.e., intervals), as shown in lemma 1.9.41. A^C is the complement of an interval, and $0 = \mathbb{1}_{\varnothing}(y_0)$ is the indicator of the complement of $\mathbb{R}$, which is the interval $(-\infty, \infty)$. Since all functions involved are bounded, they are all integrable, and lemma 1.9.34 implies

$$\inf_{\pi \in \Pi(P_1, P_0)} \int \mathbb{1}_C(y_1, y_0) d\pi(y_1, y_0) = \sup_{(\varphi,\psi) \in \Phi_c \cap \left(\mathcal{F}_c(\Phi_c^I) \times \mathcal{F}_c^c(\Phi_c^I)\right)} \int \varphi(y_1) dP_1(y_1) + \int \psi(y_0) dP_0(y_0)$$

Finally, note that $\mathcal{F}_c(\Phi_c^I) \subseteq \mathcal{F}_c$ and $\mathcal{F}_c^c(\Phi_c^I) \subseteq \mathcal{F}_c^c$, which implies the strong duality claim in display (1.95) holds. □

1.9.5.6 Special cases: $c_L(y_1, y_0) = \mathbb{1}\{y_1 - y_0 < \delta\}$ and $c_H(y_1, y_0) = \mathbb{1}\{y_1 - y_0 > \delta\}$

Lemma 1.9.43. *Let $F_1(y) = P_1(Y_1 \leq y) = \int \mathbb{1}\{y_1 \leq y\} dP_1(y_1)$ denote the cumulative distribution function (CDF) of P_1, and let F_0 the CDF of P_0. Let $c_L(y_1, y_0) = \mathbb{1}\{y_1 - y_0 < \delta\}$. Then*

$$\begin{aligned} OT_{c_L}(P_1, P_0) &= \inf_{\pi \in \Pi(P_1, P_0)} \int \mathbb{1}\{y_1 - y_0 < \delta\} d\pi(y_1, y_0) \\ &= \max\left\{ \sup_y \{F_1(y) - F_0(y - \delta)\}, P_1(Y_1 < \min\{\mathcal{Y}_0\} + \delta) \right\} \end{aligned} \tag{1.96}$$

Proof. Let $C = \{y_1 - y_0 < \delta\}$. Apply theorem 1.9.39 and lemma 1.9.40 to find that

$$OT_{c_L}(P_1, P_0) = \max\{\sup_{A \in \mathcal{A}} P_1(Y_1 \in A^{CC}) - P_0(Y_0 \in A^C), P_1(Y_1 \in C_{1m})\}$$

where

$$A^C = \{y_0 \in \mathcal{Y}_0 \,;\, \exists y_1 \in A,\ (y_1, y_0) \notin C\}, \quad A^{CC} = \left\{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0 \setminus A^C,\ (y_1, y_0) \in C\right\},$$
$$C_{1m} = \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ (y_1, y_0) \in C\}.$$

and $\mathcal{A}$ is the collection of closed, nonempty subsets of $\mathcal{Y}_1$ such that $A^C \neq \mathcal{Y}_0$.

First consider $\sup_{A\in\mathcal{A}} P_1(Y_1 \in A^{CC}) - P_0(Y_0 \in A^C)$. Let $A \in \mathcal{A}$ and $\varphi(y_1) = \mathbb{1}_A(y_1)$. Thus

$$A^C = \{y \in \mathcal{Y}_0 \,;\, \exists y_1 \in A,\ (y_1, y_0) \notin C\} = \{y_0 \in \mathcal{Y}_0 \,;\, y_0 \leq \max\{A\} - \delta\},$$
$$A^{CC} = \left\{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0 \setminus A^C,\ y_1 - y_0 < \delta\right\} = \{y_1 \in \mathcal{Y}_1 \,;\, y_1 \leq \max\{A\}\}$$

where we've used the fact that $A^C \neq \mathcal{Y}_0$ implies $\sup\{A\} < \infty$ and so $\sup\{A\} = \max\{A\}$ because A is closed. Therefore

$$\begin{aligned} J(\varphi^{cc}, \varphi^c) &= P_1(Y_1 \in A^{CC}) - P_0(Y_0 \in A^c) \\ &= P_1(Y_1 \leq \max\{A\}) - P_0(Y_0 \leq \max\{A\} - \delta) \end{aligned}$$

which takes the form $F_1(y) - F_0(y - \delta)$ for $y = \max\{A\}$.

Now consider $P_1(Y_1 \in C_{1m})$, and notice that

$$\begin{aligned} C_{1m} &= \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ (y_1, y_0) \in C\} = \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ y_1 - y_0 < \delta\} \\ &= \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ y_1 < \min\{\mathcal{Y}_0\} + \delta\} \end{aligned}$$

Thus $P_1(Y_1 \in C_{1m}) = P_1(Y_1 < \min\{\mathcal{Y}_0\} + \delta)$. The result follows. □

Remark 1.9.5. C_{1m} may be closed; e.g., let $\mathcal{Y}_1 = [0, 1] \cup [3, 10]$, let $\mathcal{Y}_0 = [2, 10]$, and $\delta = 0$. Then $C_{1m} = \{y_1 \in \mathcal{Y}_1 \,;\, y_1 < 2\} = [0, 1]$.

Corollary 1.9.44. *Let $c_L(y_1, y_0) = \mathbb{1}\{y_1 - y_0 < \delta\}$ and P_1, P_0 have continuous cumulative distribution functions $F_1(y) = P_1(Y_1 \leq y)$ and $F_0(y) = P_0(Y_0 \leq y)$ respectively. Then*

$$OT_{c_L}(P_1, P_0) = \inf_{\pi \in \Pi(P_1, P_0)} \int \mathbb{1}\{y_1 - y_0 < \delta\} d\pi(y_1, y_0) = \sup_y \{F_1(y) - F_0(y - \delta)\} \tag{1.97}$$

Proof. Continuity of the cumulative distribution functions implies $P_1(Y_1 = \delta + \min\{\mathcal{Y}_0\}) = P_0(Y_0 = \min\{\mathcal{Y}_0\}) = 0$, and thus

$$P_1(Y_1 < \delta + \min\{\mathcal{Y}_0\}) = P_1(Y_1 \leq \delta + \min\{\mathcal{Y}_0\}) - P_0(Y_0 \leq \min\{\mathcal{Y}_0\})$$

Which takes the form $F_1(y) - F_0(y - \delta)$ for $y = \delta + \min\{\mathcal{Y}_0\}$. It follows that

$$\max\left\{\sup_y \{F_1(y) - F_0(y - \delta)\}, P_1(Y_1 < \min\{\mathcal{Y}_0\} + \delta)\right\} = \sup_y \{F_1(y) - F_0(y - \delta)\}$$

and lemma 1.9.43 gives the result. □

Lemma 1.9.45. *Let $c_H(y_1, y_0) = \mathbb{1}\{y_1 - y_0 > \delta\}$. Then*

$$\begin{aligned} OT_{c_H}(P_1, P_0) &= \inf_{\pi \in \Pi(P_1, P_0)} \int \mathbb{1}\{y_1 - y_0 > \delta\} d\pi(y_1, y_0) \\ &= \max\left\{\sup_y \{P_1([y, \infty)) - P_0([y - \delta, \infty))\}, P_1((\max\{\mathcal{Y}_0\} + \delta, \infty))\right\} \end{aligned} \tag{1.98}$$

Proof. The proof is similar to that of lemma 1.9.43. Let $C = \{y_1 - y_0 > \delta\}$. Apply theorem 1.9.39 and lemma 1.9.40 to find that

$$OT_{c_L}(P_1, P_0) = \max\{\sup_{A \in \mathcal{A}} P_1(Y_1 \in A^{CC}) - P_0(Y_0 \in A^C), P_1(Y_1 \in C_{1m})\}$$

where

$$A^C = \{y_0 \in \mathcal{Y}_0 \,;\, \exists y_1 \in A,\ (y_1, y_0) \notin C\}, \quad A^{CC} = \left\{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0 \setminus A^C,\ (y_1, y_0) \in C\right\},$$
$$C_{1m} = \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ (y_1, y_0) \in C\}.$$

and $\mathcal{A}$ is the collection of closed, nonempty subsets of $\mathcal{Y}_1$ such that $A^C \neq \mathcal{Y}_0$.

Consider $\sup_{A\in\mathcal{A}} P_1(Y_1 \in A^{CC}) - P_0(Y_0 \in A^C)$. Let $A \in \mathcal{A}$ and $\varphi(y_1) = \mathbb{1}_A(y_1)$, and notice that

$$A^C = \{y \in \mathcal{Y}_0 \,;\, \exists y_1 \in A,\ (y_1, y_0) \notin C\} = \{y_0 \in \mathcal{Y}_0 \,;\, y_0 \geq \min\{A\} - \delta\},$$
$$A^{CC} = \left\{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0 \setminus A^C,\ y_1 - y_0 < \delta\right\} = \{y_1 \in \mathcal{Y}_1 \,;\, y_1 \geq \min\{A\}\}$$

Where as in the proof of lemma 1.9.43, $A^C \neq \mathcal{Y}_0$ implies $\inf\{A\} > -\infty$ and so $\inf\{A\} = \min\{A\}$ because A is closed. Thus

$$\begin{aligned} J(\varphi^{cc}, \varphi^c) &= P_1(Y_1 \in A^{CC}) - P_0(Y_0 \in A^c) \\ &= P_1(Y_1 \geq \min\{A\}) - P_0(Y_0 \geq \min\{A\} - \delta) \end{aligned}$$

which takes the form $P_1([y, \infty)) - P_0([y - \delta, \infty))$ for $y = \min\{A\}$.

Now consider $P_1(Y_1 \in C_{1m})$, and notice that

$$\begin{aligned} C_{1m} &= \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ (y_1, y_0) \in C\} = \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ y_1 - y_0 > \delta\} \\ &= \{y_1 \in \mathcal{Y}_1 \,;\, \forall y_0 \in \mathcal{Y}_0,\ y_1 > \max\{\mathcal{Y}_0\} + \delta\} \end{aligned}$$

Thus $P_1(Y_1 \in C_{1m}) = P_1(Y_1 > \max\{\mathcal{Y}_0\} + \delta)$. The result follows. □

Corollary 1.9.46. *Let $c_H(y_1, y_0) = \mathbb{1}\{y_1 - y_0 > \delta\}$ and P_1, P_0 have continuous cumulative*

distribution functions $F_1(y) = P_1(Y_1 \leq y)$ *and* $F_0(y) = P_0(Y_0 \leq y)$ *respectively. Then*

$$OT_{c_L}(P_1, P_0) = \inf_{\pi \in \Pi(P_1, P_0)} \int \mathbb{1}\{y_1 - y_0 > \delta\} d\pi(y_1, y_0) = \sup_y \{F_0(y - \delta) - F_1(y)\} \tag{1.99}$$

Proof. Continuity of the cumulative distribution functions implies that for any y,

$$\begin{aligned} P_1([y, \infty)) - P_0([y - \delta, \infty)) &= P_1((y, \infty)) - P_0((y - \delta, \infty)) \\ &= (1 - F_1(y)) - (1 - F_0(y - \delta)) \\ &= F_0(y - \delta) - F_1(y) \end{aligned}$$

and furthermore,

$$\begin{aligned} P_1(Y_1 > \delta + \max\{\mathcal{Y}_0\}) &= 1 - F_1(\delta + \min\{\mathcal{Y}_0\}) - (1 - F_0(\max\{\mathcal{Y}_0\}) \\ &= F_0(\max\{\mathcal{Y}_0\}) - F_1(\delta + \max\{\mathcal{Y}_0\}) \end{aligned}$$

equals $F_0(y - \delta) - F_1(y)$ for $y = \max\{\mathcal{Y}_0\} + \delta$. Finally, lemma 1.9.45 gives

$$\begin{aligned} OT_{c_H}(P_1, P_0) &= \max \left\{ \sup_y \{P_1([y, \infty)) - P_0([y - \delta, \infty))\}, P_1((\max\{\mathcal{Y}_0\} + \delta, \infty)) \right\} \\ &= \sup_y \{F_0(y - \delta) - F_1(y)\} \end{aligned}$$

□

1.9.6 Appendix: miscellaneous lemmas

1.9.6.1 Continuity

Lemma 1.9.47 (Continuity of maps between bounded function spaces)**.** *Let* $f : \mathbb{D}_f \subseteq \mathbb{R}^K \to \mathbb{R}^M$ *be uniformly continuous. Define the subset of bounded functions on* T *taking values in*

$\mathbb{D}_f$:

$$\ell^\infty(T,\mathbb{D}_f)=\left\{g:T\to\mathbb{R}^K\ ;\ g(t)\in\mathbb{D}_f,\ \sup_{t\in T}\|g(t)\|<\infty\right\}\subseteq\ell^\infty(T)^K$$

Let $F:\ell^\infty(T,\mathbb{D}_f)\to\ell^\infty(T)^M$ be defined pointwise as $F(g)(t)=f(g(t))$. Then F is uniformly continuous.

Proof. To see that $F:\ell^\infty(T,\mathbb{D}_f)\to\ell^\infty(T)^M$ is well defined, recall that uniform continuity of f implies f is bounded on bounded sets. Since $\{g(t)\ ;\ t\in T\}$ is bounded for any $g\in\ell^\infty(T,\mathbb{D}_f)$, this implies $\sup_t\|f(g(t))\|<\infty$ and hence $F(g)\in\ell^\infty(T)^M$.

To see uniform continuity of F, let $\varepsilon>0$ and use uniform continuity of f to choose $\delta>0$ such that for all $x,\tilde{x}\in\mathbb{D}_f$,

$$\|x-\tilde{x}\|<\delta\implies\|f(x)-f(\tilde{x})\|<\varepsilon/2$$

Now let $g,\tilde{g}\in\ell^\infty(T,\mathbb{D}_f)$ satisfy $\|g-\tilde{g}\|_T=\sup_{t\in T}\|g(t)-\tilde{g}(t)\|<\delta$. Then $\|g(t)-\tilde{g}(t)\|<\delta$ for all $t\in T$, and hence $\|f(g(t))-f(\tilde{g}(t))\|<\varepsilon/2$ for all $t\in T$, and therefore

$$\|F(g)-F(\tilde{g})\|_T=\sup_{t\in T}\|f(g(t))-f(\tilde{g}(t))\|\le\frac{\varepsilon}{2}<\varepsilon$$

which completes the proof. □

Corollary 1.9.48. *Let $f:\mathbb{D}_f\subseteq\mathbb{R}^K\to\mathbb{R}^M$ be continuous and bounded on bounded subsets of $\mathbb{D}_f$. Let $g_0\in\ell^\infty(T,\mathbb{D}_f)$ where $\ell^\infty(T,\mathbb{D}_f)$ is as defined in lemma 1.9.47. Suppose that for some $\delta>0$,*

$$g(T)^\delta\equiv\left\{x\in\mathbb{R}^K\ ;\ \inf_{t\in T}\|g_0(t)-x\|\le\delta\right\}$$

is a subset of $\mathbb{D}_f$. Then $F:\ell^\infty(T,\mathbb{D}_f)\to\ell^\infty(T)^M$ defined pointwise by $F(g)(t)=f(g(t))$ is continuous at g_0.

Proof. For any $g\in\ell^\infty(T,\mathbb{D}_f)$, we have $F(g)\in\ell^\infty(T)^M$ because $\{x\ ;\ x=g(t)$ for some $t\in$

$T\}$ is bounded and f is bounded on bounded subsets.

Let $\{g_n\}_{n=1}^{\infty} \subseteq \ell^\infty(T, \mathbb{D}_f)$ be such that $g_n \to g_0$ in $\ell^\infty(T)^K$. It suffices to show that $F(g_n) \to F(g_0)$ in $\ell^\infty(T)^M$. Let $\tilde{f} : g(T)^\delta \to \mathbb{R}^M$ be the restriction of f to $g_0(T)^\delta$; i.e., $\tilde{f}(x) = f(x)$. Note that because $g_0(T)^\delta$ is a closed and bounded subset of $\mathbb{R}^K$, it is compact, and hence $\tilde{f}$ is uniformly continuous by the Heine-Cantor theorem. Apply lemma 1.9.47 to find that

$$\tilde{F} : \ell^\infty(T, g(T)^\delta) \to \ell^\infty(T)^M, \qquad \tilde{F}(g)(t) = \tilde{f}(g(t)) = f(g(t))$$

is continuous. Since $g_n \to g_0$ in $\ell^\infty(T)^K$, there exists N such that for all $n \geq N$, $\|g_n - g_0\|_T = \sup_{t\in T}\|g_n(t) - g_0(t)\| < \delta$. Let $\tilde{g}_k = g_{k+N}$. Notice that

$$\tilde{g}_k(T) = \left\{x \in \mathbb{R}^K \ ; \ x = g_k(t) \text{ for some } t \in T\right\} \subseteq g_0(T)^\delta,$$

and hence $\tilde{g}_k \in \ell^\infty(T, g_0(T)^\delta)$. Continuity of $\tilde{F}$ and $\tilde{g}_k \to g_0$ implies $\tilde{F}(\tilde{g}_k) \to \tilde{F}(\tilde{g}_0)$. Thus

$$0 = \lim_{k\to\infty} \|\tilde{F}(\tilde{g}_k) - \tilde{F}(g_0)\|_T = \lim_{k\to\infty} \|F(g_{k+N}) - F(g_0)\|_T = \lim_{n\to\infty} \|F(g_n) - F(g_0)\|_T$$

which completes the proof. □

Lemma 1.9.49 (Uniform continuity of restricted sup)**.** *For any set X, subset $A \subseteq X$, and bounded real-valued functions $f, g \in \ell^\infty(X)$,*

$$\left|\sup_{x\in A} f(x) - \sup_{x\in A} g(x)\right| \leq \sup_{x\in A} |f(x) - g(x)| \tag{1.100}$$

and therefore $\sigma_A : \ell^\infty(X) \to \mathbb{R}$ given by $\sigma_A(f) = \sup_{x\in A} f(x)$ is uniformly continuous.

Proof. Observe that

$$\sup_{x\in A} f(x) - \sup_{x\in A} g(x) \leq \sup_{x\in A}\{f(x) - g(x)\} \leq \sup_{x\in A}|f(x) - g(x)|$$

and

$$-\left[\sup_{x\in A} f(x) - \sup_{x\in A} g(x)\right] = \sup_{x\in A} g(x) - \sup_{x\in A} f(x) \leq \sup_{x\in A}\{g(x) - f(x)\} \leq \sup_{x\in A}|f(x) - g(x)|$$

Together these inequalities imply

$$-\sup_{x\in A}|f(x) - g(x)| \leq \sup_{x\in A} f(x) - \sup_{x\in A} g(x) \leq \sup_{x\in A}|f(x) - g(x)|$$

which is equivalent to (1.100).

To see uniform continuity, let $\varepsilon > 0$ and choose $\delta = \varepsilon$. Whenever $\|f-g\|_X = \sup_{x\in X}|f(x)-g(x)| < \delta$,

$$|\sigma_A(f) - \sigma_A(g)| = \left|\sup_{x\in A} f(x) - \sup_{x\in A} g(x)\right| \leq \sup_{x\in A}|f(x) - g(x)| \leq \sup_{x\in X}|f(x) - g(x)| < \delta = \varepsilon$$

which completes the proof. □

1.9.6.2 Differentiability

This appendix reviews definitions and various facts related to Hadamard directional differentiability. The following definitions can be found in Fang & Santos (2019).

Let $\mathbb{D}$, $\mathbb{E}$ be Banach spaces (complete, normed, vector spaces), and $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$.

(i) ϕ is (fully) *Hadamard differentiable* at $x_0 \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$ if there exists

a continuous linear map $\phi'_{x_0} : \mathbb{D}_0 \to \mathbb{E}$ such that

$$\lim_{n\to\infty} \left\| \frac{\phi(x_0 + t_n h_n) - \phi(x_0)}{t_n} - \phi'_{x_0}(h) \right\|_{\mathbb{E}} = 0$$

for all sequences $\{h_n\}_{n=1}^{\infty} \subseteq \mathbb{D}$ and $\{t_n\}_{n=1}^{\infty} \subseteq \mathbb{R}$ such that $h_n \to h \in \mathbb{D}_0$ and $t_n \to 0$ as $n \to \infty$, and $x_0 + t_n h_n \in \mathbb{D}_\phi$ for all n.

(ii) ϕ is *Hadamard directionally differentiable* at $x_0 \in \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$ if there exists a continuous map $\phi'_{x_0} : \mathbb{D}_0 \to \mathbb{E}$ such that

$$\lim_{n\to\infty} \left\| \frac{\phi(x_0 + t_n h_n) - \phi(x_0)}{t_n} - \phi'_{x_0}(h) \right\|_{\mathbb{E}} = 0$$

for all sequences $\{h_n\}_{n=1}^{\infty} \subseteq \mathbb{D}$ and $\{t_n\}_{n=1}^{\infty} \subseteq \mathbb{R}_+$ such that $h_n \to h \in \mathbb{D}_0$ and $t_n \downarrow 0$ as $n \to \infty$, and $x_0 + t_n h_n \in \mathbb{D}_\phi$ for all n.

Fang & Santos (2019) proposition 2.1 shows that linearity is the key property distinguishing directional and full Hadamard differentiability. Specifically, if ϕ is Hadamard directionally differentiable at x_0 tangentially to a subspace $\mathbb{D}_0$, and ϕ'_{x_0} is linear, then ϕ is in fact fully Hadamard differentiable at x_0 tangentially to $\mathbb{D}_0$.

Hadamard directional differentiability obeys the chain rule.

Lemma 1.9.50 (Chain rule)**.** *Let $\mathbb{D}_1$, $\mathbb{D}_2$, and $\mathbb{E}$ be Banach spaces and $\phi_1 : \mathbb{D}_{\phi_1} \subseteq \mathbb{D}_1 \to \mathbb{D}_2$, $\phi_2 : \mathbb{D}_{\phi_2} \subseteq \mathbb{D}_2 \to \mathbb{E}$ be functions. Suppose*

(i) $\phi_1(\mathbb{D}_{\phi_1}) = \{y \in \mathbb{D}_2 \,;\, y = \phi_1(x) \text{ for some } x \in \mathbb{D}_{\phi_1}\} \subseteq \mathbb{D}_{\phi_2}$,

(ii) ϕ_1 *is Hadamard directionally differentiable at* $x_0 \in \mathbb{D}_{\phi_1}$ *tangentially to* $\mathbb{D}_1^T \subseteq \mathbb{D}_1$*, with derivative* $\phi'_{1,x_0}(h)$*, and*

(iii) ϕ_2 *is Hadamard directionally differentiable at* $\phi_1(x_0) \in \mathbb{D}_{\phi_2}$ *tangentially to* $\mathbb{D}_2^T \subseteq \mathbb{D}_2$*, with derivative* $\phi'_{2,\phi_1(x_0)}(h)$

Let $\mathbb{D}^T = \left\{x \in \mathbb{D}_1^T \,;\, \phi'_{1,x_0}(x) \in \mathbb{D}_2^T\right\}$. *The composition function*

$$\phi : \mathbb{D}_{\phi_1} \to \mathbb{E}, \qquad\qquad \phi(x) = \phi_2(\phi_1(\theta))$$

is Hadamard directionally differentiable at x_0 *tangentially to* $\mathbb{D}^T$, *with*

$$\phi'_{x_0} : \mathbb{D}^T \to \mathbb{E}, \qquad\qquad \phi'_{x_0}(h) = \phi'_{2,\phi_1(x_0)}(\phi'_{1,x_0}(h))$$

Proof. That ϕ is well defined is clear from assumption (i). To show its Hadamard directional differentiability, let $\{h_n\}_{n=1}^{\infty} \subseteq \mathbb{D}_{\phi_1}$ and $\{t_n\}_{n=1}^{\infty} \subseteq \mathbb{R}_+$ be such that $h_n \to h \in \mathbb{D}^T$, $t_n \downarrow 0$, and $x_0 + t_n h_n \in \mathbb{D}_{\phi_1}$ for all n. Assumption (ii) implies that

$$\lim_{n\to\infty} \left\| \frac{\phi_1(x_0 + t_n h_n) - \phi_1(x_0)}{t_n} - \phi'_{1,x_0}(h) \right\|_{\mathbb{D}_2} = 0 \tag{1.101}$$

Let $g_n = \frac{1}{t_n}\left[\phi_1(x_0 + t_n h_n) - \phi_1(x_0)\right]$, $g = \phi'_{1,x_0}(h)$, and notice that (1.101) implies $g_n \to g$ in $\mathbb{D}_2$.

Assumption (i) implies $\phi_1(x_0) + t_n g_n = \phi_1(x_0 + t_n h_n) \in \mathbb{D}_{\phi_2}$ for each n, and the definition of $\mathbb{D}^T$ implies $g \in \mathbb{D}_2^T$. Assumption (iii) implies that

$$\lim_{n\to\infty} \left\| \frac{\phi_2(\phi_1(x_0) + t_n g_n) - \phi_2(\phi_1(x_0))}{t_n} - \phi'_{2,\phi_1(x_0)}(g) \right\|_{\mathbb{E}} = 0 \tag{1.102}$$

Substitute $\phi_2(\phi_1(x_0) + t_n g_n) = \phi_2(\phi_1(x_0 + t_n h_n))$, and $g = \phi'_{1,x_0}(h)$, into (1.102) to find

$$\lim_{n\to\infty} \left\| \frac{\phi_2(\phi_1(x_0 + t_n h_n)) - \phi_2(\phi_1(x_0))}{t_n} - \phi'_{2,\phi_1(x_0)}(\phi'_{1,x_0}(h)) \right\|_{\mathbb{E}} = 0$$

which completes the proof. □

Remark 1.9.6. When defining and differentiating composition of functions, the outer function's properties determine restrictions that must be placed on the inner function to ensure

the composition function is well defined and differentiable.

A familiar example of this is that the domain of the "inner function" ϕ_1 may need to be restricted to ensure the composition map is well defined. For a simple example, x^3 is well defined and differentiable for any $x \in \mathbb{R}$, but $\log(x^3)$ is only well defined (and differentiable) for $x \in (0, \infty)$.

A less familiar example shows up only when considering Hadamard differentiability tangentially to a set. The tangent spaces of each function jointly determine the tangent space of the derivative of the composition map.

The next lemma shows that Hadamard directionally differentiable functions can be "stacked".

Lemma 1.9.51 (Stacking Hadamard differentiable functions)**.** *Let $\mathbb{D}$, $\mathbb{E}_1$, and $\mathbb{E}_2$ be Banach spaces, and $\mathbb{D}_\phi \subseteq \mathbb{D}$. Suppose $\phi^{(1)} : \mathbb{D}_\phi \to \mathbb{E}_1$ and $\phi^{(2)} : \mathbb{D}_\phi \to \mathbb{E}_2$ are Hadamard directionally differentiable tangentially to $\mathbb{D}_0 \subseteq \mathbb{D}$ at $x_0 \in \mathbb{D}_\phi$ with derivatives $\phi_{x_0}^{(1)\prime} : \mathbb{D}_0 \to \mathbb{E}_1$ and $\phi_{x_0}^{(2)\prime} : \mathbb{D}_0 \to \mathbb{E}_2$. Define*

$$\phi : \mathbb{D}_\phi \to \mathbb{E}_1 \times \mathbb{E}_2, \qquad \phi(x) = \left(\phi^{(1)}(x), \quad \phi^{(2)}(x)\right)$$

Then ϕ is Hadamard directionally differentiable tangentially to $\mathbb{D}_0$ at x_0, with derivative

$$\phi'_{x_0} : \mathbb{D}_0 \to \mathbb{E}_1 \times \mathbb{E}_2, \qquad \phi'_{x_0}(h) = \left(\phi_{x_0}^{(1)\prime}(h), \quad \phi_{x_0}^{(2)\prime}(h)\right)$$

Proof. Hadamard directional differentiability of $\phi^{(1)}$ and $\phi^{(2)}$ tangentially to $\mathbb{D}_0$ at x_0 implies that for any sequences $\{h_n\}_{n=1}^\infty \subseteq \mathbb{D}$ and $\{t_n\} \subseteq \mathbb{R}_+$ such that $h_n \to h \in \mathbb{D}_0$, $t_n \downarrow 0$, and

$x_0 + t_n h_n \in \mathbb{D}_\phi$ for all n,

$$\lim_{n\to\infty} \left\| \frac{\phi^{(1)}(x_0 + t_n h_n) - \phi^{(1)}(x_0)}{t_n} - \phi^{(1)\prime}_{x_0}(h) \right\|_{\mathbb{E}_1} = 0, \text{ and}$$
$$\lim_{n\to\infty} \left\| \frac{\phi^{(2)}(x_0 + t_n h_n) - \phi^{(2)}(x_0)}{t_n} - \phi^{(2)\prime}_{x_0}(h) \right\|_{\mathbb{E}_2} = 0$$

Since $\|(e_1, e_2) - (\tilde{e}_1, \tilde{e}_2)\|_{\mathbb{E}_1\times\mathbb{E}_2} = \|e_1 - \tilde{e}_1\|_{\mathbb{E}_1} + \|e_2 - \tilde{e}_2\|_{\mathbb{E}_2}$ metricizes $\mathbb{E}_1 \times \mathbb{E}_2$ (Aliprantis & Border (2006) lemma 3.3), we have

$$\begin{aligned}
&\left\| \frac{\phi(x_0 + t_n h_n) - \phi(x_0)}{t_n} - \phi'_{x_0}(h) \right\|_{\mathbb{E}_1\times\mathbb{E}_2} \\
&= \left\| \frac{\left(\phi^{(1)}(x_0 + t_n h_n), \quad \phi^{(2)}(x_0 + t_n h_n)\right) - \left(\phi^{(1)}(x_0), \quad \phi^{(2)}(x_0)\right)}{t_n} - \left(\phi^{(1)}_{x_0}(h), \quad \phi^{(2)}_{x_0}(h)\right) \right\|_{\mathbb{E}_1\times\mathbb{E}_2} \\
&= \left\| \left(\tfrac{\phi^{(1)}(x_0+t_n h_n) - \phi^{(1)}(x_0)}{t_n} - \phi^{(1)\prime}_{x_0}(h), \quad \tfrac{\phi^{(2)}(x_0+t_n h_n) - \phi^{(2)}(x_0+t_n h_n)}{t_n} - \phi^{(2)\prime}_{x_0} \right) \right\|_{\mathbb{E}_1\times\mathbb{E}_2} \\
&= \left\| \frac{\phi^{(1)}(x_0 + t_n h_n) - \phi^{(1)}(x_0)}{t_n} - \phi^{(1)\prime}_{x_0}(h) \right\|_{\mathbb{E}_1} + \left\| \frac{\phi^{(2)}(x_0 + t_n h_n) - \phi^{(2)}(x_0)}{t_n} - \phi^{(2)\prime}_{x_0}(h) \right\|_{\mathbb{E}_2}
\end{aligned}$$

Taking the limit as $n \to \infty$ gives the result. □

Hadamard differentiability in bounded function spaces

It is common to "rearrange" Donsker sets; i.e. view them not as scalar-valued but vector-valued with each coordinate occuring over a particular subset of functions (see van der Vaart (2007) p. 270). The following lemma shows that one direction of the equivalence can be viewed as an application of the delta method.

Lemma 1.9.52 (Rearranging Donsker sets)**.** *Suppose $\mathcal{F} = \mathcal{F}_1 \cup \ldots \cup \mathcal{F}_K$ is P-Donsker, and $\sqrt{n}(\mathbb{P}_n - P) \xrightarrow{L} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$. The map $\phi : \ell^\infty(\mathcal{F}) \to \ell^\infty(\mathcal{F}_1) \times \ldots \times \ell^\infty(\mathcal{F}_K)$ defined pointwise by*

$$\phi(g)(f_1, \ldots, f_K) = (g(f_1), \ldots, g(f_K))$$

is fully Hadamard differentiable at any $P \in \ell^\infty(\mathcal{F})$ tangentially to $\ell^\infty(\mathcal{F})$, and is its own derivative:

$$\phi'_P : \ell^\infty(\mathcal{F}) \to \ell^\infty(\mathcal{F}_1) \times \ldots \times \ell^\infty(\mathcal{F}_K), \qquad \phi'_P(h) = \phi(h)$$

and hence

$$\sqrt{n}(\phi(\mathbb{P}_n) - \phi(P)) \xrightarrow{L} \phi(\mathbb{G}) \qquad \text{in } \ell^\infty(\mathcal{F}_1) \times \ldots \times \ell^\infty(\mathcal{F}_K)$$

Proof. The map ϕ is linear; let $a, b \in \mathbb{R}$ and $g, h \in \ell^\infty(\mathcal{F})$ and notice that for any $(f_1, \ldots, f_K) \in \mathcal{F}_1 \times \ldots \times \mathcal{F}_K$,

$$\begin{aligned}
\phi(ag + bh)(f_1, \ldots, f_K) &= ((ag + bh)(f_1), \ldots, (ag + bh)(f_K)) \\
&= (ag(f_1) + bh(f_1), \ldots, ag(f_K) + bh(f_K)) \\
&= a(g(f_1), \ldots, g(f_K)) + b(h(f_1) \ldots, h(f_K)) \\
&= a\phi(g)(f_1, \ldots, f_K) + b\phi(h)(f_1, \ldots, f_K) \\
&= (a\phi(g) + b\phi(h))(f_1, \ldots, f_K)
\end{aligned}$$

hence $\phi(ag + bh) = (a\phi(g) + b\phi(h))$, as these functions agree on all of $\mathcal{F}_1 \times \ldots \times \mathcal{F}_K$.

Next observe that ϕ is continuous. Recall that the product topology on $\ell^\infty(\mathcal{F}_1) \times \ldots \times \ell^\infty(\mathcal{F}_K)$ is generated by the norm

$$\|(g_1, \ldots, g_K) - (h_1, \ldots, h_K)\|_{\mathcal{F}_1 \times \ldots \times \mathcal{F}_K} = \max\{\|g_1 - h_1\|_{\mathcal{F}_1}, \ldots, \|g_K - h_K\|_{\mathcal{F}_K}\};$$

see Aliprantis & Border (2006) lemma 3.3. Thus

$$\begin{aligned}\|\phi(g) - \phi(h)\|_{\mathcal{F}_1 \times \ldots \times \mathcal{F}_K} &= \max\left\{ \sup_{f_1 \in \mathcal{F}_1} |g(f_1) - h(f_1)|, \ldots, \sup_{f_K \in \mathcal{F}_K} |g(f_K) - h(f_K)| \right\} \\ &= \|g - h\|_{\mathcal{F}}\end{aligned}$$

and hence ϕ is continuous.

Since ϕ is linear and continuous, it is (fully) Hadamard differentiable at any point tangentially to $\ell^\infty(\mathcal{F})$ and is its own Hadamard derivative; indeed, for an: for all sequences $h_n \to h \in \ell^\infty(\mathcal{F})$ and $t_n \downarrow 0 \in \mathbb{R}$, one has $g + t_n h_n \in \ell^\infty(\mathcal{F})$ and

$$\lim_{n\to\infty} \left\| \frac{\phi(g + t_n h_n) - \phi(g)}{t_n} - \phi(h) \right\|_{\mathcal{F}_1 \times \ldots \times \mathcal{F}_K} = \lim_{n\to\infty} \|\phi(h_n) - \phi(h)\|_{\mathcal{F}_1 \times \ldots \times \mathcal{F}_K} = 0$$

Finally, since $\sqrt{n}(\mathbb{P}_n - P) \xrightarrow{L} \mathbb{G}$ in $\ell^\infty(\mathcal{F})$, the functional delta method (van der Vaart (2007) theorem 20.8) implies $\sqrt{n}(\phi(\mathbb{P}_n) - \phi(P)) \xrightarrow{L} \phi(\mathbb{G})$ in $\ell^\infty(\mathcal{F}_1) \times \ldots \times \ell^\infty(\mathcal{F}_K)$. □

Although the following lemma and its corollary are stated for functions taking values in $\mathbb{R}$, by combining it with lemma 1.9.51 a similar result can be obtained for functions taking values in $\mathbb{R}^M$, similar to the setting of lemma 1.9.47. Compare van der Vaart & Wellner (1997) lemma 3.9.25.

Lemma 1.9.53 (Hadamard differentiability of maps between bounded function spaces)**.** *Let* $f : \mathbb{D}_f \subseteq \mathbb{R}^K \to \mathbb{R}$. *Suppose that*

1. *f is continuously differentiable, and*
2. *the gradient of f,*

$$\nabla f : \mathbb{D}_f \to \mathbb{R}^K, \qquad \nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x) \quad \ldots \quad \frac{\partial f}{\partial x_K}(x) \right)^\intercal,$$

is uniformly continuous.

Define the subset of $\ell^\infty(T)^K$ taking values in $\mathbb{D}_f$,

$$\ell^\infty(T,\mathbb{D}_f) = \left\{ g : T \to \mathbb{R}^K \ ; \ g(t) \in \mathbb{D}_f, \ \sup_{t\in T} \|g(t)\| < \infty \right\} \subseteq \ell^\infty(T)^K$$

and the subset of $\ell^\infty(T,\mathbb{D}_f)$ such that composition with f defines a bounded function:

$$\ell_f^\infty(T,\mathbb{D}_f) = \left\{ g \in \ell^\infty(T,\mathbb{D}_f) \ ; \ \sup_{t\in T} |f(g(t))| < \infty \right\}$$

Then $F : \ell_f^\infty(T,\mathbb{D}_f) \to \ell^\infty(T)$ defined pointwise with $F(g)(t) = f(g(t))$ is (fully) Hadamard differentiable tangentially to $\ell^\infty(T)^K$ at any $g_0 \in \ell_f^\infty(T,\mathbb{D}_f)$, with derivative $F'_{g_0} : \ell^\infty(T)^K \to \ell^\infty(T)$ given pointwise by

$$F'_{g_0}(h)(t) = [\nabla f(g_0(t))]^\intercal \, h(t) = \sum_{k=1}^{K} \frac{\partial f}{\partial x_k}(g_0(t)) h_k(t)$$

Proof. The domain of $\ell_f^\infty(T,\mathbb{D}_f)$ ensures that $F : \ell_f^\infty(T,\mathbb{D}_f) \to \ell^\infty(T)$ is well defined.

Let $\{h_n\}_{n=1}^\infty \subseteq \ell^\infty(T)^K$ and $\{r_n\}_{n=1}^\infty \subseteq \mathbb{R}$ such that $h_n \to h \in \ell^\infty(T)^K$, $r_n \to 0$, and $g_0 + r_n h_n \in \ell_f^\infty(T,\mathbb{D}_f)$ for each n. For each n and each $t \in T$, apply the mean value theorem to find $\lambda_n(t) \in (0,1)$ such that $g_n(t) \coloneqq \lambda_n(t)(g_0(t) + r_n h_n(t)) + (1-\lambda_n(t))g_0(t)$ satisfying[15]

$$\begin{aligned} f(x_0(t) + r_n h_n(t)) - f(x_0(t)) &= [\nabla f(g_n(t))]^\intercal \, (x_0(t) + r_n h_n(t) - x_0(t)) \\ &= r_n \, [\nabla f(g_n(t))]^\intercal \, h_n(t) \end{aligned}$$

[15]The mean value theorem being invoked here is the standard result: for any $x, \tilde{x} \in \mathbb{D}_f$, let $g_{x,\tilde{x}} : [0,1] \to \mathbb{R}$ be given by $g_{x,\tilde{x}}(\lambda) = f(\lambda\tilde{x} + (1-\lambda)x)$. Then $g_{x,\tilde{x}}(0) = f(x)$ and $g_{x,\tilde{x}}(1) = f(\tilde{x})$, and the mean value theorem tells us that there exists $\lambda \in (0,1)$ such that

$$f(\tilde{x}) - f(x) = g_{x,\tilde{x}}(1) - g_{x,\tilde{x}}(0) = g'_{x,\tilde{x}}(\lambda)(1-0) = [\nabla f(\lambda\tilde{x} + (1-\lambda)x)]^\intercal \, (\tilde{x} - x)$$

Use this to see that for all n and all $t \in T$,

$$\begin{aligned}\left|\frac{f(g_0(t)+r_nh_n(t))-f(g_0(t))}{r_n}-\nabla f(g_0(t))^\intercal h(t)\right| &= |\nabla f(g_n(t))^\intercal h_n(t)-\nabla f(g_0(t))^\intercal h(t)| \\ &\leq |\nabla f(g_n(t))^\intercal h_n(t)-\nabla f(g_0(t))^\intercal h_n(t)|+|\nabla f(g_0(t))^\intercal h_n(t)-\nabla f(g_0(t))^\intercal h(t)| \\ &\leq \|\nabla f(g_n(t))-\nabla f(g_0(t))\|\times\|h_n(t)\|+\|\nabla f(g_0(t))\|\times\|h_n(t)-h(t)\|\end{aligned}$$

where the first inequality is by the triangle inequality and the second by Cauchy-Schwarz in $\mathbb{R}^K$. It follows that

$$\begin{aligned}&\sup_{t\in T}\left|\frac{f(g_0(t)+r_nh_n(t))-f(g_0(t))}{r_n}-\nabla f(g_0(t))^\intercal h(t)\right| \\ &\qquad\leq \sup_{t\in T}\|\nabla f(g_n(t))-\nabla f(g_0(t))\|\times\sup_{t\in T}\|h_n(t)\| &(1.103)\\ &\qquad\quad+\sup_{t\in T}\|\nabla f(g_0(t))\|\times\sup_{t\in T}\|h_n(t)-h(t)\| &(1.104)\end{aligned}$$

Consider the term in (1.103). Recall that for some $\lambda_n(t) \in (0,1)$,

$$\begin{aligned}g_n(t) &= \lambda_n(t)(g_0(t)+r_nh_n(t))+(1-\lambda_n(t))g_0(t) \\ &= \lambda_n(t)r_nh_n(t)+g_0(t)\end{aligned}$$

and so

$$\|g_n-g_0\|_T = \sup_{t\in T}\|\lambda_n(t)r_nh_n(t)\| \leq |r_n|\times\sup_{t\in T}\|h_n(t)\| \to 0$$

where the limit claim follows from $\sup_{t\in T}\|h_n(t)\| = \|h_n\|_T \to \|h\|_T < \infty$ (implying $\{\sup_{t\in T}\|h_n(t)\|\}_{n=1}^\infty$ is bounded) and $r_n \to 0$. Thus $g_n \to g_0$ in $\ell^\infty(T)^K$. Using this and uniform continuity of $\nabla f : \mathbb{D}_f \to \mathbb{R}^K$, lemma 1.9.47 implies $\nabla f(g_n) \to \nabla f(g_0)$ in $\ell^\infty(T)^K$, i.e.

$$\|\nabla f(g_n)-\nabla f(g_0)\|_T = \sup_{t\in T}\|\nabla f(g_n(t))-\nabla f(g_0(t))\| \to 0$$

Using once again that $\{\sup_{t\in T}\|h_n(t)\|\}_{n=1}^{\infty}$ is bounded, this implies

$$\lim_{n\to\infty}\sup_{t\in T}\|\nabla f(g_n(t)) - \nabla f(g_0(t))\| \times \sup_{t\in T}\|h_n(t)\| = 0 \tag{1.105}$$

Now consider the term in (1.104). $\sup_{t\in T}\|\nabla f(g_0(t))\| < \infty$ because $\|\nabla f(\cdot)\|$ is uniformly continuous and $\sup_{t\in T}\|g_0(t)\| < \infty$, just as in the proof of lemma 1.9.47. Furthermore, $\lim_{n\to\infty}\sup_{t\in T}\|h_n(t) - h(t)\| = 0$, so

$$\lim_{n\to\infty}\sup_{t\in T}\|\nabla f(g_0(t))\| \times \sup_{t\in T}\|h_n(t) - h(t)\| = 0 \tag{1.106}$$

Combining (1.103) through (1.106) we obtain

$$\lim_{n\to\infty}\sup_{t\in T}\left|\frac{f(g_0(t) + r_n h_n(t)) - f(g_0(t))}{r_n} - \nabla f(g_0(t))^{\intercal} h(t)\right| = 0$$

which concludes the proof. □

Remark 1.9.7. Lemma 1.9.53 specifies the domain of F as

$$\ell_f^{\infty}(T, \mathbb{D}_f) = \left\{ g \in \ell^{\infty}(T, \mathbb{D}_f) \; ; \; \sup_{t\in T}|f(g(t))| < \infty \right\}.$$

It is often straightforward to clarify the space $\ell_f^{\infty}(T, \mathbb{D}_f)$ in particular cases; for example, $\ell_f^{\infty}(T, \mathbb{D}_f) = \ell^{\infty}(T, \mathbb{D}_f)$ if f satisfies any one of the following: (i) f is bounded, (ii) f is Lipschitz, or (iii) f is bounded on bounded subsets (e.g., $f(x) = x$ is bounded on bounded subsets) See also lemma 1.9.16.

Lemma 1.9.53 requires $\nabla f(\cdot)$ be uniformly continuous, but this often stronger than necessary. When hoping to argue $F : \ell_f^{\infty}(T, \mathbb{D}_f) \to \ell^{\infty}(T)$ defined pointwise with $F(g)(t) = f(g(t))$ is (fully) Hadamard differentiable at $g_0 \in \ell_f^{\infty}(T, \mathbb{D}_f)$, it suffices that f is continuously differentiable on a closed set slightly larger than the (bounded) range of g_0. Compactness of this expanded range and the fact that continuous functions on compact sets are uniformly

continuous allow us to apply the preceding lemma. This logic is formalized in the following corollary.

Corollary 1.9.54 (Hadamard differentiability of maps between bounded function spaces, corollary)**.** *Let $f : \mathbb{D}_f \subseteq \mathbb{R}^K \to \mathbb{R}$ be continuously differentiable.*

Define the subset of $\ell^\infty(T)^K$ taking values in $\mathbb{D}_f$,

$$\ell^\infty(T, \mathbb{D}_f) = \left\{ g : T \to \mathbb{R}^K \; ; \; g(t) \in \mathbb{D}_f, \; \sup_{t \in T} \|g(t)\| < \infty \right\} \subseteq \ell^\infty(T)^K$$

and the subset of $\ell^\infty(T, \mathbb{D}_f)$ such that composition with f defines a bounded function:

$$\ell_f^\infty(T, \mathbb{D}_f) = \left\{ g \in \ell^\infty(T, \mathbb{D}_f) \; ; \; \sup_{t \in T} |f(g(t))| < \infty \right\}$$

Let $g_0 \in \ell_f^\infty(T, \mathbb{D}_f)$, and suppose that for some $\delta > 0$,

$$g_0(T)^\delta \equiv \left\{ x \in \mathbb{R}^K \; ; \; \inf_{t \in T} \|x - g_0(t)\| \leq \delta \right\} \subseteq \mathbb{D}_f.$$

Then $F : \ell_f^\infty(T, \mathbb{D}_f) \to \ell^\infty(T)$ defined pointwise by $F(g)(t) = f(g(t))$ is (fully) Hadamard differentiable at g_0 tangentially to $\ell^\infty(T)^K$, with derivative $F'_{g_0} : \ell^\infty(T)^K \to \ell^\infty(T)$ given pointwise by

$$F'_{g_0}(h)(t) = [\nabla f(g_0(t))]^\intercal \, h(t) = \sum_{k=1}^{K} \frac{\partial f}{\partial x_k}(g_0(t)) h_k(t)$$

Proof. Let $\tilde{f} : g_0(T)^\delta \to \mathbb{R}$ be the restriction of f to $g_0(T)^\delta$. Note that $\tilde{f}$ is continuously differentiable on the compact $g_0(T)^\delta \subseteq \mathbb{R}^K$, hence $\nabla \tilde{f}$ is in fact uniformly continuous by the Heine-Cantor theorem. Apply lemma 1.9.53 to find that

$$\tilde{F} : \ell_f^\infty(T, g_0(T)^\delta) \to \ell^\infty(T), \qquad \tilde{F}(g)(t) = \tilde{f}(g(t)) = f(g(t))$$

is (fully) Hadamard differentiable at g_0, with derivative $\tilde{F}'_{g_0} : \ell^\infty(T)^K \to \ell^\infty(T)$ given pointwise by $\tilde{F}'_{g_0}(h)(t) = [\nabla f(g_0(t))]^\intercal h(t)$. By definition, this means that for any sequences $\{\tilde{h}_n\}_{n=1}^\infty \subseteq \ell^\infty(T)^K$ and $\{\tilde{r}_n\}_{n=1}^\infty \subseteq \mathbb{R}$ such that $\tilde{h}_n \to \tilde{h} \in \ell^\infty(T)^K$, $\tilde{r}_n \to 0$, and $g_0 + \tilde{r}_n\tilde{h}_n \in \ell^\infty(T, g_0(T)^\delta)$ for all n,

$$\lim_{n\to\infty} \left\| \frac{\tilde{F}(g_0 + \tilde{r}_n\tilde{h}_n) - \tilde{F}(g_0)}{\tilde{r}_n} - F'_{g_0}(\tilde{h}) \right\|_T = 0 \tag{1.107}$$

Let $\{h_n\}_{n=1}^\infty \subseteq \ell^\infty(T)^K$, $\{r_n\}_{n=1}^\infty \subseteq \mathbb{R}$ be such that $h_n \to h \in \ell^\infty(T)^K$, $r_n \to 0$, and $g_0 + r_n h_n \in \ell^\infty(T, \mathbb{D}_f)$ for all n. It suffices to show that

$$\begin{aligned} &\left\| \frac{F(g_0 + r_n h_n) - F(g_0)}{r_n} - F'_{g_0}(h) \right\|_T \\ &\qquad = \sup_{t\in T} \left| \frac{f(g_0(t) + r_n h_n(t)) - f(g_0(t))}{r_n} - [\nabla f(g_0(t))]^\intercal h(t) \right| \end{aligned}$$

has limit zero.

Notice that $g_0 + r_n h_n \to g_0$ in $\ell^\infty(T)^K$, so for some N we have that for all $n \geq N$, $\|g_0 + r_n h_n - g_0\|_T = r_n \sup_{t\in T}\|h_n\| < \delta$. It follows that for $k \in \mathbb{N}$, $g_0 + r_{k+N}h_{k+N} \in \ell^\infty(T, g_0(T)^\delta)$ and hence $\tilde{r}_k = r_{k+N}$ and $\tilde{h}_k = h_{k+N}$ are sequences for which (1.107) applies. Therefore,

$$\begin{aligned} \lim_{n\to\infty} \left\| \frac{F(g_0 + r_n h_n) - F(g_0)}{r_n} - F'_{g_0}(h) \right\|_T &= \lim_{k\to\infty} \left\| \frac{F(g_0 + r_{k+N} h_{k+N}) - F(g_0)}{r_{k+N}} - F'_{g_0}(h) \right\|_T \\ &= \lim_{k\to\infty} \left\| \frac{\tilde{F}(g_0 + \tilde{r}_k\tilde{h}_k) - \tilde{F}(g_0)}{\tilde{r}_k} - F'_{g_0}(h) \right\|_T \\ &= 0 \end{aligned}$$

Where the second equality follows from $\tilde{F}(g_0 + \tilde{r}_k\tilde{h}_k) = F(g_0 + r_{k+N}h_{k+N})$ and $\tilde{F}(g_0) = F(g_0)$. □

The following lemma is lemma S.4.9 from Fang & Santos (2019), but the authors state it for a metric space. The same proof works to show that statement holds in semimetric spaces as well.

Lemma 1.9.55 (Hadamard directional differentiability of supremum)**.** *(Fang & Santos (2019) lemma S.4.9)*

Let $(\boldsymbol{A}, d)$ be a compact semimetric space, A a compact subset of $\boldsymbol{A}$, and

$$\psi : \ell^\infty(\boldsymbol{A}) \to \mathbb{R}, \qquad \psi(p) = \sup_{a \in A} p(a)$$

Then ψ is Hadamard directionally differentiable at any $p_0 \in \mathcal{C}(\boldsymbol{A}, d)$ tangentially to $\mathcal{C}(\boldsymbol{A}, d)$. $\Psi_A(p_0) = \arg\max_{a \in A} p_0(a)$ is nonempty, and the directional derivative is given by

$$\psi'_{p_0} : \mathcal{C}(\boldsymbol{A}, d) \to \mathbb{R}, \qquad \psi'_{p_0}(p) = \sup_{a \in \Psi_A(p_0)} p(a)$$

1.9.7 Appendix: extensions

This appendix briefly describes a few simple extensions.

1.9.7.1 Conditioning on $X \in A$

In many applications parameters conditional on a covariate taking a particular value are of interest. For example, the share of compliers of a particular demographic benefiting from treatment is $P(Y_1 > Y_0 \mid D_1 > D_0, \text{demographic})$. Such parameters can be written in the form

$$\gamma_A = g(\theta_A, \eta_A)$$

where for a known set $A \subseteq \mathcal{X}$,

$$\theta_A \equiv E[c(Y_1, Y_0) \mid D_1 > D_0, X \in A], \qquad \eta_A \equiv E[\eta_1(Y_1), \eta_0(Y_0) \mid D_1 > D_0, X \in A]$$

The identified set for γ_A is straightforward to characterize and estimate. First note that

$$\theta_A = E[\theta_X \mid D_1 > D_0, X \in A] = \frac{1}{s_A} \sum_{x \in A} s_x \theta_x$$

where $s_A = \sum_{x \in A} s_x$. The proof of theorem 1.4.1 shows that the sharp identified set for $(\theta_{x_1}, \ldots, \theta_{x_M})$ is in fact $[\theta^L_{x_1}, \theta^H_{x_1}] \times \ldots \times [\theta^L_{x_M}, \theta^H_{x_M}]$. It follows that the sharp identified set for θ_A is $[\theta^L_A, \theta^H_A]$, where

$$\theta^L_A = \frac{1}{s_A} \sum_{x \in A} s_x \theta^L_x, \qquad \theta^H_A = \frac{1}{s_A} \sum_{x \in A} s_x \theta^H_x$$

and the sharp identified set for γ_A is $[\gamma^L_A, \gamma^H_A]$ where

$$\gamma^L_A = \min_{t \in [\theta^L_A, \theta^H_A]} g(t, \eta_A), \qquad \gamma^H_A = \max_{t \in [\theta^L_A, \theta^H_A]} g(t, \eta_A),$$

Let $\hat{s}_x$, $\hat{\theta}^L_x$, and $\hat{\theta}^H_x$ be as defined in section 1.5. Let $\hat{s}_A = \sum_{x \in A} \hat{s}_x$ and

$$\hat{\theta}^L_A = \frac{1}{\hat{s}_A} \sum_{x \in A} \hat{s}_x \hat{\theta}^L_x, \qquad \hat{\theta}^H(A) = \frac{1}{\hat{s}_A} \sum_{x \in A} \hat{s}_x \hat{\theta}^H_x$$

$$\hat{\gamma}^L_A = \min_{t \in [\hat{\theta}^L_A, \hat{\theta}^H_A]} g(t, \hat{\eta}_A), \qquad \hat{\gamma}^H_A = \max_{t \in [\hat{\theta}^L_A, \hat{\theta}^H_A]} g(t, \hat{\eta}_A),$$

Under assumptions 1, 2, and 3, $\sqrt{n}((\hat{\gamma}^L_A, \hat{\gamma}^H_A) - (\gamma^L_A, \gamma^H_A)$ will converge weakly. With assumption 4 the straightforward bootstrap will consistently estimate its asymptotic distribution.

1.9.7.2 Multiple treatment arms with exogenous treatment

The results above are easily extended to a setting with multiple treatment arms and exogenous treatment. Let $d \in \{0, 1, \ldots, J\}$, index the mutually exclusive treatment arms, with $d = 0$ indicating control. Let Y_d be the potential outcome with treatment d, and D_d equal one if the unit has treatment d and zero otherwise. The observed outcome is

$$Y = \sum_{d=0}^{J} D_d Y_d$$

Let $D = (D_0, D_1, \ldots, D_J)$ and assume $(Y_0, Y_1, \ldots, Y_J) \perp D \mid X$. The marginal distributions of $Y_d \mid X = x$, denoted $P_{d|x}$, are identified with the relation

$$E_{P_{d|x}}[f(Y_d)] = E[f(Y_d) \mid X = x] = \frac{E[f(Y)D_d \mid X = x]}{P(D_d = 1 \mid X = x)}$$

Let $\gamma_d = g(\theta_d, \eta_d)$ where $\theta_d = E[c(Y_d, Y_0)]$. Consider estimating the sharp identified set for $(\gamma_1, \ldots \gamma_J)$. For example, an RCT with two treatment arms may have similar average treatment effects. The treatment arms may be further distinguished by comparing $P(Y_1 - Y_0 > 0)$ with $P(Y_2 - Y_0 > 0)$, or $\text{Cov}(Y_1 - Y_0, Y_0)$ with $\text{Cov}(Y_2 - Y_0, Y_0)$.

Let $\theta_{d,x} = E[c(Y_1, Y_0) \mid X = x]$. The sharp identified set for $(\theta_{1,x}, \ldots, \theta_{J,x})$ is given by

$$[\theta^L_{1,x}, \theta^H_{1,x}] \times \ldots \times [\theta^L_{J,x}, \theta^H_{J,x}]$$

where $\theta^L_{d,x} = \theta^L(P_{d|x}, P_{0|x})$ and $\theta^H_{d,x} = \theta^H(P_{d|x}, P_{0|x})$ as in section 1.4.[16] The sharp identified set for θ_d is $[\theta^L_d, \theta^H_d]$ where $\theta^L_d = \sum_x s_x \theta^L_{d,x}$ and $\theta^H_d = \sum_x s_x \theta^H_{d,x}$, and the sharp identified set for $(\gamma_1, \ldots \gamma_J)$ is

$$[\gamma^L_1, \gamma^H_1] \times \ldots \times [\gamma^L_J, \gamma^H_J]$$

[16] This follows from existing results and the *gluing lemma*, found in Villani (2009) (pp. 11-12).

Sample analogues $(\hat{\gamma}_1^L, \hat{\gamma}_1^H, \ldots, \hat{\gamma}_J^L, \hat{\gamma}_J^H)$ can be formed just as in section 1.5. Under natural adjustments to assumptions 2, 3, and 4, the same arguments work to show

$$\sqrt{n}((\hat{\gamma}_1^L, \hat{\gamma}_1^H, \ldots, \hat{\gamma}_J^L, \hat{\gamma}_J^H) - (\gamma_1^L, \gamma_1^H, \ldots, \gamma_J^L, \gamma_J^H))$$

is asymptotically Gaussian and the bootstrap consistently estimates its asymptotic distribution.

CHAPTER 2

Robustness to Missing Data: Breakdown Point Analysis

2.1 Introduction

Virtually every economic dataset is plagued by missing and incomplete records. Survey nonresponse is the most visible cause, and appears to be worsening over time. Bollinger et al. (2019) report that the Current Population Survey's Annual Social and Economics Supplement item and whole nonresponse has been increasing, reaching 43% in 2015. By linking these data with the Social Security Administration Detailed Earnings Record, the authors show that the distribution of nonreponders differs from that of responders even after conditioning on a large set of covariates.

Samples with missing or incomplete observations fail to identify the population distribution (Manski, 2005). To make progress, researchers commonly apply standard procedures to the complete observations. This practice is typically justified by assuming the data are "missing completely at random" (MCAR); that is, incomplete observations follow the same distribution as that of the complete observations. In many settings such an assumption is implausible. Without it, the conclusions drawn are uncomfortably qualified as being about the distribution of the complete observations, rather than the actual distribution of interest.

This paper proposes a method to investigate the robustness of a conclusion when asserted about the whole population. Results are more robust when overturning them would require more selection. To make this intuition precise, selection is measured with the squared

Hellinger divergence between the distribution of complete observations and that of the incomplete observations. Although many different statistical divergences could be used to measure selection, squared Hellinger is interpretable as a measure of how well the variables under study would predict an observation being complete. This gives the values of the selection measure context, allowing researchers to guage how much selection can be expected in a given setting.

The *breakdown point* is the minimum amount of selection needed to overturn a conclusion. Readers who doubt the setting exhibits that much selection will find the conclusion compelling. In models identified with the generalized method of moments (GMM), the breakdown point is the constrained minimum of the value function of a convex optimization problem. Estimators of the breakdown point are constructed from the dual of this convex inner problem, and shown to be $\sqrt{n}$-consistent and asymptotically normal. Lower confidence intervals are simple to construct. Reporting the point estimates and lower confidence intervals of the breakdown point is a simple, concise way to communicate a result's robustness.

This approach has a number of advantages over existing methods for incomplete datasets. Sample selection models consider regressions with samples where the dependent variable is sometimes missing, and obtain point identification by modeling the selection process (Heckman, 1979; Das et al., 2003). These models require the data include a variable changing the probability of observation but not the dependent variable. This "exclusion restriction" is difficult to satisfy in many applications. The breakdown point approach proposed here can be used on most common GMM models (including but not restricted to regressions with missing outcomes), and requires no additional data. The breakdown point can be estimated even if the incomplete observations are in fact completely missing, a distinct possibility when using survey data.

The econometric literature on missing data has also explored bounding the parameter of interest based on the support (Manski, 2005; Horowitz & Manski, 2006). If all parameter values within these "worst-case" bounds satisfy the researcher's conclusion, then the conclu-

sion is undoubtedly robust. Unfortunately, the bounds may be uninformative in practice. Proponents of this approach are well aware these bounds are conservative, and propose this exercise as a place to begin an investigation rather than end one. Additional identifying assumptions should then be considered, in order to make plain to readers what needs to be assumed to reach a given conclusion (Manski, 2013). The breakdown approach proposed here is a simple version of this exercise, as the assumption that selection is less than the breakdown point leads one to conclude the hypothesis under investigation.

A growing literature advocates for breakdown analysis as a general, tractable method to assess the sensitivity of a result to relaxations of identifying assumptions. The term "identification breakdown point" can be found as early as Horowitz & Manski (1995) in the context of corrupted data. Masten & Poirier (2020) advocates for the approach generally, and illustrates it with the potential outcomes framework. Diegert et al. (2022) define and study breakdown points in linear regressions suffering from omitted variable bias.

This paper is not the first to notice the appeal of breakdown point analysis in the context of missing data. Kline & Santos (2013) consider a setting with a missing scalar, propose measuring selection with the maximal Kolmogorov-Smirnov (KS) distance between the conditional distributions of complete and incomplete observations across all values of covariates, and advocate for "reporting the minimal level of selection necessary to undermine a hypothesis," (p. 233). The methodology proposed here has some notable advantages. First, measuring selection with the maximal KS distance limits researchers to the case where only a scalar is missing, while measuring selection as proposed here allows any number of variables to be missing. Second, in a given setting it is easier to gauge whether the variables under study are likely to be good predictors of missingness than what share of the missing data is missing at random. This makes squared Hellinger a more natural measure of selection than KS distance. Which approach is more tractable will depend on the parameter of interest. Kline & Santos (2013) derives sharp, closed form bounds to the conditional quantiles of the missing variable, and frame the conclusion to be investigated in terms of those quan-

tiles. This paper assumes the parameter of interest is identified with GMM and uses the model directly, giving up closed form solutions. In thoery this could lead to computational difficulties, but the simulations in section 2.5 present no issue.

The remainder of this paper is structured as follows: section 2.2 formalizes the setting, the proposed measure of selection, and the breakdown point. The dual problem is presented and discussed in section 2.3. Section 2.4 defines the estimator and states the main results on estimation and inference, which are proven in the appendix. Section 2.5 presents a simulation study investigating the finite sample performance of these estimators. Section 2.6 concludes.

2.2 Measuring selection and breakdown analysis

Suppose the available data is the i.i.d. sample $\{(D_i, D_iY_i, X_i)\}_{i=1}^n$, where $Z_i \equiv (Y_i, X_i) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ contains the variables of interest and $D_i \in \{0, 1\}$ indicates whether Y_i is observed. Note that Y_i may be a vector, and X_i may be empty. Let $p_D \equiv P(D = 1)$ denote the probability of observing Y, P_1 the distribution of Z conditional on $D = 1$, and P_0 the distribution of Z conditional on $D = 0$. P_1 and P_0 are called the *complete case* and *incomplete case* distributions respectively. The distribution of interest is the unconditional distribution of Z, given by $p_D P_1 + (1 - p_D)P_0$. When X is nonempty, the marginal distribution of X conditional on $D = 0$ is denoted P_{0X}. For simplicitly, X is assumed to have the same finite support when $D = 0$ as when $D = 1$, which greatly simplifies asymptotic analysis. Remark 2.2.3 below discusses this assumption further.

To fix ideas, consider data collection via survey. Y is a vector of data the survey hopes to collect, which is observed only if the recipient responds ($D = 1$). The survey's response rate, $p_D = P(D = 1)$, is essentially always less than one in practice. It is common for administrative data to provide basic information about a survey recipient (such as age, occupation, etc.), which is collected in X.

Analyses based on the complete observations may not convince researchers who worry

that P_0 differs from P_1. Such concerns are common, as few settings plausibly satisfy the missing completely at random assumption. However, it is often similarly implausible that P_0 differs greatly from P_1. Researchers who convincingly argue that P_0 is not too different from P_1 can still convince their audience of conclusions drawn from an analysis of P_1.[1]

A quantitative measure of the difference between P_1 and P_0 is needed to make this argument formal and convincing. The statistics literature provides a natural solution in the form of *divergences*: functions mapping two probability distributions to the nonnegative real line that take value zero if and only if the two distributions are the same. There are many such functions. To be useful as a measure of selection, a divergence should have a tractable interpretation, so that researchers can gauge whether a given amount of selection is reasonable for their setting.

2.2.1 An interpretable measure of selection

Missing data cause greater concern when researchers expect the variables of interest (Z) to be a good predictor of incompleteness (D). Consider again the example of data collection via survey. Researchers are rightfully more concerned about survey nonresponse when asking about the respondent's arrest record than when asking for opinions on recent television programming. People with criminal records may be less willing to answer questions about that record.[2] This suggests that the distribution of responders may look quite different from the distribution of nonresponders, and that criminal records would be a good predictor of nonresponse.

To illustrate this more formally, let f_1 and f_0 be densities of P_1 and P_0 with respect to

[1] In some cases, such as correctly specified regression models, it suffices that the conditional distributions $f_{Y|X=x,D=0}(y \mid x)$ are the same as the identified $f_{Y|X=x,D=1}(y \mid x)$. This weaker "missing at random" (MAR) assumption is also rarely plausible in practice, and analyses based on this assumption often rely heavily on the model being correctly specified.

[2] For example, Brame et al. (2012) estimate the cumulative prevalence of arrest from ages 8 to 23 from a survey directly asking about prior arrests. The authors report upper and lower bounds derived by assuming the entire set of nonresponders had or had not been arrested, essentially the worst-case bounds advocated for by Manski (2005).

$p_D P_1 + (1 - p_D)P_0$ respectively:

$$f_1(z) = \frac{P(D = 1 \mid Z = z)}{p_D}, \qquad f_0(z) = \frac{(1 - P(D = 1 \mid Z = z))}{1 - p_D}$$

An optimist may assume D is independent of Z, implying that $P(D = 1 \mid Z = z) = P(D = 1) = p_D$ and $f_0 = f_1 = 1$. In contrast, a pessimist may assume D is close to a deterministic function of Z, allowing Z to predict D well. This would imply $P(D = 1 \mid Z = z)$ is close to 1 or 0 for many values of z, and that f_1 differs greatly from f_0.

As in the survey example, the setting often makes it clear whether Z would be a good predictor of D. This hueristic is useful to identify and discuss selection concerns. The following lemma shows that measuring selection as the squared Hellinger distance between P_0 and P_1 captures this intuition, with larger values corresponding to Z having greater capability of predicting D.[3]

Lemma 2.2.1. *Let $(Z, D) \in \mathbb{R}^{d_z} \times \{0, 1\}$ be random variables with $p_D = P(D = 1) \in (0, 1)$. Let $Z \mid D = 1 \sim P_1$ and $Z \mid D = 0 \sim P_0$. Then*

$$H^2(P_0, P_1) = 1 - \frac{E\left[\sqrt{Var(D \mid Z)}\right]}{\sqrt{Var(D)}} \tag{2.1}$$

where the expectation is taken with respect to $p_D P_1 + (1 - p_D)P_0$, the marginal distribution of Z.

All results are proven in the appendix.

Equation (2.1) states that the squared Hellinger distance between P_0 and P_1 is the expected percent of the standard deviation of D reduced by conditioning on Z. In the extreme

[3]The Hellinger distance between probability measures Q and P is

$$H(Q, P) \equiv \left(\frac{1}{2} \int \left(\sqrt{\frac{dQ}{d\lambda}(z)} - \sqrt{\frac{dP}{d\lambda}(z)} \right)^2 d\lambda(z) \right)^{1/2}$$

where λ is any measure dominating both P and Q.

case where $\text{Var}(D \mid Z) = \text{Var}(D)$, equation (2.1) implies $H^2(P_0, P_1) = 0$ and the conditional distributions are the same. As the ability of Z to predict D grows, the variance of D conditional on Z decreases and $H^2(P_0, P_1)$ grows toward one.

Remark 2.2.1. It's straightforward to see that $\text{Var}(D \mid X) \geq \text{Var}(D \mid X, Y)$ implies

$$H^2(P_0, P_1) = 1 - \frac{E\left[\sqrt{\text{Var}(D \mid Y, X)}\right]}{\sqrt{\text{Var}(D)}} \geq 1 - \frac{E[\sqrt{\text{Var}(D \mid X)}]}{\sqrt{\text{Var}(D)}} = H^2(P_{0X}, P_{1X})$$

where P_{0X}, P_{1X} are the marginal distributions of X conditonal on $D = 0$ and $D = 1$ respectively. This lower bound on the selection is identified from the sample, and motivates the common practice of comparing the distribution of X conditional on $D = 0$ with that of X conditional on $D = 1$; the distributions P_0 and P_1 can only be "further" apart.

2.2.2 Divergences

Squared Hellinger provides an intuitive measure of selection, but there are many other options. A function $d(\cdot\|\cdot)$ mapping two probability distributions P and Q to $\mathbb{R}$ is called a *divergence* if 1. $d(Q\|P) \geq 0$, and 2. $d(Q\|P) = 0$ if and only if $P = Q$. Divergences need not be symmetric nor satisfy the triangle inequality. The set of *f-divergences* are particularly well behaved. Given a convex function $f : \mathbb{R} \to [0, \infty]$ satisfying $f(t) = \infty$ for $t < 0$ and taking a unique minimum of $f(1) = 0$, the corresponding f-divergence is given by

$$d_f(Q\|P) \equiv \begin{cases} \int f\left(\frac{dQ}{dP}\right) dP & \text{if } Q \ll P \\ \infty & \text{otherwise} \end{cases} \tag{2.2}$$

Many popular divergences are equal to f-divergences when P dominates Q.

Table 2.1: Common f-divergences

Name	Common formula	$f(t)$
Squared Hellinger	$H^2(Q,P) = \frac{1}{2}\int \left(\sqrt{\frac{dQ}{dP}(z)} - 1\right)^2 dP(z)$	$\frac{1}{2}(\sqrt{t}-1)^2$
Kullback-Leibler (KL)	$KL(Q\|P) = \int \log\left(\frac{dQ}{dP}(z)\right) dQ(z)$	$t\log(t) - t + 1$
"Reverse" KL	$KL(P\|Q) = \int \log\left(\frac{dP}{dQ}(z)\right) dP(z)$	$-\log(t) + t - 1$
Cressie-Read	–	$\frac{t^\gamma - \gamma t + \gamma - 1}{\gamma(\gamma-1)}$, $\gamma < 1$

Although squared Hellinger has intuitive appeal outlined in Section 2.2.1, the breakdown point analysis proposed in this paper remains tractable for any f-divergence listed in Table 2.1.[4] Precise assumptions regarding the f-divergence are collected in Assumption 5 below.

Remark 2.2.2. Measuring selection with an f-divergence facilitates estimation and inference, as the space of distributions Q with $d_f(Q\|P_1) < \infty$ corresponds to the set of densities with respect to P_1. In substance, this assumes $P_0 \ll P_1$ and rules out selection mechanisms that "truncate" data.

2.2.3 Breakdown analysis in GMM models

Suppose a preliminary analysis supports an alternative hypothesis H_1 over a null hypothesis H_0.[5] The breakdown point is the minimum amount of selection needed to overturn such a conclusion. When selection is measured in terms of the squared Hellinger distance, the breakdown point translates the claim that H_0 is true into a claim about the ability of Z to

[4]It is worth noting that the Cressie-Read divergence nests the other three as special cases. Squared Hellinger corresponds to $\frac{1}{2}f_{1/2}$. l'Hôpital's rule shows that Kullback-Leibler corresponds to $\lim_{\gamma\to 1} f_\gamma$ and Reverse Kullback-Leibler to $\lim_{\gamma\to 0} f_\gamma$. See Broniatowski & Keziou (2012) for additional discussion.

[5]For example, such an analysis may be based on the complete observations assuming MCAR, or using imputation and assuming Y is MAR conditional on X.

predict D. Specifically, if H_0 were true then $1 - \frac{E[\sqrt{\text{Var}(D|Z)}]}{\sqrt{\text{Var}(D)}}$ would be weakly larger than the breakdown point. If this is implausible, then H_0 is similarly implausible.

This section formalizes this idea for generalized method of moment (GMM) models. Suppose the parameter of interest $\beta \in \mathbf{B} \subseteq \mathbb{R}^{d_b}$ is characterized as the unique solution to a finite set of moment conditions,

$$E[g(Z, \beta)] = 0 \in \mathbb{R}^{d_g}$$

where the expectation is taken with respect to the unconditional distribution, $p_D P_1 + (1 - p_D)P_0$. The conclusion to be investigated is that β falls outside a particular set $\mathbf{B}_0 \subset \mathbf{B}$, motivating the null and alternative hypotheses

$$H_0 \ : \ \beta \in \mathbf{B}_0, \qquad\qquad H_1 \ : \ \beta \in \mathbf{B} \setminus \mathbf{B}_0$$

Recall that the observed data is $\{(D_i, D_i Y_i, X_i)\}_{i=1}^n$, where $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$. The sample identifies P_1, p_D, and P_{0X}. A hypothetical distribution of the incomplete observations Q *rationalizes* the parameter b if it has the identified marginal distribution of X, $Q_X = P_{0X}$, and the implied unconditional distribution $p_D P_1 + (1 - p_D)Q$ solves the moment conditions for b. The set of such distributions implying finite selection is

$$\mathbf{P}^b \equiv \{Q \ ; \ Q \ll P_1, \ Q_X = P_{0X}, \ p_D E_{P_1}[g(Z, b)] + (1 - p_D) E_Q[g(Z, b)] = 0\} . \tag{2.3}$$

The *breakdown point* δ^{BP} is the minimum selection needed to rationalize the null hypothesis:

$$\delta^{BP} \equiv \inf_{b \in \mathbf{B}_0} \inf_{Q \in \mathbf{P}^b} d_f(Q \| P_1) \tag{2.4}$$

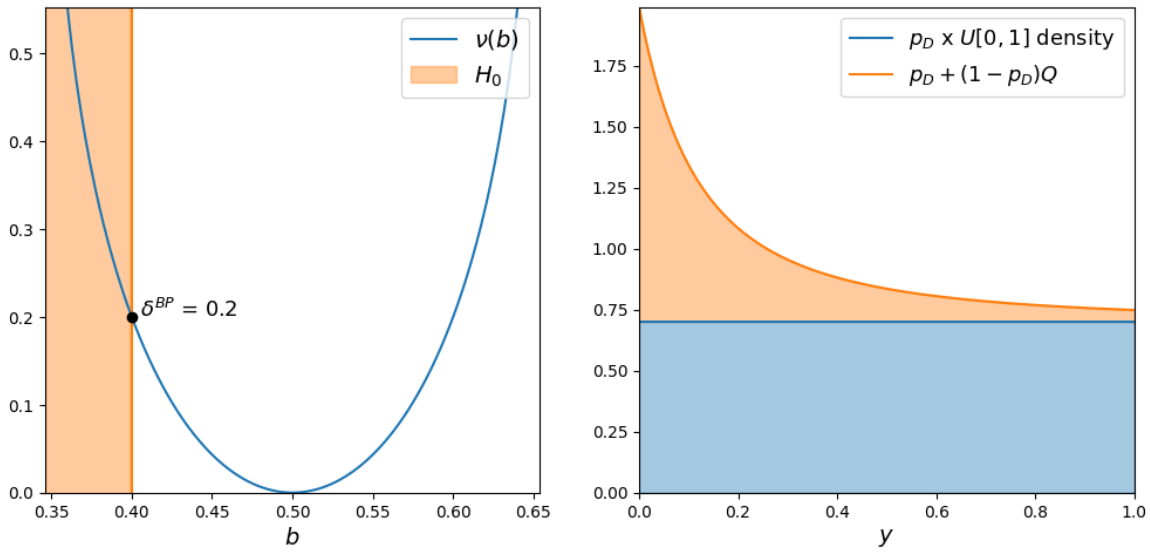
where the infimum over the empty set is understood as ∞. A simple example illustrates the idea.

Example 2.2.1. *Let $Y \in \mathbb{R}$ and $\beta = E[Y] = p_D E_{P_1}[Y] + (1 - p_D)E_{P_0}[Y]$. Let $p_D = 0.7$ and P_1 be $\mathcal{U}[0, 1]$. The claim to support is $H_1 : \beta > 0.4$, and selection is measured with squared Hellinger. $\boldsymbol{P}^b$ is the set of continuous distributions on $[0, 1]$ with expectation $\frac{b - p_D/2}{1 - p_D}$, so that $Q \in \boldsymbol{P}^b$ implies*

$$p_D E_{P_1}[Y] + (1 - p_D)E_Q[Y] = \frac{p_D}{2} + (1 - p_D)\frac{b - p_D/2}{1 - p_D} = b$$

The inner minimization in (2.4) chooses the distribution that minimizes selection while rationalizing b. The outer minimization chooses the parameter that minimizes selection while rationalizing $H_0 : \beta \leq 0.4$. Unsurprisingly, the outer minimization is solved by $b = 0.4$. The breakdown point δ^{BD} is slightly above 0.2. A researcher convinced $H^2(P_0, P_1) \leq 0.2$ should conclude $\beta > 0.4$.

Figure 2.1: $\nu(b)$ and $p_D P_1 + (1 - pD)Q^*$, where $Q^* \in \mathbf{P}^{0.4}$ minimizes selection.



Breakdown analysis can also be framed as an exercise in partial identification, as in Kline & Santos (2013), Masten & Poirier (2020), and Diegert et al. (2022). In this framing, the researcher considers assumptions of the form $d_f(P_0, P_1) \leq \delta$ for some $\delta > 0$, which

continuously relax the assumption $P_0 = P_1$. The identified set for β grows with δ. As long as the identified set is a subset of $\mathbf{B} \setminus \mathbf{B}_0$, it is clear the researcher's conclusion holds. The breakdown point δ^{BP} can then be defined as either the largest δ for which the identified set is contained in $\mathbf{B} \setminus \mathbf{B}_0$, or the smallest δ for which the identified set has nontrivial intersection with $\mathbf{B}_0$ (the latter of which corresponds to the definition given in (2.4)). For further discussion of this equivalent framing of the breakdown point, see appendix 2.7.2.2.

The remainder of this paper constructs a $\sqrt{n}$-consistent and asymptotically normal estimator of δ^{BP}, and constructs a lower confidence interval for δ^{BP}. Researchers working with partially complete datasets should discuss the plausible amount of selection in their setting, and report the point estimate and the lower confidence interval for δ^{BP} for each asserted conclusion. This will make plain to readers which conclusions are more sensitive to missing data concerns, and whether crucial results are sufficiently robust.

2.2.4 Preview of results

The estimation proceeds by separating the optimizations in (2.4). Define the *primal problem*

$$\nu(b) = \inf_{Q \in \mathbf{P}^b} d_f(Q \| P_1) \tag{2.5}$$

and notice that $\delta^{BP} = \inf_{b \in \mathbf{B}_0} \nu(b)$. The first step is to estimate the value function ν over a set $B \subseteq \mathbf{B}$ large enough that $\inf_{b \in \mathbf{B}_0} \nu(b) = \inf_{b \in B \cap \mathbf{B}_0} \nu(b)$, while the second step estimates δ^{BP} through a simple plug-in estimator.

The primal problem is an infinite dimensional convex optimization problem over the space of probability distributions, but one that is very well studied in convex analysis. In particular, when $\mathbf{P}^b$ defined in (2.3) is characterized by a finite number of moment conditions, (2.5) has a well behaved, finite-dimensional dual problem with the same value function (Borwein & Lewis, 1991, 1993; Csiszár et al., 1999; Broniatowski & Keziou, 2006). Section 2.3 discusses this dual problem and the assumptions needed to make use of it. Under regularity conditions

discussed in section 2.4, sample analogue estimators based on this dual problem are uniformly consistent and asymptotically Gaussian on compact sets. Differentiability of the infimum then implies convergence in distribution of the plug-in estimator.

To conclude this section, Assumption 5 collects conditions on the setting, the GMM model, and the f-divergence used to measure selection.

Assumption 5 (Setting)**.** $\{(D_i, D_iY_i, X_i)\}_{i=1}^n$ *is an i.i.d. sample from a distribution satisfying*

(i) $p_D = P(D = 1) \in (0, 1)$,

(ii) $X \mid D = 1$ *and* $X \mid D = 0$ *have the same finite support* $\{x_1, \ldots, x_K\}$,

(iii) $E\left[\sup_{b \in \boldsymbol{B}} \|g(Z, b)\| \mid D = 1\right] < \infty$, *where* $Z = (Y, X)$, *and*

(iv) $f : \mathbb{R} \to [0, \infty]$ *is closed, proper, strictly convex, essentially smooth, takes its unique minimum of* $f(t) = 0$ *at* $t = 1$, *and satisfies* $f(t) = \infty$ *for all* $t < 0$. *The interior of* $dom(f) \equiv \{t \in \mathbb{R} \,;\, f(t) < \infty\}$, *denoted* (ℓ, u), *satisfies* $\ell < 1 < u$, *and* f *is twice continuously differentiable on* (ℓ, u).

The finite support condition in (ii) ensures that $\mathbf{P}^b$ defined in (2.3) is characterized by a finite number of moments (see remark 2.2.3 below for additional discussion). Condition (iv) ensures the f-divergence used to measure selection is well behaved, and is satisfied by every divergence in Table 2.1.[6] In particular, strict convexity of f ensures the primal problem (2.5) has a unique solution (P_1-almost surely). f is required to be essentially smooth to ensure the dual problem has a unique solution. The requirements that $f(x)$ take a unique minimum of 0 at $x = 1$ and $f(x) = \infty$ for $x < 0$ ensures that $d_f(Q\|P)$ is a well defined f-divergence.

Remark 2.2.3. If X is not finitely valued, it is easy to see that requiring Q_X match a finite number of moments of P_{0X} will estimate a value no larger than δ^{BP}. If this value is large

[6]See appendix 2.7.3 for definitions of the convex analysis terms used in Assumption 5 (iv).

enough to assuage missing data concerns, the breakdown point can only be larger. When the distribution of X is characterized by a countable set of moments, it may be possible to increase the number of moments with the sample size to estimate δ^{BP} directly. This is left for future research.

2.3 Duality

As defined in display (2.5), $\nu(b)$ is the value function of an infinite dimensional convex optimization problem. Fortunately, when selection is measured with an f-divergence, (2.5) becomes a well-studied problem known by various names: maximal entropy (Csiszár et al. (1999)), partially finite programming (Borwein & Lewis (1991)), or simply f-divergence projection (Broniatowski & Keziou (2006)). The convex analysis results in these papers connect the primal problem in (2.5) to a finite dimensional dual problem that is much easier to study and estimate. Under mild conditions, the value function of this dual problem coincides with the value function of the primal.

To state the dual problem, first note that the primal can be viewed as a problem over the set of densities with respect to P_1:

$$\begin{aligned}\nu(b) = &\inf_q E\left[f(q(Y,X)) \mid D=1\right] \\ &\text{s.t. } E[h(Y,X,b)q(Y,X) \mid D=1] = c(b)\end{aligned}$$

where

$$h(z,b) \equiv h(y,x,b) \equiv \begin{pmatrix} g(y,x,b) \\ \mathbb{1}\{x = x_1\} \\ \vdots \\ \mathbb{1}\{x = x_K\} \end{pmatrix}, \qquad c(b) \equiv \begin{pmatrix} \frac{-p_D}{1-p_D} E[g(Y,X,b) \mid D=1] \\ P(X = x_1 \mid D=0) \\ \vdots \\ P(X = x_K \mid D=0) \end{pmatrix}, \tag{2.6}$$

The dual problem corresponding to the problem in display (2.5) is given by

$$V(b) \equiv \sup_{\lambda \in \mathbb{R}^{d_g+K}} \lambda^\intercal c(b) - E\left[f^*\left(\lambda^\intercal h(Y, X, b)\right) \mid D = 1\right] \tag{2.7}$$

where f^* is the convex conjugate of f, given by $f^*(r) \equiv \sup_{t \in \mathbb{R}}\{rt - f(t)\}$. For convenience, table 2.2 summarizes the convex conjugate for several common divergences.

Table 2.2: Common f-divergence conjugates and effective domains

Name	$f(t)$	ℓ, u	$f^*(r)$	ℓ^*, u^*
Squared Hellinger	$\frac{1}{2}(\sqrt{t}-1)^2$	$\ell = 0$, $u = \infty$	$\frac{1}{2}\left(\frac{1}{1-2r} - 1\right)$	$\ell^* = -\infty$, $u^* = 1/2$
Kullback-Leibler	$t\log(t) - t + 1$	$\ell = 0$, $u = \infty$	$\exp(r) - 1$	$\ell^* = -\infty$, $u^* = \infty$
"Reverse" KL	$-\log(t) + t - 1$	$\ell = 0$, $u = \infty$	$-\log(1-r)$	$\ell^* = -\infty$, $u^* = 1$

Remark 2.3.1. To ensure q corresponds to a probability density, the constraints must enforce $\int q(z)dP(z) = 1$. This is implied by the constraints ensuring $Q_X = P_{0X}$ when X is present. If there are no always-observed variables, set $h(z, b) = \begin{pmatrix} g(z,b)^\intercal & 1 \end{pmatrix}^\intercal \in \mathbb{R}^{d_g+1}$ and $c(b) = \begin{pmatrix} \frac{-p_D}{(1-p_D)} E[g(Y, X, b) \mid D = 1]^\intercal & 1 \end{pmatrix}^\intercal$.

2.3.1 Weak and strong duality

Assumption 5 suffices to show $V(b) \leq \nu(b)$. This fact is known as *weak duality*, and implies that

$$\inf_{b \in B \cap \mathbf{B}_0} V(b) \leq \inf_{b \in B \cap \mathbf{B}_0} \nu(b) = \delta^{BP} \tag{2.8}$$

for any $B \subseteq \mathbf{B}$. This inequality shows that using the dual problem for estimation of the breakdown point is at worst conservative: if $\inf_{b \in B \cap \mathbf{B}_0} V(b)$ is large enough to assuage selection concerns, researchers are assured that the breakdown point can only be larger.

Assuming only slightly more ensures *strong duality* holds, that is, $V(b) = \nu(b)$. Recall from Assumption 5 (iv) that the interior of $\text{dom}(f) = \{t \in \mathbb{R} \,;\, f(t) < \infty\}$ is denoted (ℓ, u).

Assumption 6 (Strong duality)**.** *$B \subseteq \boldsymbol{B}$ is convex, compact, and satisfies $\inf_{b \in \boldsymbol{B}_0} \nu(b) = \inf_{b \in B \cap \boldsymbol{B}_0} \nu(b)$. Furthermore, for each $b \in B$,*

(i) there exists $Q^b \in \boldsymbol{P}^b$ such that $\ell < \frac{\partial Q^b}{\partial P_1}(z) < u$, almost surely P_1, and

(ii) $\lambda(b)$ solving (2.7) is in the interior of $\{\lambda \,;\, E[|f^*(\lambda^\intercal h(Z,b))| \mid D = 1] < \infty\}$.

That strong duality holds under these conditions is a well known result.[7]

Theorem 2.3.1 (Strong duality)**.** *Suppose assumptions 5 and 6 hold. Then for each $b \in B$, $\nu(b) = V(b)$, with dual attainment.*

The first order condition of the dual problem (2.7) provides intuition. Exchanging expectation and differentiation, the first order condition is

$$\begin{pmatrix} \frac{-p_D}{1-p_D} E_{P_1}[g(Y,X,b)] \\ P(X = x_1 \mid D = 0) \\ \vdots \\ P(X = x_K \mid D = 0) \end{pmatrix} = E_{P_1}\left[(f^*)'\left(\lambda(b)^\intercal h(Y,X,b)\right) \begin{pmatrix} g(Y,X,b) \\ \mathbb{1}\{X = x_1\} \\ \vdots \\ \mathbb{1}\{X = x_K\} \end{pmatrix} \right]$$

where $\lambda(b) \in \mathbb{R}^{d_g + K}$ solves the dual problem. Consider $(f^*)'(\lambda(b)^\intercal h(y,x,b))$ as a density with respect to P_1. Notice that the first d_g equations of the first order condition ensure $p_D E_{P_1}[g(Y,X,b)] + (1-p_D) E_{P_1}[(f^*)'\left(\lambda(b)^\intercal h(Y,X,b)\right) g(Y,X,b)] = 0$, while the remaining K equalities ensure the marginal distribution of X matches P_{0X}. In fact, the proof of theorem 2.3.1 shows that under assumptions 5 and 6, $(f^*)'\left(\lambda(b)^\intercal h(y,x,b)\right)$ is the P_1-density of the solution to the primal.

[7] To the authors knowledge, the first to show strong duality holds under similar conditions was Borwein & Lewis (1991). The proof of theorem 2.3.1, found in appendix 2.7.4, uses a result due to Csiszár et al. (1999).

Assumption 6 ensures the set on which ν is estimated is large enough to estimate the breakdown point, but not so large as to contain parameter values that cannot be rationalized with a well behaved P_1-density. To illustrate, consider again example 2.2.1: Y is a scalar, $\beta = E[Y] = p_D E_{P_1}[Y] + (1 - p_D)E_{P_0}[Y]$, and P_1 is $\mathcal{U}[0,1]$, but for tractability suppose that Kullback-Leibler is used to measure selection. Since P_0 takes values on $[0,1]$, the Manski bounds for β are $\left[\frac{p_D}{2}, 1 - \frac{p_D}{2}\right]$. Appendix 2.7.6.1 shows that strong duality is satisfied whenever $b \in \left(\frac{p_D}{2}, 1 - \frac{p_D}{2}\right)$. Thus for this example, B can be any convex, compact set in the interior of the Manski bounds.

Assumption 6 is maintained throughout the remainder of the paper. Accordingly, the notation ν will be used for the value function of the dual problem as well.

2.4 Estimation

2.4.1 The estimator

The sample analogue of the dual problem provides an estimator of the value function, and suggests a simple plug-in estimator of the breakdown point. The asymptotic properties of these estimators are easier to study if the objective of the dual problem is expressed with a single unconditional expectation, which comes at the cost of additional notation.

First define the matrix $J(D) \equiv \begin{bmatrix} -DI_{d_g} & 0 \\ 0 & (1-D)I_K \end{bmatrix}$ where I_{d_g} and I_K are identity matrices. Notice that $E\left[\frac{J(D)h(DY,X,b)}{(1-p_D)}\right] = c(b)$ and

$$\nu(b) = \sup_{\lambda \in \mathbb{R}^{d_g+K}} E\left[\frac{\lambda^\intercal J(D)h(DY,X,b)}{1-p_D} - \frac{Df^*\left(\lambda^\intercal h(DY,X,b)\right)}{p_D}\right] \tag{2.9}$$

Define

$$\varphi(D, DY, X, b, \lambda, p) \equiv \frac{\lambda^\intercal J(D)h(DY,X,b)}{1-p} - \frac{D}{p} f^*(\lambda^\intercal h(DY,X,b)) \tag{2.10}$$

and observe that the dual problem is $\sup_{\lambda \in \mathbb{R}^{d_g+K}} E[\varphi(D, DY, X, b, \lambda, p_D)]$. The estimator of the value function is defined pointwise by

$$\hat{\nu}_n(b) \equiv \sup_{\lambda \in \mathbb{R}^{d_g+K}} \frac{1}{n} \sum_{i=1}^{n} \varphi(D_i, D_i Y_i, X_i, b, \lambda, \hat{p}_{D,n}) \tag{2.11}$$

where $\hat{p}_{D,n} \equiv \frac{1}{n} \sum_{i=1}^{n} D_i$ estimates p_D. Finally, $\hat{\delta}_n^{BP} \equiv \inf_{b \in B \cap \mathbf{B}_0} \hat{\nu}_n(b)$ estimates the breakdown point.

2.4.2 Asymptotic normality

The following assumption suffices for $\hat{\delta}_n^{BP}$ to be $\sqrt{n}$-consistent and asymptotically normal. First observe that the estimands $\theta_0(b) = (\nu(b), \lambda(b), p_D)$ solve the moment conditions $E[\phi(D, DY, X, b, \theta_0(b))] = 0$, where

$$\phi(D, DY, X, b, \theta) = \phi(D, DY, X, b, v, \lambda, p) = \begin{pmatrix} \varphi(D, DY, X, b, \lambda, p) - v \\ \nabla_\lambda \varphi(D, DY, X, b, \lambda, p) \\ D - p \end{pmatrix}, \tag{2.12}$$

Let $\mathrm{Gr}(\theta_0) \equiv \{(b, \theta(b)) \; ; \; b \in B\}$ denote the graph of θ_0. For $\eta > 0$, the closed η-expansion about this graph is $\mathrm{Gr}(\theta_0)^\eta \equiv \big\{(b, \theta) \in B \times \mathbb{R}^{d_g+K+2} \; ; \; \inf_{(b', \theta') \in \mathrm{Gr}(\theta_0)} \|(b, \theta) - (b', \theta')\| \leq \eta\big\}$.

Assumption 7 (Estimation)**.** *Suppose that*

(i) $\boldsymbol{B}_0$ *is closed,*

(ii) $\min_{b \in B \cap \boldsymbol{B}_0} \nu(b)$ *has a unique solution,*

(iii) the matrix $E[h(Y, X, b) h(Y, X, b)^\intercal \mid D = 1]$ *is nonsingular for each* $b \in B$*,*

(iv) $g(y, x, b)$ *is continuously differentiable with respect to* b *for each* (y, x)*, and*

(v) there exists a convex, compact set Θ^B *containing* $Gr(\theta_0)^\eta$ *for some* $\eta > 0$ *satisfying*

$$\max\left\{E\left[\sup_{(b,\theta)\in\Theta^B}\|\phi(D,DY,X,b,\theta)\|^2\right], E\left[\left(\sup_{(b,\theta)\in\Theta^B}\|\nabla_{(b,\theta)}\phi(D,DY,X,b,\theta)\|\right)^2\right]\right\} < \infty.$$

As previewed in section 2.2.4, $\hat{\delta}_n^{BP}$ is viewed as a two-step estimator where $\hat{\nu}_n$ estimates ν in the first step, and $\hat{\delta}_n^{BP} = \inf_{b\in B\cap\mathbf{B}_0}\hat{\nu}_n(b)$ is a simple plug-in estimator for $\delta^{BP} = \inf_{b\in B\cap\mathbf{B}_0}\nu(b)$. Conditions (iii), (iv), and (v) imply $\sqrt{n}(\hat{\nu}_n - \nu)$ converges weakly in the space of bounded functions on B, to a limiting process that is almost surely continuous. This is shown by linearizing $0 = \frac{1}{n}\sum_{i=1}^n \phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b))$ uniformly over $b \in B$. Conditions (i) and (ii) ensure minimization over $B\cap\mathbf{B}_0$ is a differentiable map on the set of continuous functions of B. The delta method then implies $\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})$ converges in distribution to a normal distribution.

Assumption 7 (i) and (iv) are easily verified by inspection of $\mathbf{B}_0$ and g respectively. Conditions (iii) and (v) are similar to conditions required of generalized empirical likelihood estimators (see, e.g., Antoine & Dovonon (2021) assumption 1 (v) and assumption 3 (iv), (vii)). Assumption 7 (ii) deserves additional scrutiny. When $\mathbf{B}_0$ is a convex set, condition (ii) holds when ν is a strictly convex function. The following lemma shows that this is the case when $g(y,x,b)$ describes a linear model with an occasionally missing outcome.

Lemma 2.4.1 (Convex value function, linear models)**.** *Suppose assumptions 5 and 6 hold, the sample is* $\{D_i, D_iY_i, X_{i1}, X_{i2}\}_{i=1}^n$ *where* $Y_i \in \mathbb{R}$, $X_{i1} \in \mathbb{R}^{d_{x1}}$, *and* $X_{i2} \in \mathbb{R}^{d_{x2}}$, *and the parameter* β *is identified by*

$$E[(Y - X_1^\intercal\beta)X_2] = 0$$

Then $\hat{\nu}_n$ *and* ν *are convex. If in addition* $E[X_2X_1^\intercal]$ *has full column rank, then* ν *is strictly convex.*

Lemma 2.4.1 covers instrumental variable models directly, and ordinary least squares as a

special case (by setting $X_2 = X_1$). It also covers parameters of the form $\beta = E[\tilde{g}(Y, X)]$, because the OLS regression of $\tilde{g}(Y, X)$ on a constant recovers $E[\tilde{g}(Y, X)]$. Simulation evidence presented in appendix 2.7.6 suggests data generating processes and models not covered by lemma 2.4.1 also produce convex ν. Remark 2.4.1 below discusses an approach to relaxing assumption 7 (ii), at the cost of additional complexity.

Theorem 2.4.2 below formally states the convergence in distribution result along with consistency of an estimator of the asymptotic variance. The variance depends on the jacobian term $\Phi(b) \equiv E[\nabla_\theta \phi(D, DY, X, b, \theta_0(b))]$, which is estimated with

$$\hat{\Phi}_n(b) \equiv \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \phi(D, DY, X, b, \hat{\theta}_n(b)), \tag{2.13}$$

where $\hat{\theta}_n(b) \equiv (\hat{\nu}_n(b), \hat{\lambda}_n(b), \hat{p}_{D,n})$ and $\hat{\lambda}_n(b) \equiv \arg\max_{\lambda \in \mathbb{R}^{d_g+K}} \frac{1}{n} \sum_{i=1}^{n} \varphi(D_i, D_i Y_i, X_i, b, \lambda, \hat{p}_{D,n})$. Equations (2.21) and (2.22) in appendix 2.7.1.1 contain expressions for $\nabla_\theta \phi(D, DY, X, b, \theta)$.

Theorem 2.4.2 (Asymptotic normality)**.** *Suppose assumptions 5, 6, and 7 hold. Let* $\hat{b}_n \equiv \arg\min_{b \in B \cap \boldsymbol{B}_0} \hat{\nu}_n(b)$ *and*

$$\hat{\sigma}_n^2 \equiv \frac{1}{n} \sum_{i=1}^{n} \left((\hat{\Phi}_n(\hat{b}_n)^{-1})^{(1)} \phi(D, DY, X, \hat{b}_n, \hat{\theta}_n(\hat{b}_n)) \right)^2$$

where $(\hat{\Phi}_n(\hat{b}_n)^{-1})^{(1)}$ *is the first row of the matrix* $\hat{\Phi}_n(\hat{b}_n)^{-1}$*. Then* $\frac{\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})}{\hat{\sigma}_n} \xrightarrow{d} N(0, 1)$.

2.4.3 Inference

A large breakdown point implies the incomplete distribution P_0 would have to differ greatly from P_1 to rationalize the null hypothesis. If δ^{BD} is larger than the plausible amount of selection in the setting, the null hypothesis is similarly implausible. Skeptical readers following this argument may worry the point estimate $\hat{\delta}_n^{BP}$ is larger than δ^{BP} due to sample noise – but the force of the argument is only strengthened if $\hat{\delta}_n^{BP}$ falls below δ^{BP}.

To address these concerns, researchers should report lower confidence intervals along with point estimates of the breakdown point. Theorem 2.4.2 implies that under assumptions 5, 6, and 7,

$$\widehat{CI}_L \equiv \hat{\delta}_n - \frac{\hat{\sigma}_n}{\sqrt{n}} c_{1-\alpha} \tag{2.14}$$

satisfies $\lim_{n\to\infty} P(\widehat{CI}_L \leq \delta^{BP}) = 1 - \alpha$ when $c_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution.

Remark 2.4.1. Assumption 7 (ii) can be relaxed at the cost of additional complexity. Without assumption 7 (ii), $\sqrt{n}(\hat{\nu}_n - \nu)$ still converges in $\ell^\infty(B)$ to $\mathbb{G}_\nu$, a tight Gaussian process on B, and minimization of a function over $B \cap \mathbf{B}_0$ remains a (Hadamard) directionally differentiable map on the set of continuous functions of B. The delta method continues to imply $\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})$ converges in distribution to $\inf_{b\in\mathbf{m}(\nu)} \mathbb{G}_\nu(b)$, where $\mathbf{m}(\nu)$ is the set of minimizers of ν over $B \cap \mathbf{B}_0$.

Given a bootstrap $\hat{\nu}_n^*$ such that $\sqrt{n}(\hat{\nu}_n^* - \hat{\nu}_n)$ converges weakly in probability conditional on $\{D_i, D_iY_i, X_i\}_{i=1}^n$ to $\mathbb{G}_\nu$, confidence intervals can still be constructed by utilizing the tools developed in Fang & Santos (2019). One approach is to estimate the set $\mathbf{m}(\nu)$ through "near minimizers" of $\hat{\nu}_n$ and using this estimated set to form an estimator of the map $h \mapsto \inf_{b\in\mathbf{m}(\nu)} h(b)$. The confidence interval for δ^{BP} is formed by replacing $c_{1-\alpha}$ in equation (2.14) with the $1-\alpha$ quantile of this estimated function applied to the bootstrap sample; see Fang & Santos (2019) theorem 3.2 and appendix lemma S.4.8. As most cases of interest appear to satisfy assumption 7 (ii), this extension is left for future research.

2.5 Simulations

This section presents simulation results on a variety of different data generating processes. This serves both to illustrate the wide scope of models which can make use of breakdown point analysis and to investigate the finite sample properties of the proposed estimators. In each case, selection is measured using squared Hellinger divergence.

2.5.1 Expectation

Recall example 2.2.1. The parameter of interest is the mean of a scalar random variable Y, $\beta = E[Y] = p_D E_{P_1}[Y] + (1 - p_D)E_{P_0}[Y]$, and the sample is $\{D_i, D_i Y_i\}_{i=1}^n$. The distribution of $Y \mid D = 1$ is the uniform distribution on $[0, 1]$. The probability of observing Y is $p_D = P(D = 1) = 0.7$. To support the claim $H_1 : \beta > 0.4$, let $H_0 : \beta \leq 0.4$. Recall that the true breakdown point, δ^{BP}, of this example is just over 0.2.

The following table summarizes 1,000 simulations for several different sample size.[8]

Table 2.3: Simulations, expectation

n	Bias	St. Dev.	Coverage	Ave. CI Length
1,000	0.005	0.056	98.5	0.090
3,000	0.002	0.032	96.3	0.051
5,000	0.001	0.025	95.8	0.039
10,000	0.001	0.017	95.8	0.028

The simulations show little bias. Coverage is slightly above the targeted 95% significance level in smaller samples.

2.5.2 Linear models

Linear models are the among the most common tools used by empirical researchers. This subsection uses simulations to investigate linear regression with exogenous regressors.[9]

Consider the model

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 Y_2 + \beta_3 X_2 + \varepsilon = W^\intercal \beta + \varepsilon, \tag{2.15}$$

[8] Here Ave. CI Length $= \frac{1}{500} \sum_{s=1}^{500} (\hat{\delta}_{n,s}^{BP} - \widehat{CI}_{L,s})$.

[9] Lemma 2.4.1 shows that when the outcome of a regression is the only missing variable, $\nu(\cdot)$ is convex. Appendix 2.7.6.2 shows simulation evidence that the $\nu(\cdot)$ of the following DGP is convex.

where $W = \begin{pmatrix} 1 & X_1 & Y_2 & X_2 \end{pmatrix}^\intercal$ are the exogenous regressors: $E[W\varepsilon] = 0$. Here Y_1 is a continuous outcome variable, $X_1 = \{0, 1\}$ is the regressor of interest, Y_2 is a continuously distributed control, and $X_2 \in \{0, 1, 2\}$ is a discrete control. The conclusion to be investigated is that the coefficient on X_1 is positive:

$$H_0 \ : \ \beta_1 \leq 0, \qquad\qquad H_1 \ : \ \beta_1 > 0 \tag{2.16}$$

The researcher observes the sample $\{D_i, D_i Y_{1i}, D_i Y_{2i}, X_{1i}, X_{2i}\}_{i=1}^n$ and uses squared Hellinger to measure selection.

The data generating process specification takes inspiration from mincerian wage equations. For worker i, let Y_{1i} be i's log-income, X_{1i} an indicator for i being a college graduate, Y_{2i} be i's work experience, and X_{2i} the number of parents with college degrees (0, 1, or 2). Specifically, let X_2 be multinomial, $X_1 \sim \text{Binomial}\left(\frac{X_2+1}{4}\right)$, and $Y_2 \sim \text{Beta}(3 - X_1, 3)$.[10] Let $\tilde{\varepsilon} \sim U[-1, 1]$ (independent of all other variables), and $\varepsilon = (X_1 + 1)\tilde{\varepsilon}$. The coefficients are specified as $\beta_0 = \beta_1 = \beta_2 = 1$ and $\beta_3 = 0.5$. Finally, Y_1 is generated according to equation (2.15). Notice the support of (Y_1, Y_2, X_1, X_2) is compact, ensuring the moment conditions in assumption 7 (v) are satisfied.

For the missing data process, let $D = \mathbb{1}\{\varepsilon X_1 + 10X_1 + 5(X_2 - 1) > \eta\}$, where $\eta \sim N(\mu_\eta, \sigma_\eta^2)$. The population value of the breakdown point is approximated as the point estimate obtained from a sample with one million observations. This sample reveals $P(D = 1)$ is about 0.71, and suffers from selection. Specifically, ignoring the incomplete observations is equivalent to solving $\frac{1}{n}\sum_{i=1}^n \frac{D_i}{\hat{p}_{D,n}}(Y_i - W_i^\intercal \hat{\beta}_n^{MCAR})W_i = 0$ for $\hat{\beta}_n^{MCAR}$, which results in $\hat{\beta}_n^{MCAR} = (1.04, 1.37, 1.02, 0.42)$. This large sample suggests the breakdown point of the conclusion $\beta_1 > 0$ is about 0.163.

The following table summarizes 1,000 simulations for several different sample sizes.

[10] $P(X_2 = 0) = 0.5$, $P(X_2 = 1) = 0.3$, and $P(X_2 = 2) = 0.2$

Table 2.4: Simulations, OLS

n	Bias	St. Dev.	Coverage	Ave. CI Length
1,000	0.016	0.049	98.7	0.076
3,000	0.007	0.026	95.8	0.041
5,000	0.004	0.019	95.5	0.031
10,000	0.003	0.013	94.5	0.022

Once again the simulations show little bias, with coverage slightly above the targeted 95% significance level in smaller samples.

2.5.3 Logistic regression

The logistic model is a popular choice for estimating the conditional probability of an event. Let $Z = (Z_1, Z_{-1}) \in \{0,1\} \times \mathbb{R}^d$ and suppose that $P(Z_1 = 1 \mid Z_{-1}) = \Lambda(Z_{-1}^\intercal \beta)$, where $\Lambda(t) \equiv \frac{\exp(t)}{1+\exp(t)}$. Since the log-likelihood is concave, estimating this model through maximum likelihood is equivalent to solving the first order condition

$$E[(Z_1 - \Lambda(Z_{-1}^\intercal \beta))Z_{-1}] = 0$$

and so can be viewed as nonlinear GMM, with moment function $g(z,b) = (z_1 - \Lambda(z_{-1}^\intercal b))z_{-1}$.

This simulation considers the model's prediction for $P(Z_1 = 1 \mid Z_{-1} = \bar{z}) = \Lambda(\bar{z}^\intercal \beta)$ for a known $\bar{z}$, and investigates the robustness of the conclusion that this conditional probability is at least 0.5. The corresponding null and alternative hypotheses are

$$H_0 : \Lambda(\bar{z}^\intercal \beta) \leq 0.5, \qquad H_1 : \Lambda(\bar{z}^\intercal \beta) > 0.5 \tag{2.17}$$

Simulation evidence presented in appendix 2.7.6.3 suggests the data generating process described below produces a convex value function. Since H_0 is equivalent to $\bar{z}^\intercal \beta \leq \ln(0.5) -$

$\ln(1-0.5)=0$ and is therefore convex, this suggests that assumption 7 (ii) holds.

The data generating process is one where the outcome is always observed, and the explanatory variables are sometimes missing. Specifically, $Y = Z_{-1} \in \mathbb{R}^3$ is constructed by drawing $\tilde{Y} \sim N(0,\Omega)$ and setting $Y^{(j)} = 2 \times (\Phi(\tilde{Y}^{(j)}) - 0.5)$; the result is that each $Y^{(j)}$ has uniform marginal distributions on $[-1,1]$, and nontrivial covariance matrix. The always-observed variable is the outcome, $X = Z_1$. The true underlying coefficients are $\beta = (1,-1,0.1)$, and the missing data process is conditionally binomial with $P(D=1 \mid X = x, Y = y) = \max\{0.8 - X, Y^{(3)}/2 + 0.5\}$; that is, the probability of a complete observation is at least 0.8 when $X = 1$ and grows weakly with $Y^{(3)}$.

The resulting samples suffer from selection. A sample with one million observations suggests that $P(D=1)$ is about 0.65. Ignoring the incomplete observations is equivalent to solving $\frac{1}{n}\sum_{i=1}^{n} \frac{D}{\hat{p}_{D,n}} g(D_i Y_i, X_i, \hat{\beta}_n^{MCAR}) = 0$, which results in $\hat{\beta}_n^{MCAR} = (1,-1,0.79)$. The estimated squared Hellinger distance between P_{0X} and P_{1X} is 0.076. The covariate value of interest is $\bar{y} = (-0.35,-0.25,0.5)$. The true value for $\Lambda(\bar{y}^\intercal \beta)$ is 0.488, while the estimate assuming using the complete observations of the large sample above is $\Lambda(\bar{y}^\intercal \hat{\beta}_n^{MCAR}) = 0.573$. The point estimate for the breakdown point described by (2.17) using this large sample is 0.108. This is treated as the truth when evaluating the 1,000 simulations per sample size summarized in the following table:

Table 2.5: Simulations, logistic

n	Bias	St. Dev.	Coverage	Ave. CI Length
1,000	0.003	0.018	94.5	0.029
3,000	-0.000	0.010	96.1	0.017
5,000	0.001	0.008	94.8	0.013
10,000	-0.000	0.005	95.9	0.009

These simulations shows essentially zero bias and approximately correct coverage.

2.6 Conclusion

This paper proposes breakdown point analysis as a tractable approach to assessing the sensitivity of a researcher's conclusion to the common MCAR assumption. When defined with squared Hellinger, the breakdown point δ^{BP} has a natural interpretation: if the result were false, the variables under study (Z) would have to predict an observation being selected into the sample (D) at least well enough that $H^2(P_0, P_1) = 1 - E[\sqrt{\text{Var}(D \mid Z)}]/\sqrt{\text{Var}(D)} \geq \delta^{BP}$. Estimators based on the sample analogue of the dual problem are shown $\sqrt{n}$-consistent and asymptotically normal, which facilitates the construction of lower confidence intervals. Researchers working with incomplete datasets should report the breakdown point estimate and lower confidence interval along with standard results, making transparent to their audience how robust the conclusion is to relaxing the MCAR assumption.

2.7 Appendix

2.7.1 Appendix: notation

This appendix summarizes notation and facts used throughout the paper and appendices.

2.7.1.1 Calculations

A number of expressions are useful for verifying conditions in proofs and programming estimators. These are collected here for convenience.

Recall that $\theta_0(b) = (\nu(b), \lambda(b), p_D)$, where $\nu(b)$ is the population value of the value function, $\lambda(b)$ is the corresponding Lagrange multiplier, and $p_D = P(D = 1)$. The notation

$\theta = (v, \lambda, p) \in \mathbb{R}^{d_g+K+2}$ refers to a vector in Euclidean space.

$$\varphi(D, DY, X, b, \lambda, p) \equiv \frac{\lambda^\intercal J(D) h(DY, X, b)}{1-p} - \frac{D}{p} f^*(\lambda^\intercal h(DY, X, b)) \tag{2.18}$$

$$\phi(D, DY, X, b, \theta) = \phi(D, DY, X, b, v, \lambda, p) \equiv \begin{pmatrix} \varphi(D, DY, X, b, \lambda, p) - v \\ \nabla_\lambda \varphi(D, DY, X, b, \lambda, p) \\ D - p \end{pmatrix}, \tag{2.19}$$

$$\Phi(b) = E[\nabla \phi(D, DY, X, b, \theta_0(b))] \tag{2.20}$$

$$\begin{aligned} \nabla_\theta \phi(d, dy, x, b, \theta) &= \nabla_{v,\lambda,p} \begin{pmatrix} \varphi(d, dy, x, b, \lambda, p) - v \\ \nabla_\lambda \varphi(d, dy, b, \lambda, p) \\ d - p \end{pmatrix} \\ &= \begin{bmatrix} -1 & \nabla_\lambda \varphi(d, dy, x, b, \lambda, p)^\intercal & \nabla_p \varphi(d, dy, x, b, \lambda, p) \\ 0 & \nabla^2_\lambda \varphi(d, dy, x, b, \lambda, p) & \nabla_p \nabla_\lambda \varphi(d, dy, x, b, \lambda, p) \\ 0 & 0 & -1 \end{bmatrix} \end{aligned} \tag{2.21}$$

$$\nabla_\lambda \varphi(d, dy, x, b, \lambda, p) = \frac{J(d) h(dy, x, b)}{1-p} - \frac{d}{p} (f^*)'(\lambda^\intercal h(dy, x, b)) h(dy, x, b) \tag{2.22}$$

$$\nabla^2_\lambda \varphi(d, dy, x, b, \lambda, p) = -\frac{d}{p} (f^*)''(\lambda^\intercal h(dy, x, b)) h(dy, x, b) h(dy, x, b)^\intercal$$

$$\nabla_p \varphi(d, dy, x, b, \lambda, p) = \frac{\lambda^\intercal J(d) h(dy, x, b)}{(1-p)^2} + \frac{d}{p^2} f^*(\lambda^\intercal h(dy, x, b))$$

$$\nabla_p \nabla_\lambda \varphi(d, dy, x, b, \lambda, p) = \frac{J(d) h(dy, x, b)}{(1-p)^2} + \frac{d}{p^2} (f^*)'(\lambda^\intercal h(dy, x, b)) h(dy, x, b)$$

2.7.1.2 Graphs

Let $X \subseteq \mathbb{R}^{d_x}$ and $Y \subseteq \mathbb{R}^{d_y}$. For a function $f : X \to Y$, the *graph* of the function refers to the set $\mathrm{Gr}(f) = \{(x, f(x)) \; ; \; x \in X\} \subseteq X \times Y$. Define the closed δ-expansion of the graph

of f:

$$\mathrm{Gr}(f)^\delta \equiv \left\{(x,y) \in X \times Y \; ; \; \inf_{(x',y') \in \mathrm{Gr}(f)} \|(x,y) - (x',y')\| \leq \delta \right\}$$

Note that $\mathrm{Gr}(f)^\delta$ is closed, and bounded if $\mathrm{Gr}(f)$ is bounded.

Given $Z \subseteq \mathbb{R}^{d_z}$ and $g : X \to Z$, one can view (f,g) as a function from X to (Y,Z):

$$(f,g) : X \to (Y,Z), \qquad\qquad (f,g)(x) = (f(x), g(x))$$

Define the graph of this function, $\mathrm{Gr}(f,g) = \{(x, f(x), g(x)) \; ; \; x \in X\} \subseteq X \times Y \times Z$, and the closed δ-expansion about this graph:

$$\mathrm{Gr}(f,g)^\delta = \left\{(x,y,z) \in X \times Y \times Z \; ; \; \inf_{(x',y',z') \in \mathrm{Gr}(f,g)} \|(x,y,z) - (x',y',z')\| \leq \delta \right\}$$

Several easily constructed subsets of $\mathrm{Gr}(f,g)^\delta$ imply useful inequalities. For example,

$$\begin{aligned} \inf_{(x',y',z') \in \mathrm{Gr}(f,g)} \|(x,y,g(x)) - (x',y',z')\| &\leq \inf_{(x',y',z') \in \mathrm{Gr}(f,g)} \|(x,y,g(x)) - (x',y',g(x))\| \\ &= \inf_{(x,y) \in \mathrm{Gr}(f)^\delta} \|(x,y) - (x',y')\| \end{aligned}$$

implies $\left\{(x,y,g(x)) \; ; \; (x,y) \in \mathrm{Gr}(f)^\delta\right\} \subseteq \mathrm{Gr}(f,g)^\delta$. It follows that for a function $h : X \times Y \times Z \to \mathbb{R}$,

$$\sup_{(x,y) \in \mathrm{Gr}(f)^\delta} h(x,y,g(x)) \leq \sup_{(x,y,z) \in \mathrm{Gr}(f,g)^\delta} h(x,y,z).$$

Similarly,

$$\begin{aligned} \|(x,y,z) - (x',y',z')\| &\leq \|(x,y,z) - (x',y',z)\| + \|(x',y',z) - (x',y',z')\| \\ &= \|(x,y) - (x',y')\| + \|z - z'\| \end{aligned}$$

implies that

$$\inf_{(x',y')\in \mathrm{Gr}(f)^{\delta/2},\ z'\in \mathrm{Gr}(g)^{\delta/2}} \|(x,y,z)-(x',y',z')\| \\ \leq \inf_{(x',y')\in \mathrm{Gr}(f)^{\delta/2}} \|(x,y)-(x',y')\| + \inf_{z'\in \mathrm{Gr}(g)^{\delta/2}} \|z-z'\|.$$

It follows that $\left\{(x,y,z)\ ;\ (x,y)\in \mathrm{Gr}(f)^{\delta/2},\ (x,z)\in \mathrm{Gr}(g)^{\delta/2}\right\} \subseteq \mathrm{Gr}(f,g)^{\delta}$, and hence for a function $h : X \times Y \times Z \to \mathbb{R}$,

$$\sup_{(x,y)\in \mathrm{Gr}(f)^{\delta/2},\ z\in \mathrm{Gr}(g)^{\delta/2}} h(x,y,) \leq \sup_{(x,y,z)\in \mathrm{Gr}(f,g)^{\delta}} h(x,y,z).$$

Finally, note that any constant $\bar{y} \in Y$ can be viewed as a trivial function of X. The graph of this function is the set $\mathrm{Gr}(\bar{y}) = \{(x,\bar{y})\ ;\ x \in X\}$ and $\mathrm{Gr}(\bar{y})^{\delta}$ is the set $\{(x,y)\ ;\ x \in X, \|y-\bar{y}\| \leq \delta\}$.

2.7.1.3 Spaces of bounded functions

For any set T, $\ell^{\infty}(T) = \{f : T \to \mathbb{R}\ ;\ \sup_{t\in T}|f(t)| < \infty\}$ denotes the set of real-valued bounded functions on T. $\ell^{\infty}(T)$ is equipped with the sup-norm: for $f \in \ell^{\infty}(T)$, $\|f\|_{\infty} = \|f\|_T = \sup_{t\in T}|f(t)|$. The space of bounded functions taking values in $\mathbb{R}^K$ for some $K \in \mathbb{N}$ is the product space $\ell^{\infty}(T)^K = \underbrace{\ell^{\infty}(T) \times \ldots \times \ell^{\infty}(T)}_{K \text{ times}}$, but can also be viewed as a process on $\ell^{\infty}(T \times \{1,\ldots,K\})$. The latter notation makes it clear that standard empirical process results, typically stated for scalar-valued processes, apply.

If (T,d) is a compact metric space, the extreme value theorem implies the set of continuous functions on T are also bounded and hence form of a subpace of $\ell^{\infty}(T)$. This subspace is denoted

$$\mathcal{C}(T,d) = \{f : T \to \mathbb{R}\ ;\ f \text{ is continuous}\}$$

the notation $\mathcal{C}(T)$ will be used to mean $\mathcal{C}(T, d)$ when the metric d is clear from context.

Some results will refer to subsets of bounded functions whose graphs falls into a particular set. Specifically, let $E^t \subseteq \mathbb{R}^{d_E}$ for each $t \in T$, $E^T = \{(t, e)\ ;\ e \in E^t\}$, and $\ell^\infty(T, E^T)^{d_E}$ be the subset of $\ell^\infty(T)^{d_E}$ whose graph is a subset of E^T:

$$\ell^\infty(T, E^T)^{d_E} = \left\{ g : T \to \mathbb{R}^{d_E}\ ;\ g(t) \in E^t,\ \sup_{t \in T} \|g(t)\| < \infty \right\} \subset \ell^\infty(T)^{d_E}$$

For an example of how this will be used, let $\bar{x} > 0$ and note that the function $f(t, e) = \ln(t+e)$ is uniformly continuous on the set $\{(t, e)\ ;\ t + e \geq \bar{x}\}$. Defining $E^t = \{e \in \mathbb{R}\ ;\ e \geq \bar{x} - t\}$ and E^T as above, we have that $f(t, e)$ is uniformly continuous on this set. This implies that $\tilde{f} : \ell^\infty(T, E^T) \to \ell^\infty(T)$ given by $\tilde{f}(g)(t) = f(t, g(t)) = \ln(t + g(t))$ is continuous (see lemma 2.7.4).

2.7.1.4 Matrices

For a matrix $A \in \mathbb{R}^{J \times K}$, let $\|A\|_o = \sup_{x\ ;\ \|x\|_2 = 1} \|Ax\|_2$ be the operator norm of A, and $\|A\|_{\max} = \max_{ij} |a_{ij}|$, where $a_{ij} \in \mathbb{R}$ is the entry in the i-th row and j-th column of A. Let $\sigma_1(A) \geq \ldots \geq \sigma_K(A) \geq 0$ be the ordered singular values of A. For a square $K \times K$ real matrix A, let $\alpha_1(A) \geq \ldots \geq \alpha_K(A)$ be the ordered eigenvalues of A.

Recall that all norms on finite dimensional real vector spaces are strongly equivalent, meaning that if $\|\cdot\|_1$ and $\|\cdot\|_2$ are any norms on $\mathbb{R}^{J \times K}$, there exist constants $c, C > 0$ such that $c\|A\|_1 \leq \|A\|_2 \leq C\|A\|_1$ for any matrix $A \in \mathbb{R}^{J \times K}$. If $A : T \to \mathbb{R}^{J \times K}$ for some set T, it follows that $E[\sup_t \|A(t)\|] < \infty$ for any norm if and only if $E[\sup_t \|A(t)\|_{\max}] < \infty$. Notice that strong equivalence with $\|\cdot\|_{\max}$ implies that, for any submatrix $\tilde{A}(t)$ of $A(t)$, $E[\sup_t \|A(t)\|] < \infty$ implies $E[\sup_t \|\tilde{A}(t)\|] < \infty$.

Recall that the singular values of a matrix $A \in \mathbb{R}^{J \times K}$ are related to the eigenvalues of

the $K \times K$ square matrix $A^\intercal A$ by $\sigma_k(A) = \sqrt{\alpha_k(A^\intercal A)}$. The operator norm of a matrix is equal to its largest singular value, $\|A\|_o = \sigma_1(A)$, and for invertible matrix A and any $k = 1, \ldots, K$, $\frac{1}{\sigma_k(A)}$ is a singular value of A^{-1}. These imply $\|A^{-1}\|_o = \frac{1}{\sigma_K(A)}$. Finally, for a vector $x \in \mathbb{R}^K$, $\|xx^\intercal\|_o = \|x^\intercal x\|_o = \|x\|_2$.

2.7.2 Appendix: measuring selection and breakdown analysis

2.7.2.1 Measuring selection

Lemma 2.2.1 is found in subsection 2.2.1.

Lemma 2.2.1. *Let $(Z, D) \in \mathbb{R}^{d_z} \times \{0, 1\}$ be random variables with $p_D = P(D = 1) \in (0, 1)$. Let $Z \mid D = 1 \sim P_1$ and $Z \mid D = 0 \sim P_0$. Then*

$$H^2(P_0, P_1) = 1 - \frac{E\left[\sqrt{Var(D \mid Z)}\right]}{\sqrt{Var(D)}} \tag{2.1}$$

where the expectation is taken with respect to $p_D P_1 + (1 - p_D)P_0$, the marginal distribution of Z.

Proof. The marginal, unconditional distribution of Z is $P = p_D P_1 + (1 - p_D)P_0$. This distribution dominates P_1 and P_0, which have densities

$$f_1(z) = \frac{P(D = 1 \mid Z = z)}{p_D}, \qquad f_0(z) = \frac{(1 - P(D = 1 \mid Z = z))}{1 - p_D},$$

with respect to P. This implies

$$\begin{aligned} H^2(P_0, P_1) &= \frac{1}{2}\int \left(\sqrt{f_0(z)} - \sqrt{f_1(z)}\right)^2 dP(z) = \frac{1}{2}\left[\int f_0(z) + f_1(z) - 2\sqrt{f_1(z)f_0(z)}dP(z)\right] \\ &= 1 - \frac{\int \sqrt{P(D=1 \mid Z=z)(1-P(D=1 \mid Z=z))}dP(z)}{\sqrt{p_D(1-p_D)}} \\ &= 1 - \frac{E_P\left[\sqrt{\text{Var}(D \mid Z)}\right]}{\sqrt{\text{Var}(D)}}. \end{aligned}$$

□

2.7.2.2 Nominal identified sets

The exercise proposed in section 2.2.3 can also be understood with a framework of nominally identified sets. This approach to exposition is used in Kline & Santos (2013), Masten & Poirier (2020), and Diegert et al. (2022), and described for the current setting in this appendix.

Under the assumption $d(P_0 \| P_1) \leq \delta$ and $P_0 \ll P_1$, the identified set for β_P is a function of δ:

$$\mathbf{B}_{ID}(\delta) = \left\{ b \in \boldsymbol{B} \; ; \; \exists Q, \;\; p_D \mathbb{E}_{P_1}[g(Z,b)] + (1-p_D)\mathbb{E}_Q[g(Z,b)] = 0, \text{ and } \; d(Q \parallel P_1) \leq \delta \right\} \tag{2.23}$$

Notice $\mathbf{B}_{ID}(\delta)$ is always growing with δ, in the sense that $\delta < \delta' \implies \mathbf{B}_{ID}(\delta) \subseteq \mathbf{B}_{ID}(\delta')$.

The researcher is primarily interested in testing $H_0 \; : \; \beta \in \mathbf{B}_0$ against $H_1 \; : \; \beta \in \mathbf{B}_1 = \mathbf{B} \setminus \mathbf{B}_0$. Naturally, if $\mathbf{B}_{ID}(\delta)$ has trivial intersection with $\mathbf{B}_0$ she is confident in rejecting H_0. This leads to the question "what is the largest value of δ such that $\mathbf{B}_{ID}(\delta)$ has empty intersection with $\mathbf{B}_0$?" Formally, define the **breakdown point** as

$$\bar{\delta}^{BD} = \sup\left\{\delta \in \mathbb{R}_+ \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 = \varnothing\right\} \tag{2.24}$$

if $\mathbf{B}_{ID}(0) \cap \mathbf{B}_0 = \varnothing$, otherwise define $\bar{\delta}^{BD} := 0$.

2.7.2.3 Characterization through a value function

Let

$$\mathbf{P}^b = \{Q \; ; \; Q \ll P_1, \; Q_X = P_{0X}, \; p_D E_{P_1}[g(Z,b)] + (1-p_D)E_Q[g(Z,b)] = 0\},$$

be the set of distributions that "rationalizes" $\beta = b$. Notice that if there exists $Q \in \mathbf{P}^b$ such that $d(Q \parallel P_1) \leq \delta$, then $b \in \boldsymbol{B}_{ID}(\delta)$. This suggests the identified sets can be characterized through the value function

$$\nu(b) = \inf_{Q \in \mathbf{P}_b(Q)} d(Q \parallel P_1), \tag{2.25}$$

where the infimum over the empty set is defined to be $+\infty$. Observe that $\nu(b) < \delta$ implies $b \in \mathbf{B}_{ID}(\delta)$, and if the infimum is attained at some minimum, then $\nu(b) \leq \delta$ if and only if $b \in \mathbf{B}_{ID}(\delta)$.

Lemma 2.7.1 shows that the definition of the breakdown point given in (2.24) is equivalent to that given by (2.4).

Lemma 2.7.1 (Characterization of breakdown point)**.**

$$\inf_{b \in \boldsymbol{B}_0} \nu(b) = \bar{\delta}^{BD}$$

Proof. Define the "robust region" as the set of $\delta \in \mathbb{R}_+$ where the identified set has trivial intersection with the null hypothesis:

$$RR = \{\delta \in \mathbb{R}_+ \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 = \varnothing\}$$

and let $RR^c = \mathbb{R}_+ \setminus RR = \{\delta \in \mathbb{R}_+ \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing\}$ be its compliment in $\mathbb{R}_+$. Notice

that

$$\bar{\delta}^{BD} = \begin{cases} \sup RR & \text{if } RR \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

The proof consists of two steps:

1. Showing that

$$\bar{\delta}^{BD} = \inf RR^c \tag{2.26}$$

where the infimum over the empty set is defined to be ∞.

2. Arguing that

$$\inf_{b \in \boldsymbol{B}_0} \nu(b) \leq \inf RR^c, \qquad \text{and} \qquad \inf_{b \in \boldsymbol{B}_0} \nu(b) \geq \inf RR^c,$$

Step 1. is a consequence of $\mathbf{B}_{ID}(\delta)$ being a growing set (in the sense that $\delta \leq \delta' \implies \mathbf{B}_{ID}(\delta) \subseteq \mathbf{B}_{ID}(\delta')$). Define $\bar{\delta}^* = \inf RR^c = \inf\{\delta \in \mathbb{R}_+ \,;\, \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing\}$. There are three possibilities:

(i) $\delta^{BD} = 0$. Then RR^c contains $(0, \infty)$, hence $0 \leq \bar{\delta}^* = \inf RR^c \leq \inf(0, \infty) = 0$.

(ii) $\delta^{BD} \in (0, \infty)$. Notice that $\delta \leq \delta' \implies \mathbf{B}_{ID}(\delta) \subseteq \mathbf{B}_{ID}(\delta')$ implies that $\delta \leq \delta' \implies (\mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0) \subseteq (\mathbf{B}_{ID}(\delta') \cap \mathbf{B}_0)$, from which it follows that

$$\begin{aligned} &\delta \leq \delta' \text{ and } \delta' \in RR &&\implies &&\delta \in RR \\ &\delta \leq \delta' \text{ and } \delta \in RR^c &&\implies &&\delta' \in RR^c \end{aligned}$$

since $\bar{\delta}^{BD} \in (0, \infty)$, we have RR contains $[0, \bar{\delta}^{BD})$. Similarly, RR^c contains $(\bar{\delta}^*, \infty)$, and since $RR \cap RR^c = \varnothing$, we have $\bar{\delta}^{BD} \leq \bar{\delta}^*$. For $n \in \mathbb{N}$, let $\delta_n \coloneqq \bar{\delta}^* - \frac{1}{n} \geq 0$, and

notice that $\mathbf{B}_{ID}(\delta_n) \cap B = \varnothing$, equivalently, $\delta_n \in RR$. Therefore

$$\bar{\delta}^* - \frac{1}{n} \leq \bar{\delta}^{BD} \leq \bar{\delta}^*$$

let $n \to \infty$ to see that $\delta^{BD} = \delta^*$.

(iii) $\delta^{BD} = \infty$. Then the argument above implies RR contains $[0, \infty)$, so $RR^c = \varnothing$ and $\delta^* = \infty$.

Therefore (2.26) holds.

For step 2., first notice that

$$\inf_{b \in \boldsymbol{B}_0} \nu(b) = \inf_{b \in \boldsymbol{B}_0} \inf_{Q \in \mathbf{P}^b} d(Q \parallel P_1) = \inf \bigcup_{b \in \boldsymbol{B}_0} \left\{ d(Q \parallel P_1) \; ; \; Q \in \mathbf{P}^b \right\} \tag{2.27}$$

If δ is such that $\mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing$, then there exists $b \in \mathbf{B}_0$ and $Q \in \mathbf{P}^b$ such that $d(Q \parallel P_1) \leq \delta$. This implies

$$\inf RR^c = \inf \left\{ \delta \in \mathbb{R}_+ \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing \right\} \geq \inf \bigcup_{b \in \mathbf{B}_0} \left\{ d(Q \parallel P_1) \; ; \; K \in \mathbf{P}^b \right\}$$

Conversely, for each real number a satisfying $a = d(Q \parallel P_1)$ for some $Q \in \mathbf{P}^b$, $b \in \mathbf{B}_0$, we have that $a \in \{\delta \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing\}$. This implies

$$\inf \bigcup_{b \in \mathbf{B}_0} \{ d(Q \parallel P_1) \; ; \; Q \in \mathbf{P}_b \} \geq \inf \{ \delta \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing \} = \inf RR^c$$

Putting (2.26), (2.27), and these two inequalities together we obtain

$$\inf_{b \in \mathbf{B}_0} \nu(b) = \inf \bigcup_{b \in \mathbf{B}_0} \left\{ d(Q \parallel P_1) \; ; \; Q \in \mathbf{P}^b \right\} = \inf \{ \delta \; ; \; \mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing \} = \delta^{BD}$$

as was claimed. □ □

2.7.3 Appendix: additional duality discussion

This short appendix contains no original results, but collects definitions and useful facts related to convex analysis.

2.7.3.1 Definitions

For reference, see Broniatowski & Keziou (2012), or Rockafellar (1970).

Let $f : \mathbb{R} \to (-\infty, \infty]$. The *effective domain* of f is $\text{dom}(f) = \{x \in \mathbb{R} \; ; \; f(x) < \infty\}$. f is called *proper* if $\text{dom}(f)$ is nonempty. f is called *convex* if $\text{dom}(f)$ is a convex set. For a convex $f : E \subsetneq \mathbb{R} \to \mathbb{R}$, f can be extended to $\mathbb{R}$ by setting $f(x) = \infty$ for all $x \notin E$. This extended function is still convex.

Now consider a convex $f : \mathbb{R} \to (-\infty, \infty]$. Notice that convexity implies $\text{dom}(f)$ is a subset of $\mathbb{R}$ with interior of the form (ℓ, u). ℓ or u may be infinite, and $\lim_{x \to \ell^+} f(x)$ or $\lim_{x \to u^-} f(u)$ may be finite. f is called *closed* if 1. $\lim_{x \to \ell^+} f(x) = \infty$ if $\ell > -\infty$, and 2. $\lim_{x \to u^-} f(x) = \infty$ if $u < \infty$. f is called *essentially smooth* if 1. f is differentiable on (ℓ, u), 2. $\lim_{x \to \ell^+} f'(x) = -\infty$ if $\ell > -\infty$, and 3. $\lim_{x \to u^-} f'(x) = \infty$ if $u < \infty$. The *convex conjugate* or *Legendre-Fenchel transform* of a convex function f is defined as $f^*(y) = \sup_{x \in \mathbb{R}} \{xy - f(x)\}$.

2.7.3.2 Results

Now let f be closed, proper, and convex. The following are results not proven here; see footnotes for references.

- f^* is a closed, proper, convex function.[11]
- $(f^*)^* = f$; that is, the convex conjugate of f^* is f.[12]

[11] Rockafellar (1970), p. 104.
[12] Rockafellar (1970) p. 104, theorem 12.2

- f is strictly convex if and only if f^* is essentially smooth.[13]
- f is essentially smooth if and only if f^* is strictly convex.[14]
- If f is strictly convex and essentially smooth, then f' is one-to-one and $(f')^{-1}(y) = (f^*)'(y)$ for all $y \in \text{dom}(f^*)$.[15]
- If f is strictly convex, essentially smooth, and twice differentiable, then f^* is twice differentiable and $(f^*)''(y) = \frac{1}{f''((f')^{-1}(y))}$.[16]
- If f is strictly convex and essentially smooth with $\text{dom}(f) \subseteq [0, \infty)$, then $(f')^{-1}(x) \geq 0$.[17]
- If f is convex, $f(x) = 0$ at $x = 1$, and f is strictly convex on a neighborhood of 1 then $\int f(k(z))dP(z) = 0$ if and only if $k(z) = 1$, P-a.s..[18]

2.7.4 Appendix: proofs of duality results

Lemma 2.7.2 (Unique primal solution)**.** *Suppose f is strictly convex on its domain, $p_D \in (0,1)$, and the infimum in* (2.5) *is finite. Then any solution attaining the infimum in* (2.5) *is unique, P_1-almost surely.*

Proof. Let $Q^0, Q^1 \in \mathbf{P}^b$ attain the finite infimum in (2.5), and let q^0 and q^1 denote their densities with respect to P_1. We have that $\frac{-p_D}{(1-p_D)} E_{P_1}[g(Y,X,b)] = E_{Q^0}[g(Y,X,b)] = E_{Q^1}[g(Y,X,b)]$ and $Q^0_X = Q^1_X = P_{0X}$. For any $\alpha \in (0,1)$, the measure $Q^\alpha = \alpha Q^1 + (1-\alpha)Q^0 \in \mathbf{P}^b$ is feasible in (2.5), and characterized by the P_1-density $\alpha q^1 + (1-\alpha)q^0$.

[13] Borwein & Lewis (1993) p. 251, or Rockafellar (1970) theorem 26.3 on p. 253.
[14] This follows from the two preceding facts.
[15] Broniatowski & Keziou (2012) p. 2559. See also Rockafellar (1970) corollary 23.5.1 on p. 219, corollary 26.3.1 on p. 254, and theorem 26.5 on p. 258.
[16] Broniatowski & Keziou (2012), p. 2559. See also the preceding fact.
[17] Broniatowski & Keziou (2012), p. 2557.
[18] Broniatowski & Keziou (2012), p. 2556.

Suppose for contradiction that Q^0 and Q^1 differ on a set of positive P_1-measure. Strict convexity implies that for any (y, x) in that set,

$$f(\alpha q^1(y,x) + (1-\alpha)q^0(y,x)) < \alpha f(q^1(y,x)) + (1-\alpha)q^0(y,x)$$

Integrating with respect to P_1 reveals $d_f(\alpha Q^1 + (1-\alpha)Q^0 \| P_1) < \alpha d_f(Q^1\|P_1) + (1-\alpha)d_f(Q^0\|P_1)$, contradicting optimality of Q^0, Q^1. □

Lemma 2.7.3 (Weak duality)**.** *Let $\nu(b)$ and $V(b)$ be as defined in* (2.5) *and* (2.7)*, respectively. If assumption 5 holds, then $V(b) \leq \nu(b)$ for any $b \in \boldsymbol{B}$.*

Proof. First note that if $\nu(b) = \infty$ the inequality holds trivially.

Suppose $\nu(b) < \infty$. Then $\mathbf{P}^b \neq \varnothing$, hence there exists at least one density $q(z) = \frac{dQ}{dP_1}(z)$ satisfying $\int h(z,b)q(z)dP_1(z) = c(b)$. Notice that $f^*(r) = \sup_{t\in\mathbb{R}}\{rt - f(t)\}$ implies $f(t) + f^*(r) \geq f(t) + rt - f(t) = rt$. Use this to see that for any $Q \in \mathbf{P}^b$ with P_1-density q,

$$\begin{aligned} f(q(z)) + f^*(\lambda^\intercal h(z,b)) &\geq \lambda^\intercal h(z,b)q(z) \\ \implies f(q(z)) &\geq \lambda^\intercal h(z,b)q(z) - f^*(\lambda^\intercal h(z,b)) \end{aligned}$$

integrating over z with respect to P_1 gives

$$\begin{aligned} &\int f(q(z))dP_1(z) \geq \lambda^\intercal \underbrace{\int h(z,b)q(z)dP_1(z)}_{=c(b)} - \int f^*(\lambda^\intercal h(z,b))dP_1(z) \\ &\implies d_f(Q\|P_1) \geq \lambda^\intercal c(b) - E\left[f^*(\lambda^\intercal h(z,b)) \mid D = 1\right] \end{aligned}$$

the left hand side of the last inequality doesn't depend on $\lambda \in \mathbb{R}^{d_g+K}$, while the right hand

side doesn't depend on $Q \in \mathbf{P}^b$. Hence,

$$\nu(b) = \inf_{Q \in \mathbf{P}^b} d_f(Q \| P_1) \geq \sup_{\lambda \in \mathbb{R}^{d_g + K}} \left\{ \lambda^\intercal c(b) - E\left[f^*(\lambda^\intercal h(Z, b)) \mid D = 1 \right] \right\} = V(b).$$

□

Theorem 2.3.1 (Strong duality)**.** *Suppose assumptions 5 and 6 hold. Then for each $b \in B$, $\nu(b) = V(b)$, with dual attainment.*

Proof. Let $\mathbf{M}$ be the set of measurable functions mapping $z = (x, y) \mapsto \mathbb{R}$. Consider the relaxed problem

$$\tilde{\nu}(b) = \inf_{q \in \tilde{\mathbf{P}}^b} \int f(q(y, x)) dP_1(y, x)$$
$$\tilde{\mathbf{P}}^b = \left\{ q \in \mathbf{M} \, ; \, \int h(y, x, b) q(y, x) dP_1(y, x) = c(b) \right\}$$

for any $q \in \tilde{\mathbf{P}}^b$, $K(\psi) = \int \psi(y, x) q(y, x) dP_1(y, x)$ is a (possibly signed) measure with total measure one. Notice this problem has the same objective as the primal problem (2.5), but a larger feasible set (the set of finite signed measures with total measure one).

Now apply Theorem II.2 of Csiszár et al. (1999), with trivial $\mathcal{K} = \{c(b)\}$. The dual of the relaxed problem is (2.7). Assumption 6 (i) is the "constraint qualification" of Csiszár et al. (1999) Theorem II.2, implying strong duality holds for the relaxed problem, $\tilde{\nu}(b) = V(b)$, and the dual problem's value is attained at a maximum. Let $\lambda(b)$ solve the dual problem. Assumption 6 (ii) allows application of the second part of Theorem II.2, implying the solution to the relaxed problem is given by

$$q^b(y, x) = (f')^{-1}(\lambda(b)^\intercal h(y, x, b)) = (f^*)'(\lambda(b)^\intercal h(y, x, b))$$

By assumption 5 (iv) and Lemma 2.7.2, this solution is unique P_1-almost surely.

Now we show that q^b in fact solves the primal problem, (2.5). Notice q^b is nonnegative, because f' is only defined on the non-negative reals. Furthermore,

$$\begin{aligned}\int q^b(x,y)dP_1(x,y) &= \int \sum_{k=1}^{K} \mathbb{1}\{x = x_k\} q^b(x,y)dP_1(x,y) \\ &= \sum_{k=1}^{K} \int \mathbb{1}\{x = x_k\} q^b(x,y)dP_1(x,y) \\ &= \sum_{k=1}^{K} P(X = x_k \mid D = 0) \\ &= 1\end{aligned}$$

where the third equality follows from $\int h(y,x,b)q^b(x,y)dP_1(x,y) = c(b)$. So the measure Q^b given by $Q^b(\psi) = \int \psi(y,x)q^b(y,x)dP_1(y,x)$ is a probability distribution dominated by P_1. Therefore $Q^b \in \mathbf{P}^b$ is feasible in the primal problem (2.5). Being feasible in the primal and solving the relaxed problem, Q^b must also solve the primal problem. □

2.7.5 Appendix: proofs of estimation results

2.7.5.1 Technical lemmas

These results are self contained, with notation not related to the present paper.

Lemma 2.7.4 (Uniform continuity of maps between bounded functions)**.** *Let T be a set, $E^t \subseteq \mathbb{R}^{d_E}$ for each $t \in T$, $E^T = \{(t,e)\ ;\ t \in T, e \in E^t\}$, and $\ell^\infty(T,E^T)^{d_E}$ be the subset of $\ell^\infty(T)^{d_E}$ whose graph is a subset of E^T:*

$$\ell^\infty(T,E^T)^{d_E} = \left\{ g : T \to \mathbb{R}^{d_E}\ ;\ g(t) \in E^t,\ \sup_{t\in T} \|g(t)\| < \infty \right\} \subset \ell^\infty(T)^{d_E}$$

Let $f : E^T \to \mathbb{R}^{d_f}$ be such that $\sup_{t\in T}\|f(t, g(t))\| < \infty$ for any $g \in \ell^\infty(T, E^T)^{d_E}$, and define

$$\tilde{f} : \ell^\infty(T, E^T)^{d_E} \to \ell^\infty(T)^{d_f}, \qquad \tilde{f}(g)(t) = f(t, g(t)).$$

If $\{f(t, \cdot)\}_{t\in T}$ is uniformly equicontinuous, then $\tilde{f}$ is uniformly continuous.

Proof. Let $\varepsilon > 0$, and use uniform equicontinuity to choose $\delta > 0$ such that

$$|e_1 - e_2| < \delta \implies |f(t, e_1) - f(t, e_2)| < \varepsilon/2$$

for any $t \in T$. Notice that if $g_1, g_2 \in \ell^\infty(T, E^T)^{d_E}$ with $\|g_1 - g_2\|_T = \sup_{t\in T}|g_1(t) - g_2(t)| < \delta$, then

$$\|\tilde{f}(g_1) - \tilde{f}(g_2)\|_T = \sup_{t\in T}|f(t, g_1(t) - f(t, g_2(t))| \leq \varepsilon/2 < \varepsilon$$

and hence $\|g_1 - g_2\|_T < \delta \implies \|\tilde{f}(g_1) - \tilde{f}(g_2)\|_T < \varepsilon$. □

Remark 2.7.1. Lemma 2.7.4 implies many simpler special cases. For example, suppose that for all $t, t' \in T$, $f(t, e) = f(t', e)$ and $E^t = E \subseteq \mathbb{R}$. Then lemma 2.7.4 simplifies to: if $f : E \to \mathbb{R}$ is uniformly continuous, then $\tilde{f} : \ell^\infty(T) \to \ell^\infty(T)$ defined pointwise by $\tilde{f}(g)(t) = f(g(t))$ is continuous.

Lemma 2.7.5 (Restricted infimum is uniformly continuous)**.** *For any $A \subseteq T$ and any $f, g \in \ell^\infty(T)$,*

$$\left|\inf_{t\in A} f(t) - \inf_{t\in A} g(t)\right| \leq \sup_{t\in A}|f(t) - g(t)|$$

as a result, $\iota : \ell^\infty(T) \to \mathbb{R}$ given by $\iota(h) = \inf_{t\in A} h(t)$ is uniformly continuous.

Proof. Notice that

$$\begin{aligned}
\sup_{t\in A} f(t) - \sup_{t\in A} g(t) &\le \sup_{t\in A}\{f(t)-g(t)\} \le \sup_{t\in A}|f(t)-g(t)|, \text{ and} \\
-\left[\sup_{t\in A} f(t) - \sup_{t\in A} g(t)\right] = \sup_{t\in A} g(t) - \sup_{t\in A} f(t) &\le \sup_{t\in A}\{g(t)-f(t)\} \le \sup_{t\in A}|g(t)-f(t)| \\
&= \sup_{t\in A}|f(t)-g(t)|,
\end{aligned}$$

hence $-\sup_{t\in A}|f(t)-g(t)| \le \sup_{t\in A} f(t) - \sup_{t\in A} g(t) \le \sup_{t\in A}|f(t)-g(t)|$, or equivalently

$$\left|\sup_{t\in A} f(t) - \sup_{t\in A} g(t)\right| \le \sup_{t\in A}|f(t)-g(t)|.$$

Use this to see the claimed inequality:

$$\begin{aligned}
\left|\inf_{t\in A} f(t) - \inf_{t\in A} g(t)\right| &= \left|-\sup_{t\in A}\{-f(t)\} - \left(-\sup_{t\in A}\{-g(t)\}\right)\right| = \left|\sup_{t\in A}\{-g(t)\} - \sup_{t\in A}\{-f(t)\}\right| \\
&\le \sup_{t\in A}|-g(t) - \{-f(t)\}| = \sup_{t\in A}|f(t)-g(t)|.
\end{aligned}$$

Regarding the continuity claim, let $\varepsilon > 0$ and set $\delta = \varepsilon$. Then

$$|\iota(f) - \iota(g)| \le \sup_{t\in A}|f(t)-g(t)| \le \sup_{t\in T}|f(t)-g(t)| = \|f-g\|_T,$$

hence $\|f-g\|_T < \delta$ implies $|\iota(f)-\iota(g)| < \varepsilon$. □

Lemma 2.7.6 (Restricted infimum is Hadamard directionally differentiable)**.** *Let (T,d) be a metric space, A a compact subset of T, and*

$$\iota : \ell^\infty(T) \to \mathbb{R}, \qquad\qquad \iota(f) = \inf_{t\in A} f(t)$$

Then ι is Hadamard directionally differentiable at any $f \in \mathcal{C}(T,d)$ tangentially to $\mathcal{C}(T,d)$.

$\Psi_A(f) = \arg\min_{t\in A} f(t)$ *is nonempty, and the directional derivative is given by*

$$\iota'_f : \mathcal{C}(T,d) \to \mathbb{R}, \qquad\qquad \iota'_f(h) = \inf_{t\in\Psi_A(f)} h(t)$$

If $\Psi_A(f)$ *is the singleton* $\{t_f\}$*, then* ι *is fully Hadamard differentiable at* f *tangentially to* $\mathcal{C}(T,d)$ *and* $\iota'_f(h) = h(t_f)$*.*

Proof. The result is essentially a corollary of Fang & Santos (2019) Lemma S.4.9, which shows that $\phi : \ell^\infty(A) \to \mathbb{R}$ given by $\phi(f) = \sup_{t\in A} f(t)$ is Hadamard directionally differentiable at any $f \in \mathcal{C}(A,d)$ tangentially to $\mathcal{C}(A,d)$, with directional derivative

$$\phi'_f : \mathcal{C}(A,d) \to \mathbb{R}, \qquad\qquad \phi'_f(h) = \sup_{t\in\Psi_A(f)} h(t).$$

See Fang & Santos (2019) definition 2.1 for definitions of Hadamard directionally differentiable and fully Hadamard differentiable.

Let $f \in \mathcal{C}(T,d)$ and note that $\Psi_A(f) = \arg\min_{t\in A} f(t)$ is nonempty by the extreme value theorem. Let $\{h_n\}_{n=1}^\infty \subseteq \ell^\infty(T)$ and $\{r_n\}_{n=1}^\infty \subseteq \mathbb{R}_+$ be such that $h_n \to h \in \mathcal{C}(T,d)$ and $r_n \downarrow 0$. For $g \in \ell^\infty(T)$, let $g_A : A \to \mathbb{R}$ be the restriction of g to A, given by $g_A(t) = g(t)$. Observe that $g \in \mathcal{C}(T,d)$ implies $g_A \in \mathcal{C}(A,d)$. Now notice that

$$\begin{aligned}
&\left|\frac{\iota(f + r_n h_n) - \iota(f)}{r_n} - \iota'_a(h)\right| \\
&\quad= \left|\frac{\inf_{t\in A}\{f(t) + r_n h_n(t)\} - \inf_{t\in A} f(t)}{r_n} - \inf_{t\in\Psi_A(f)} h(t)\right| \\
&\quad= \left|\frac{-\sup_{t\in A}\{-f(t) - r_n h_n(t)\} - (-\sup_{t\in A}\{-f(t)\})}{r_n} - \left(-\sup_{t\in\Psi_A(f)}\{-h(t)\}\right)\right| \\
&\quad= \left|\frac{\sup_{t\in A}\{-f(t) + r_n\left(-h_n(t)\right)\} - (\sup_{t\in A}\{-f(t)\})}{r_n} - \left(\sup_{t\in\Psi_A(f)}\{-h(t)\}\right)\right| \\
&\quad= \left|\frac{\phi(-f_A + r_n(-h_{n,A})) - \phi(-f_A)}{r_n} - \phi'_f(-h_A)\right|,
\end{aligned}$$

where the last equality follows from the definitions and the fact that $\Psi_A(f) = \arg\min_{a\in A} f(a) = \arg\max_{a\in A}\{-f_A(a)\}$.

$h_n \to h \in h \in \mathcal{C}(T,d)$ implies $-h_{n,A} \to -h_A \in \mathcal{C}(A,d)$. Thus Fang & Santos (2019) Lemma S.4.9 and the definition of Hadamard directional differentiability implies

$$\lim_{n\to\infty}\left|\frac{\iota(f+t_nh_n)-\iota(f)}{t_n} - \iota'_a(h)\right| = \lim_{n\to\infty}\left|\frac{\phi(-f_A+t_n(-h_{n,A}))-\phi(-f_A)}{t_n} - \phi'_f(-h_A)\right| = 0.$$

Finally, if $\Psi_A(f) = \{t_f\}$ then $\inf_{t\in\Psi_A(f)}(h) = h(t_f)$ is linear in h, and hence ι is fully Hadamard differentiable at f. □

Lemma 2.7.7 (Uniform consistency of estimated moments)**.** *Let* $\mathcal{X} \subseteq \mathbb{R}^{d_X}$, $T \subseteq \mathbb{R}^{d_T}$, $E^t \subseteq \mathbb{R}^{d_E}$, $E^T = \{(t,e)\ ;\ t\in T, e\in E^t\}$,

$$\hat\gamma_n, \gamma : T \to \mathbb{R}^{d_E}, \qquad \text{and} \qquad f : \mathcal{X}\times E^T \to \mathbb{R}^{K\times J}.$$

Suppose that

(i) $\{X_i\}_{i=1}^n$ *is i.i.d.,*

(ii) $\sup_{t\in T}\|\hat\gamma_n(t)-\gamma(t)\| \xrightarrow{p} 0$,

(iii) $Gr(\gamma) = \{(t,\gamma(t))\ ;\ t\in T\}$ *is bounded, and*

(iv) there exists a finite $\varepsilon > 0$ *such that* $(t,e)\mapsto f(x,t,e)$ *is continuous on*

$$Gr(\gamma)^\varepsilon \equiv \left\{(t,e)\in E^T\ ;\ \inf_{(t',e')\in Gr(\gamma)}\|(t,e)-(t',e')\| \le \varepsilon\right\}$$

for all $x\in\mathcal{X}$, *and*

$$E\left[\sup_{(t,e)\in Gr(\gamma)^\varepsilon}\|f(X,t,e)\|\right] < \infty.$$

Then

$$\sup_{t\in T}\left\|\frac{1}{n}\sum_{i=1}^{n} f(X_i,t,\hat{\gamma}_n(t)) - E[f(X,t,\gamma(t))]\right\| \xrightarrow{p} 0$$

Proof. The triangle inequality implies

$$\begin{aligned}
&\sup_{t\in T}\left\|\frac{1}{n}\sum_{i=1}^{n} f(X_i,t,\hat{\gamma}_n(t)) - E[f(X,t,\gamma(t))]\right\| \\
&\qquad \leq \sup_{t\in T}\left\|\frac{1}{n}\sum_{i=1}^{n} f(X_i,t,\hat{\gamma}_n(t)) - E[f(X,t,\hat{\gamma}_n(t))]\right\| \\
&\qquad\qquad + \sup_{t\in T}\|E[f(X,t,\hat{\gamma}_n(t))] - E[f(X,t,\gamma(t))]\|
\end{aligned}$$

Consider the second term first. The dominated convergence theorem, $(t,e)\mapsto f(x,t,e)$ being continuous, and $E\left[\sup_{(t,e)\in \mathrm{Gr}(\gamma)^\varepsilon}\|f(X,t,e)\|\right] < \infty$ implies that

$$\psi : \mathrm{Gr}(\gamma)^\varepsilon \to \mathbb{R}^{K\times J}, \qquad\qquad \psi(t,e) = E[f(X,t,e)]$$

is continuous. $\mathrm{Gr}(\gamma)^\varepsilon$ is a closed and bounded subset of $\mathbb{R}^{d_T}\times\mathbb{R}^{d_E}$, hence compact by the Heine-Borel theorem. Thus ψ is in fact uniformly continuous by the Heine-Cantor theorem. Lemma 2.7.4 then implies

$$\Psi : \ell^\infty(T,\mathrm{Gr}(\gamma)^\varepsilon) \to \ell^\infty(T)^{K\times J}, \qquad\qquad \Psi(g)(t) = \psi(t,g(t))$$

is continuous. $\sup_{t\in T}\|E[f(X,t,\hat{\gamma}_n(t))] - E[f(X,t,\gamma(t))]\| \xrightarrow{p} 0$ follows from $\sup_{t\in T}\|\hat{\gamma}_n(t) - \gamma(t)\| \xrightarrow{p} 0$ and the continuous mapping theorem.

Now consider the first term. Compactness of $\mathrm{Gr}(\gamma)^\varepsilon$, continuity of $(t,e)\mapsto f(x,t,e)$ on $\mathrm{Gr}(\gamma)^\varepsilon$, and $E\left[\sup_{(t,e)\in \mathrm{Gr}(\gamma)^\varepsilon}\|f(X,t,e)\|\right] < \infty$ implies that $\{f(X,t,e)\ ;\ (t,e)\in \mathrm{Gr}(\gamma)^\varepsilon\}$ is Glivenko-Cantelli by van der Vaart (2007) example 19.8. With probability approaching one,

$\sup_{t\in T}\|\hat{\gamma}_n(t)-\gamma(t)\|<\varepsilon$ and when this holds,

$$\sup_{t\in T}\left\|\frac{1}{n}\sum_{i=1}^{n}f(X_i,t,\hat{\gamma}_n(t))-E[f(X,t,\hat{\gamma}_n(t))]\right\| \leq \sup_{(t,g)\in \mathrm{Gr}(\gamma)^{\varepsilon}}\left\|\frac{1}{n}\sum_{i=1}^{n}f(X_i,t,e)-E[f(X,t,e)]\right\| \xrightarrow{p} 0.$$

This concludes the proof. □

Lemma 2.7.8 (Uniform consistency of matrix inverses)**.** *Let* $\hat{\Phi}_n, \Phi : T \to \mathbb{R}^{K\times K}$*. If*

(i) $\Phi(t)^{-1}$ *exists for all* $t \in T$*,*

(ii) $\sup_{t\in T}\|\Phi(t)\|_o<\infty$ *and* $\sup_{t\in T}\|\Phi(t)^{-1}\|_o<\infty$*, and*

(iii) $\sup_{t\in T}\|\hat{\Phi}_n(t)-\Phi(t)\|_o \xrightarrow{p} 0$*,*

then with probability approaching one, the function mapping T *to* $\hat{\Phi}_n(t)^{-1}$ *is well defined and*

$$\sup_{t\in T}\|\hat{\Phi}_n(t)^{-1}-\Phi(t)^{-1}\|_o \xrightarrow{p} 0$$

Proof. It suffices to show that the singular values of $\hat{\Phi}_n(t)$ converge in probability to the singular values of $\Phi(t)$, uniformly over $t \in T$:

$$\sup_{t\in T}\max_{k}|\sigma_k(\hat{\Phi}_n(t))-\sigma_k(\Phi(t))| \xrightarrow{p} 0. \tag{2.28}$$

To see why, notice that $\infty > \sup_{t\in T}\|\Phi(t)^{-1}\|_o = \sup_{t\in T}\frac{1}{\sigma_K(\Phi(t))} = \frac{1}{\inf_{t\in T}\sigma_K(\Phi(t))}$ implies $\varepsilon \equiv \inf_{t\in T}\sigma_K(\Phi(t)) > 0$. Condition (2.28) implies that with probability approaching one,

$$\sup_{t\in T}\max_{k}\left|\sigma_k(\hat{\Phi}_n(t))-\sigma_k(\Phi(t))\right|<\varepsilon/2,$$

and on this event the function mapping T to $\hat{\Phi}_n(t)^{-1}$ is well defined. Then notice that

$$\begin{aligned}\left\|\hat{\Phi}_n(t)^{-1} - \Phi(t)^{-1}\right\|_o &= \left\|\hat{\Phi}_n(t)^{-1}(\Phi(t) - \hat{\Phi}_n(t))\Phi(t)^{-1}\right\|_o \\ &\leq \left\|\hat{\Phi}_n(t)^{-1}\right\|_o \left\|\Phi(t) - \hat{\Phi}_n(t)\right\|_o \left\|\Phi(t)^{-1}\right\|_o\end{aligned}$$

implying

$$\sup_{t\in T}\left\|\hat{\Phi}_n(t)^{-1} - \Phi(t)^{-1}\right\|_o \leq \sup_{t\in T}\left\|\hat{\Phi}_n(t)^{-1}\right\|_o \sup_{t\in T}\left\|\Phi(t) - \hat{\Phi}_n(t)\right\|_o \sup_{t\in T}\left\|\Phi(t)^{-1}\right\|_o \tag{2.29}$$

$\sup_{t\in T}\|\Phi(t)^{-1}\|_o < \infty$ and $\sup_{t\in T}\|\hat{\Phi}_n(t) - \Phi(t)\|_o \xrightarrow{p} 0$ are assumed, the latter implying $\sup_{t\in T}\|\hat{\Phi}_n(t)\|_o = O_p(1)$ by the continuous mapping theorem. Thus (2.29) implies $\sup_{t\in T}\left\|\hat{\Phi}_n(t)^{-1} - \Phi(t)^{-1}\right\|_o \xrightarrow{p} 0$.

The argument that (2.28) holds is broken into three steps:

1. Show that $\hat{\Phi}_n(t)^\intercal\hat{\Phi}_n(t)$ is uniformly consistent for $\Phi(t)^\intercal\Phi(t)$.

   Notice that

$$\begin{aligned}&\sup_{t\in T}\left\|\hat{\Phi}_n(t)^\intercal\hat{\Phi}_n(t) - \Phi(t)^\intercal\Phi(t)\right\|_o \\ &\quad\leq \sup_{t\in T}\left\|\hat{\Phi}_n(t)^\intercal\hat{\Phi}_n(t) - \hat{\Phi}_n(t)^\intercal\Phi(t)\right\|_o + \sup_{t\in T}\left\|\hat{\Phi}_n(t)^\intercal\Phi(t) - \Phi(t)^\intercal\Phi(t)\right\|_o \\ &\quad\leq \sup_{t\in T}\left\|\hat{\Phi}_n(t)^\intercal\right\|_o \sup_{t\in T}\left\|\hat{\Phi}_n(t) - \Phi(t)\right\|_o + \sup_{t\in T}\left\|\hat{\Phi}_n(t)^\intercal - \Phi(t)^\intercal\right\|_o \sup_{t\in T}\|\Phi(t)\|_o\end{aligned}$$

   Recall that for any square matrix $A \in \mathbb{R}^{K\times K}$,

$$\|A^\intercal\|_{\max} = \|A\|_{\max} \leq \|A\|_o \leq K\|A\|_{\max} = K\|A^\intercal\|_{\max}.$$

Use this to see that

$$\sup_{t\in T}\|\hat{\Phi}_n(t)^\intercal\|_o \leq K \sup_{t\in T}\|\hat{\Phi}_n(t)\|_o$$
$$\text{and } \sup_{t\in T}\left\|\hat{\Phi}_n(t)^\intercal - \Phi(t)^\intercal\right\|_o \leq K \sup_{t\in T}\left\|\hat{\Phi}_n(t) - \Phi(t)\right\|_o,$$

and therefore

$$\begin{aligned}&\sup_{t\in T}\left\|\hat{\Phi}_n(t)^\intercal\hat{\Phi}_n(t) - \Phi(t)^\intercal\Phi(t)\right\|_o \\ &\qquad \leq K\left(\sup_{t\in T}\|\hat{\Phi}_n(t)\| + \sup_{t\in T}\|\Phi(t)\|\right)\sup_{t\in T}\|\hat{\Phi}_n(t) - \Phi(t)\| \end{aligned} \tag{2.30}$$

$\sup_{t\in T}\|\Phi(t)^{-1}\|_o < \infty$ and $\sup_{t\in T}\|\hat{\Phi}_n(t) - \Phi(t)\|_o \xrightarrow{p} 0$ by assumption, implying $\sup_{t\in T}\|\hat{\Phi}_n(t)\|_o = O_p(1)$ by the continuous mapping theorem, and thus $\sup_{t\in T}\left\|\hat{\Phi}_n(t)^\intercal\hat{\Phi}_n(t) - \Phi(t)^\intercal\Phi(t)\right\|_o \xrightarrow{p} 0$ by (2.30).

2. Show the eigenvalues of $\hat{\Phi}_n(t)^\intercal\hat{\Phi}_n(t)$ converge to the eigenvalues of $\Phi(t)^\intercal\Phi(t)$ uniformly over $t \in T$.

   Apply Weyl's perturbation theorem, found in Bhatia (1997) as corollary III.2.6: for Hermitian matrices A and B,

$$\max_k|\alpha_k(A) - \alpha_k(b)| \leq \|A - B\|_o$$

   For real matrices Hermitian is equivalent to symmetric, so Weyl's perturbation theorem implies

$$\begin{aligned}&\sup_{t\in T}\max_k|\alpha_k(\hat{\Phi}_n(t)^\intercal\hat{\Phi}_n(t)) - \alpha_k(\Phi(t)^\intercal\Phi(t))| \\ &\qquad \leq \sup_{t\in T}|\hat{\Phi}_n(t)^\intercal\hat{\Phi}_n(t) - \Phi(t)^\intercal\Phi(t)\|_o \xrightarrow{p} 0\end{aligned}$$

   In other words, the eigenvalues of $\hat{\Phi}_n(t)^\intercal\hat{\Phi}_n(t)$ converge to the eigenvalues of $\Phi(t)^\intercal\Phi(t)$

uniformly over $t \in T$. These eigenvalues are the squared singular values of $\Phi_n(t)$.

3. Apply the continuous mapping theorem to conclude (2.28) holds.

   Let $\ell^\infty(T, [0,\infty))$ denote the subset of $\ell^\infty(T)$ consisting of functions h taking nonnegative real values: $h : T \to [0,\infty)$. Lemma 2.7.4 shows that if $f : [0,\infty) \to \mathbb{R}$ is uniformly continuous, then $\tilde{f} : \ell^\infty(T, [0,\infty)) \to \ell^\infty(T)$ given pointwise by $\tilde{f}(h)(t) = f(h(t))$ is continuous. It is well known that the square root funtion $x \mapsto \sqrt{x}$ is uniformly continuous on $[0,\infty)$. Thus (2.28) follows by the continuous mapping theorem.

□

2.7.5.2 Consistency

Lemma 2.7.9 (Unique dual solution)**.** *Suppose assumptions 5 and 6 hold, $b \in B$, and $E[h(Y,X,b)h(Y,X,b)^\intercal \mid D=1]$ is nonsingular. Then $M(b,\lambda) \equiv E[\varphi(D,DY,X,b,\lambda,p_D)]$ is strictly concave in λ and $\lambda(b) = \arg\max_{\lambda \in \mathbb{R}^{d_g+K}} M(b,\lambda)$ is unique.*

Proof. Let $\lambda \neq \tilde{\lambda}$ and $\alpha \in (0,1)$. Since f is essentially smooth, f^* is strictly convex and as a result,

$$f^*((\alpha\tilde{\lambda} + (1-\alpha)\lambda)^\intercal h(y,x,b)) < \alpha f^*(\tilde{\lambda}^\intercal h(y,x,b)) + (1-\alpha) f^*(\lambda^\intercal h(y,x,b)) \tag{2.31}$$

for any (y,x) where $\tilde{\lambda}^\intercal h(y,x,b) \neq \lambda^\intercal h(y,x,b)$, equivalently, where $(\lambda - \tilde{\lambda})^\intercal h(y,x,b) \neq 0$. Since $\lambda - \tilde{\lambda} \neq 0$, nonsingularity of $E[h(Y,X,b)h(Y,X,b)^\intercal \mid D=1]$ implies

$$0 < (\lambda - \tilde{\lambda})^\intercal E[h(Y,X,b)h(Y,X,b)^\intercal \mid D=1](\lambda - \tilde{\lambda}) = E[\big((\lambda - \tilde{\lambda})^\intercal h(Y,X,b)\big)^2 \mid D=1]$$

implies $\left\{(y,x)\ ;\ (\lambda-\tilde{\lambda})^\intercal h(y,x,b)\neq 0\right\}$ is a P_1-nonnegligible set. It follows that

$$E\left[\frac{D}{p_D}f^*((\alpha\tilde{\lambda}+(1-\alpha)\lambda)^\intercal h(DY,X,b))\right]$$
$$<\alpha E\left[\frac{D}{p_D}f^*(\tilde{\lambda}^\intercal h(DY,X,b))\right]+(1-\alpha)E\left[\frac{D}{p_D}f^*(\lambda^\intercal h(Y,X,b))\right]$$

and hence

$$\begin{aligned}
M(b,\alpha\tilde{\lambda}+(1-\alpha)\lambda) &= E[\varphi(D,DY,X,b,\alpha\tilde{\lambda}+(1-\alpha)\lambda,p_D)]\\
&= E\left[\frac{(\alpha\tilde{\lambda}+(1-\alpha)\lambda)^\intercal J(D)h(DY,X,b)}{1-p_D}-\frac{D}{p_D}f^*((\alpha\tilde{\lambda}+(1-\alpha)\lambda)^\intercal h(DY,X,b))\right]\\
&> \alpha E\left[\frac{\tilde{\lambda}^\intercal J(D)h(DY,X,b)}{1-p_D}\right]+(1-\alpha)E\left[\frac{\lambda^\intercal J(D)h(DY,X,b)}{1-p_D}\right]\\
&\quad -\alpha E\left[\frac{D}{p_D}f^*(\tilde{\lambda}^\intercal h(DY,X,b))\right]-(1-\alpha)E\left[\frac{D}{p_D}f^*(\lambda^\intercal h(Y,X,b))\right]\\
&= \alpha M(b,\tilde{\lambda})+(1-\alpha)M(b,\lambda)
\end{aligned}$$

Therefore $M(b,\cdot)$ is strictly concave. $M(b,\cdot)$ attains a maximum by Theorem 2.3.1, and strict concavity implies this maximizer is unique. □

Lemma 2.7.10 (Continuous dual solution and value function)**.** *Suppose assumptions 5, 6, and 7 hold. Then* $\lambda(b)=\arg\max_{\lambda\in\mathbb{R}^{d_g+\kappa}}M(b,\lambda)$*,* $\nu(b)=M(b,\lambda(b))$*, and* $\nabla^2_\lambda M(b,\lambda(b))$ *are all continuous.*

Proof. Jensen's inequality and assumption 7 (v) imply that

$$E\left[\sup_{(b,\nu,\lambda,p)\in\mathrm{Gr}(\theta_0)^\eta}\|\nabla_{(b,\nu,\lambda,p)}\phi(D,DY,X,b,\nu,\lambda,p)\|\right]<\infty$$

and, therefore, the following inequalities as well:

$$E\left[\sup_{(b,\lambda)\in \mathrm{Gr}(\lambda)^\eta} \|\nabla_\lambda \varphi(D, DY, X, b, \lambda, p_D)\|\right] < \infty, \text{ and}$$
$$E\left[\sup_{(b,\lambda)\in \mathrm{Gr}(\lambda)^\eta} \|\nabla^2_\lambda \varphi(D, DY, X, b, \lambda, p_D)\|\right] < \infty$$

where $\mathrm{Gr}(\lambda) = \{(b, \lambda(b)) \; ; \; b \in B\}$ and $\mathrm{Gr}(\lambda)^\eta = \left\{(b, \lambda) \; ; \; \inf_{(b',\lambda')\in \mathrm{Gr}(\lambda)} \|(b, \lambda) - (b', \lambda')\| \leq \eta\right\}$. The dominated convergence theorem implies $M(b, \lambda) = E[\varphi(D, DY, X, b, \lambda, p_D)]$ is twice continuously differentiable with respect to λ in a neighborhood of $\lambda(b)$ for every $b \in B$, with $\nabla_\lambda M(b, \lambda) = E\left[\nabla_\lambda \varphi(D, DY, X, b, \lambda, p_D)\right]$ and $\nabla^2_\lambda M(b, \lambda) = E\left[\nabla^2_\lambda \varphi(D, DY, X, b, \lambda, p_D)\right]$.

$\lambda(b)$ must therefore solve the first order condition

$$0 = \nabla_\lambda M(b, \lambda(b)) = E\left[\nabla_\lambda \varphi(D, DY, X, b, \lambda, p_D)\right].$$

Apply the implicit function theorem to this equation. The maps $(b, \lambda) \mapsto \nabla_\lambda M(b, \lambda)$ and $(b, \lambda) \mapsto \nabla^2_\lambda M(b, \lambda)$ exist and are continuous on an open neighborhood of $(b, \lambda(b))$. Moreover, strict concavity of $M(b, \cdot)$ shown in lemma 2.7.9 implies $\nabla^2_\lambda M(b, \lambda(b))$ is negative definite and hence invertible. It follows from the implicit function theorem (found in Zeidler (1986) as theorem 4.B) that $\lambda(b)$ is continuous in a neighborhood of b. Since this holds for every $b \in B$, the function $\lambda : B \to \mathbb{R}^{d_g+K}$ is continuous.

Assumption 7 (v) and the dominated convergence theorem implies $M(b, \lambda) = E\left[\varphi(D, DY, X, b, \lambda, p_D)\right]$ and $(b, \lambda) \mapsto \nabla^2_\lambda M(b, \lambda)$ are continuous. This implies $\nu(b) = M(b, \lambda(b))$ and $b \mapsto \nabla^2_\lambda M(b, \lambda(b))$ are the composition of continuous functions and hence continuous. □

Lemma 2.7.11 (Uniform consistency of the dual objective)**.** *Suppose assumptions 5, 6, and*

7 hold, and let $\hat{M}_n(b,\lambda) \equiv \frac{1}{n}\sum_{i=1}^n \varphi(D_i, D_iY_i, X_i, b, \lambda, \hat{p}_{D,n})$. *Then*

$$\sup_{(b,\lambda)\in Gr(\lambda)^{\eta/2}} |\hat{M}_n(b,\lambda) - M(b,\lambda)| \xrightarrow{p} 0$$

where

$$Gr(\lambda)^{\eta/2} = \left\{(b,\lambda)\ ;\ \inf_{(b',\lambda')\in Gr(\lambda)} \|(b,\lambda)-(b',\lambda')\| \le \eta/2\right\}$$

and $Gr(\lambda) = \{(b,\lambda(b))\ ;\ b\in B\}$.

Proof. Note that

$$\begin{aligned}&\sup_{(b,\lambda)\in \mathrm{Gr}(\lambda_0)^{\eta/2}} |\hat{\nu}_n(b,\lambda) - \nu(b,\lambda)| \\ &\quad = \sup_{(b,\lambda)\in \mathrm{Gr}(\lambda)^{\eta/2}} \left|\frac{1}{n}\sum_{i=1}^n \varphi(D_i, D_iY_i, X_i, b, \lambda, \hat{p}_{D,n}) - E[\varphi(D, DY, X, b, \lambda, p_D)\right|\end{aligned}$$

and so the claim can be shown by applying technical lemma 2.7.7, with $T \equiv \mathrm{Gr}(\lambda)^{\eta/2}$ indexed by $t = (b,\lambda)$, and the constant map $\gamma(t) = p_D$ for all $t \in T$. Verify the conditions of lemma 2.7.7:

(i) $\{D_i, D_iY_i, X_i\}_{i=1}^n$ is i.i.d. by assumption 5.

(ii) $\sup_{(b,\lambda)\in \mathrm{Gr}(\lambda)^{\eta/2}} |\hat{p}_{D,n} - p_D| = |\hat{p}_{D,n} - p_D| \xrightarrow{p} 0$ by the law of large numbers.

(iii) $\mathrm{Gr}(p_D) = \{(b,\lambda,p_D)\ ;\ (b,\lambda)\in \mathrm{Gr}(\lambda)^{\eta/2}\} = \{(b,\lambda(b),p_D)\ ;\ b\in B\}$ is bounded because λ is continuous (by lemma 2.7.10) and B is compact by assumption 6.

(iv) $p_D \in (0,1)$ implies $\varepsilon \equiv \min\{\min\{p_D, 1-p_D\}, \eta\}/2 > 0$. Let $\mathrm{Gr}(p_D) \equiv \left\{(b,\lambda,p_D)\ ;\ (b,\lambda)\in \mathrm{Gr}(\lambda)^{\eta/2}\right\}$ and

$$\begin{aligned}\mathrm{Gr}(p_D)^{\varepsilon} &\equiv \left\{(b,\lambda,p)\ ;\ \inf_{(b',\lambda',p')\in \mathrm{Gr}(p_D)} \|(b,\lambda,p)-(b',\lambda',p')\| \le \varepsilon\right\} \\ &= \left\{(b,\lambda,p)\ ;\ (b,\lambda)\in \mathrm{Gr}(\lambda)^{\eta/2},\ |p-p_D|\le\varepsilon\right\}.\end{aligned}$$

Observe that

$$(b,\lambda,p) \mapsto \varphi(d,dy,x,b,\lambda,p) = \frac{\lambda^\intercal J(d) h(dy,x,b)}{1-p} - \frac{d}{p} f^*(\lambda^\intercal h(dy,x,b))$$

is continuous on $\mathrm{Gr}(p_D)^\varepsilon$ for each (d,dy,x). Moreover, $(b,\lambda,p) \in \mathrm{Gr}(p_D)^\varepsilon$ implies

$$\begin{aligned}\inf_{(b',\lambda',p')\in \mathrm{Gr}(p_D)} \|(b,\lambda,p) - (b',\lambda',p')\| &\leq \inf_{(b,\lambda)\in \mathrm{Gr}(\lambda)^{\eta/2}} \|(b,\lambda) - (b',\lambda')\| + |p - p_D| \\ &\leq \eta/2 + \varepsilon \leq \eta\end{aligned}$$

and hence $\{(b,\nu(b),\lambda,p)\ ;\ (b,\lambda,p) \in \mathrm{Gr}(p_D)^\varepsilon\} \subseteq \mathrm{Gr}(\theta_0)^\eta$. This implies

$$\begin{aligned}E\left[\sup_{(b,\lambda,p)\in \mathrm{Gr}(p_D)^\varepsilon} |\varphi(D,DY,X,b,\lambda,p)|\right] &\leq E\left[\sup_{(b,v,\lambda,p)\in \mathrm{Gr}(\theta_0)^\eta} |\varphi(D,DY,X,b,\lambda,p)|\right] \\ &\leq E\left[\sup_{(b,v,\lambda,p)\in \Theta^B} |\varphi(D,DY,X,b,\lambda,p)|\right] < \infty.\end{aligned}$$

Thus the result follows from lemma 2.7.7. □

Lemma 2.7.12 (Uniform consistency of the first stage)**.** *Suppose assumptions 5, 6, and 7 hold, and let* $\hat{\lambda}_n(b) = \arg\max_{\lambda\in\mathbb{R}^{d_g+K}} \hat{M}_n(b,\lambda)$*. Then*

$$\sup_{b\in B} \|(\hat{\nu}_n(b), \hat{\lambda}_n(b), \hat{p}_{D,n}) - (\nu(b), \lambda(b), p_D)\| \xrightarrow{p} 0$$

Proof. Let $\Lambda(b) \equiv \{\lambda\ ;\ \|\lambda - \lambda(b)\| \leq \eta/2\}$ and $\bar{\lambda}_n(b) = \arg\max_{\lambda\in\Lambda(b)} \hat{M}_n(b,\lambda)$. The proof consists of three steps:

1. Show $\sup_{b\in B} \|\bar{\lambda}_n(b) - \lambda(b)\| \xrightarrow{p} 0$.

   The following argument shows that for any $\varepsilon > 0$ there exists $\xi > 0$ such that $\sup_{b\in B} M(b,\lambda(b)) - M(b,\bar{\lambda}_n(b)) \leq \xi$ implies $\sup_{b\in B} \|\bar{\lambda}_n(b) - \lambda(b)\| < \varepsilon$, and the probability of the former event converges to one. Let $\varepsilon > 0$, and recall that $M(b,\lambda)$ and

$\lambda(b)$ are continuous by lemma 2.7.10. This implies $M(b,\lambda(b)) - M(b,\lambda)$ is continuous in (b,λ) and

$$\bar{\Lambda}^{B,\varepsilon} \equiv \left\{(b,\lambda) \in \mathrm{Gr}(\lambda)^{\eta/2} \; ; \; \|\lambda - \lambda(b)\| \geq \varepsilon/2\right\}$$

is compact. It follows by the extreme value theorem that $\sup_{(b,\lambda)\in\bar{\Lambda}^{B,\varepsilon}} M(b,\lambda(b)) - M(b,\lambda)$ is attained, say by (b^s,λ^s). Lemma 2.7.9 shows $\lambda(b)$ is the unique maximizer of $M(b,\cdot)$ over $\Lambda(b)$, which is a subset of $\left\{\lambda \; ; \; (b,\lambda) \in \mathrm{Gr}(\lambda)^{\eta/2}\right\}$, and therefore $\xi \equiv M(b^s,\lambda(b^s)) - M(b^s,\lambda^s) > 0$. Observe that $M(b,\lambda(b)) - M(b,\bar{\lambda}_n(b)) < \xi$ implies $\|\bar{\lambda}_n(b) - \lambda(b)\| < \varepsilon/2$, and thus

$$\sup_{b\in B} M(b,\lambda(b)) - M(b,\bar{\lambda}_n(b)) < \xi \qquad \Longrightarrow \qquad \sup_{b\in B} \|\bar{\lambda}_n(b) - \lambda(b)\| \leq \varepsilon/2 < \varepsilon \tag{2.32}$$

Now notice that

$$\begin{aligned}
&\sup_{b\in B} M(b,\lambda(b)) - M(b,\bar{\lambda}_n(b)) \\
&\qquad \leq \sup_{b\in B} \left\{ M(b,\lambda(b)) - \hat{M}_n(b,\lambda(b)) \right\} + \underbrace{\sup_{b\in B} \left\{ \hat{M}_n(b,\lambda(b)) - \hat{M}_n(b,\bar{\lambda}_n(b)) \right\}}_{\leq 0 \text{ by defn of } \bar{\lambda}_n(b)} \\
&\qquad\qquad + \sup_{b\in B} \left\{ \hat{M}_n(b,\bar{\lambda}_n(b)) - M(b,\bar{\lambda}_n(b)) \right\} \\
&\qquad \leq \sup_{b\in B} \left| \hat{M}_n(b,\lambda(b)) - M(b,\lambda(b)) \right| + \sup_{b\in B} \left| \hat{M}_n(b,\bar{\lambda}_n(b)) - M(b,\bar{\lambda}_n(b)) \right| \\
&\qquad \leq 2 \sup_{(b,\lambda)\in \mathrm{Gr}(\lambda)^{\eta/2}} \left| \hat{M}_n(b,\lambda) - M(b,\lambda) \right|.
\end{aligned} \tag{2.33}$$

Lemma 2.7.11 implies that $\sup_{(b,\lambda)\in \mathrm{Gr}(\lambda)^{\eta/2}} \left| \hat{M}_n(b,\lambda) - M(b,\lambda) \right| < \xi/2$ holds with probability approaching one. When it does, (2.32) and (2.33) imply $\sup_{b\in B} \|\bar{\lambda}_n(b) - \lambda(b)\| < \varepsilon$. Therefore $\sup_{b\in B} \|\bar{\lambda}_n(b) - \lambda(b)\| \xrightarrow{p} 0$.

2. Show $\sup_{b\in B}|\hat{M}_n(b,\bar{\lambda}_n(b)) - M(b,\lambda(b))| \xrightarrow{p} 0$.

   The claim follows from lemma 2.7.11, because

$$\begin{aligned}\sup_{b\in B}|\hat{M}_n(b,\bar{\lambda}_n(b)) - M(b,\lambda(b))| &= \sup_{b\in B}\left|\sup_{\lambda\in\Lambda^b}\hat{M}_n(b,\bar{\lambda}_n(b)) - \sup_{\lambda\in\Lambda^b}M(b,\bar{\lambda}_n(b))\right| \\ &\leq \sup_{b\in B}\sup_{\lambda\in\Lambda^b}\left|\hat{M}_n(b,\lambda) - M(b,\lambda)\right| \\ &\leq \sup_{(b,\lambda)\in \mathrm{Gr}(\lambda)^{\eta/2}}\left|\hat{M}_n(b,\lambda) - M(b,\lambda)\right| \xrightarrow{p} 0.\end{aligned}$$

3. Show that with probability approaching one, $\sup_{b\in B}\|\hat{\lambda}_n(b) - \bar{\lambda}_n(b)\| = 0$.

   This follows from an argument similar to the proof of Theorem 2.7 in Newey & McFadden (1994). With probability approaching one, $\sup_{b\in B}\|\bar{\lambda}_n(b) - \lambda(b)\| < \eta/2$ and on this event, $\bar{\lambda}_n(b) \in \mathrm{int}(\Lambda(b)) = \{\lambda\ ;\ \|\lambda - \lambda(b)\| < \eta/2\}$ for every $b \in B$. Since

$$\begin{aligned}\hat{M}_n(b,\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\varphi(D_i, D_iY_i, X_i, b, \lambda, \hat{p}_{D,n}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda^\intercal J(D_i)h(D_iY_i, X_i, b)}{1-\hat{p}_{D,n}} - \frac{D_i}{\hat{p}_{D,n}}f^*(\lambda^\intercal h(D_iY_i, X_i, b))\end{aligned}$$

   is concave in λ, no λ outside of $\mathrm{int}(\Lambda(b))$ could make the objective larger than $\bar{\lambda}_n(b)$. Thus when $\sup_{b\in B}\|\bar{\lambda}_n(b) - \lambda(b)\| < \eta/2$ holds, $\hat{\lambda}_n(b) = \bar{\lambda}_n(b)$ for every $b \in B$ or equivalently, $\sup_{b\in B}\|\hat{\lambda}_n(b) - \bar{\lambda}_n(b)\| = 0$.

□

Theorem 2.7.13 (Consistency of δ^{BP})**.** *Suppose assumptions 5, 6, and 7 hold. Then* $\hat{\delta}_n^{BP} \xrightarrow{p} \delta^{BP}$.

Proof. Lemma 2.7.12 implies $\hat{\nu}_n$ converges in probability to ν in $\ell^\infty(B)$, and lemma 2.7.5 shows $\iota : \ell^\infty(B) \to \mathbb{R}$ given by $\iota(f) = \inf_{b\in B\cap \mathbf{B}_0} f(b)$ is continuous. Since $\hat{\delta}_n^{BP} = \iota(\hat{\nu}_n)$ and $\delta^{BP} = \iota(\nu)$, the result follows from the continuous mapping theorem. □

2.7.5.3 Inference

Lemma 2.7.14 (Bounds on Jacobian terms)**.** *Suppose assumption 5, 6, and 7 hold. Then* $\sup_{b\in B}\|\Phi(b)\|_o < \infty$ *and* $\sup_{b\in B}\|\Phi(b)^{-1}\|_o < \infty$.

Proof. Recall that $\Phi(b) = E[\nabla_\theta \phi(D, DY, X, b, \theta_0(b))]$. Jensen's inequality and convexity of norms implies

$$\begin{aligned}\sup_{b\in B}\|\Phi(b)\|_o = \sup_{b\in B}\|E[\nabla_\theta\phi(D, DY, X, b, \theta)]\|_o &\leq E\left[\sup_{b\in B}\|\nabla_\theta\phi(D, DY, X, b, \theta)\|_o\right] \\ &\leq E\left[\sup_{(b,\theta)\in\Theta^B}\|\nabla_\theta\phi(D, DY, X, b, \theta)\|_o\right],\end{aligned}$$

and $E\left[\sup_{(b,\theta)\in\Theta^B}\|\nabla_\theta\phi(D, DY, X, b, \theta)\|\right] < \infty$ is implied by assumption 7 (v) and Jensen's inequality. Therefore $\sup_{b\in B}\|\Phi(b)\|_o < \infty$.

To establish $\sup_{b\in B}\|\Phi(b)^{-1}\|_o < \infty$, first use expression (2.21) to see that

$$\begin{aligned}\Phi(b) &= E\left[\nabla_\theta\phi(D, DY, X, b, \nu(b), \lambda(b), p_D)\right] \\ &= \begin{bmatrix} -1 & 0 & E[\nabla_p\varphi(D, DY, X, b, \lambda(b), p_D)] \\ 0 & E[\nabla^2_\lambda\varphi(D, DY, X, b, \lambda(b), p_D)] & E[\nabla_p\nabla_\lambda\varphi(D, DY, X, b, \lambda(b), p_D)] \\ 0 & 0 & -1 \end{bmatrix}\end{aligned}$$

where $E[\nabla_\lambda\varphi(D, DY, X, b, \lambda(b), p_D)]^\intercal = 0$ is the first order condition of the dual problem.

The middle matrix, $E[\nabla^2_\lambda\varphi(D, DY, X, b, \lambda(b), p_D)]$, is invertible for each $b \in B$. To see this, first recall that lemma 2.7.10, this matrix equals $\nabla^2_\lambda M(b, \lambda(b)$, which is also continuous in b. The mapping from matrices to eigenvalues is continuous (see Bhatia (1997) corollary III.2.6 or its application in the proof of lemma 2.7.8), so the extreme value theorem implies $\sup_{b\in B}\alpha_1(\nabla^2_\lambda M(b, \lambda(b)))$ is attained by some $\bar{b} \in B$. Lemma 2.7.9 argues that $M(\bar{b}, \lambda)$ is strictly concave in λ, hence $\nabla^2_\lambda M(\bar{b}, \lambda(\bar{b}))$ is negative definite and thus $\alpha_1(\nabla^2_\lambda M(\bar{b}, \lambda(\bar{b}))) < 0$.

To summarize,

$$\sup_{b\in B}\alpha_1(\nabla^2_\lambda M(b,\lambda(b))) = \alpha_1(\nabla^2_\lambda M(\bar{b},\lambda(\bar{b}))) < 0$$

which implies $\nabla M(b,\lambda(b)) = E[\nabla^2_\lambda \varphi(D,DY,X,b,\lambda(b),p_D)]$ is invertible for each $b \in B$. With this invertibility claim, it is straightforward to verify that for each $b \in B$, $\Phi(b)^{-1}$ exists and is given by

$$\Phi(b)^{-1} = \begin{bmatrix} -1 & 0 & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & -1 \end{bmatrix}$$

where

$$\begin{aligned} A_{13} &\equiv -E[\nabla_p \varphi(D,DY,X,b,\lambda(b),p_D)] \\ A_{22} &\equiv E[\nabla^2_\lambda \varphi(D,DY,X,b,\lambda(b),p_D)]^{-1} \\ A_{23} &\equiv E[\nabla^2_\lambda \varphi(D,DY,X,b,\lambda(b),p_D)]^{-1} E[\nabla_p \nabla_\lambda \varphi(D,DY,X,b,\lambda(b),p_D)] \end{aligned}$$

To see that $\sup_{b\in B}\|\Phi(b)^{-1}\|_o$ is finite, first recall that for conformable matrices,

$$\left\| \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \right\|_o \leq \|A_{11}\|_o + \|A_{12}\|_o + \|A_{21}\|_o + \|A_{22}\|_o, \text{ and}$$

$$\|AB\|_o \leq \|A\|_o \|B\|_o.$$

Apply these inequalities to find that

$$\begin{aligned}
&\sup_{b\in B}\|\Phi(b)^{-1}\|_o \\
&\quad\leq 2+\sup_{b\in B}|E[\nabla_p\varphi(D,DY,X,b,\lambda(b),p_D)]|+\sup_{b\in B}\|E[\nabla^2_\lambda\varphi(D,DY,X,b,\lambda(b),p_D)]^{-1}\|_o \\
&\qquad+\sup_{b\in B}\|E[\nabla^2_\lambda\varphi(D,DY,X,b,\lambda(b),p_D)]^{-1}\|_o\times\sup_{b\in B}\|E[\nabla_p\nabla_\lambda\varphi(D,DY,X,b,\lambda(b),p_D)]\|
\end{aligned} \tag{2.34}$$

$|E[\nabla_p\varphi(D,DY,X,b,\lambda(b),p_D)]|$ and $\|E[\nabla_p\nabla_\lambda\varphi(D,DY,X,b,\lambda(b),p_D)]\|$ are the operator norms of submatrices of $\Phi(b)$. Thus $\sup_{b\in B}\|\Phi(b)\|_o<\infty$, argued above, implies

$$\sup_{b\in B}|E[\nabla_p\varphi(D,DY,X,b,\lambda(b),p_D)]|<\infty,\text{ and }\sup_{b\in B}\|E[\nabla_p\nabla_\lambda\varphi(D,DY,X,b,\lambda(b),p_D)]\|<\infty. \tag{2.35}$$

Finally, since $E[\nabla^2_\lambda\varphi(D,DY,X,b,\lambda(b),p_D)]=\nabla^2_\lambda M(b,\lambda(b))$ is symmetric and negative definite, $\|E[\nabla^2_\lambda\varphi(D,DY,X,b,\lambda(b),p_D)]^{-1}\|_o=\|\nabla^2_\lambda M(b,\lambda(b))^{-1}\|_o=\frac{1}{|\alpha_1(\nabla^2_\lambda M(b,\lambda(b)))|}$ and

$$\begin{aligned}
&\sup_{b\in B}\|E[\nabla^2_\lambda\varphi(D,DY,X,b,\lambda(b),p_D)]^{-1}\|_o \\
&\qquad=\sup_{b\in B}\|\nabla^2_\lambda M(b,\lambda(b))^{-1}\|_o=\sup_{b\in B}\frac{1}{|\alpha_1(\nabla^2_\lambda M(b,\lambda(b)))|} \\
&\qquad=\frac{1}{\inf_{b\in B}|\alpha_1(\nabla^2_\lambda M(b,\lambda(b)))|}=\frac{1}{|\sup_{b\in B}\alpha_1(\nabla^2_\lambda M(b,\lambda(b)))|}<\infty,
\end{aligned} \tag{2.36}$$

where the final claim follows from $\sup_{b\in B}\alpha_1(\nabla^2_\lambda M(b,\lambda(b)))<0$ as argued above. Taken together, (2.34), (2.35), and (2.36) show that $\sup_{b\in B}\|\Phi(b)^{-1}\|_o<\infty$. □

Lemma 2.7.15 (Donsker influence functions)**.** *Suppose assumptions 5, 6, and 7 hold. Then the class of functions*

$$\left\{\phi(D,DY,X,b,\theta)\ ;\ (b,\theta)\in\Theta^B\right\}$$

is Donsker.

Proof. By verifying the conditions of van der Vaart (2007) example 19.7.

Θ^B is a compact subset of a finite dimensional space, hence bounded. Let $(b_1, \theta_1), (b_2, \theta_2) \in \Theta^B$ and apply the mean value inequality (e.g., Coleman (2012) Corollary 3.2) to find

$$\begin{aligned}
&\|\phi(d, dy, x, b_1, \theta_1) - \phi(d, dy, x, b_2, \theta_2)\| \\
&\quad \leq \left[\sup_{t \in (0,1)} \left\| \nabla_{(b,\theta)} \phi(d, dy, x, tb_1 + (1-t)b_2, t\theta_1 + (1-t)\theta_2) \right\|_o \right] \|(b_1, \theta_1) - (b_2, \theta_2)\| \\
&\quad \leq \left[\sup_{(b,\theta) \in \Theta^B} \left\| \nabla_{(b,\theta)} \phi(d, dy, x, b, \theta) \right\|_o \right] \|(b_1, \theta_1) - (b_2, \theta_2)\|
\end{aligned}$$

Assumption 7 (v) includes $E\left[\left(\sup_{(b,\theta)\in\Theta^B} \left\|\nabla_{(b,\theta)}\phi(d, dy, x, b, \theta)\right\|_o\right)^2\right] < \infty$. Therefore the class $\left\{\phi(D, DY, X, b, \theta)\ ;\ (b, \theta) \in \Theta^B\right\}$ is a special case of van der Vaart (2007) example 19.7, and thus Donsker. □

Lemma 2.7.16 (Weak convergence of the first stage)**.** *Suppose assumptions 5, 6, and 7 hold. Let* $\mathcal{I} = \{1, \ldots, d_g + K + 2\}$*, and view* $\hat{\theta}_n(b) \equiv (\hat{\nu}_n(b), \hat{\lambda}_n(b), \hat{p}_{D,n})$ *and* $\theta_0(b) \equiv (\nu(b), \lambda(b), p_D)$ *as functions mapping* $B \times \mathcal{I}$ *to* $\mathbb{R}$*. Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} \mathbb{G}$$

where $\mathbb{G}$ *is a tight, mean zero Gaussian process in* $\ell^\infty(B \times \mathcal{I})$*. The covariance function of* $\mathbb{G}$ *is given by*

$$\begin{aligned}
&Cov(\mathbb{G}(b_1, i_1), \mathbb{G}(b_2, i_2)) \\
&\quad = E\left[(\Phi(b_1)^{-1})^{(i_1)} \phi(D, DY, X, b_1, \theta_0(b_1)) \left\{(\Phi(b_2)^{-1})^{(i_2)} \phi(D, DY, X, b_2, \theta_0(b_2))\right\}\right].
\end{aligned}$$

where $(\Phi(b)^{-1})^{(i)}$ *is the* i*-th row of the matrix* $\Phi(b)^{-1} = E[\nabla_\theta \phi(D, DY, X, b, \theta_0(b))]^{-1}$*.*

Proof. For legibility, the proof is presented in six steps:

1. Mean value theorem.

   For each $b \in B$, apply the mean value theorem to each coordinate of $0 = \frac{1}{n}\sum_{i=1}^{n} \phi(D_i, D_iY_i, X_i, b, \hat{\theta}_n(b))$ and stack the results to obtain

$$0 = \frac{1}{n}\sum_{i=1}^{n} \phi(D_i, D_iY_i, X_i, b, \theta_0(b)) + \underbrace{\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \phi^{(1)}(D_i, D_iY_i, X_i, b, \bar{\theta}_n^1(b)) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \phi^{(d_g+K+2)}(D_i, D_iY_i, X_i, b, \bar{\theta}_n^{d_g+K+2}(b)) \end{bmatrix}}_{\equiv \bar{\Phi}_n(b)} (\hat{\theta}_n(b) - \theta_0(b)) \quad (2.37)$$

   where $\nabla_\theta \phi^{(j)}(D_i, D_iY_i, X_i, b, \theta)$ is the j-th coordinate of the vector $\nabla_\theta \phi(D_i, D_iY_i, X_i, b, \theta)$, and $\bar{\theta}_n^j = \theta_0(b) + a_n^j(b) \times (\hat{\theta}_n(b) - \theta_0(b)) \in \mathbb{R}^{d_g+K+2}$ for some $a_n^j(b) \in (0,1)$.[19] Notice

$$\begin{aligned} \|\bar{\theta}_n^j(b) - \theta_0(b)\| &= \left\| \theta_0(b) + a_n^j(b) \times (\hat{\theta}_n(b) - \theta_0(b)) - \theta_0(b) \right\| \\ &= a_n^j(b) \times \|\hat{\theta}_n(b) - \theta_0(b)\| \\ &\leq \|\hat{\theta}_n(b) - \theta_0(b)\| \end{aligned}$$

   and $\sup_{b\in B}\|\hat{\theta}_n(b) - \theta_0(b)\| \xrightarrow{p} 0$, shown in lemma 2.7.12, implies $\sup_{b\in B}\|\bar{\theta}_n^j(b) - \theta_0(b)\| \xrightarrow{p} 0$.

2. Show $\sup_{b\in B}\|\bar{\Phi}_n(b) - \Phi(b)\| \xrightarrow{p} 0$.

   First, notice that

$$\bar{\Phi}_n(b) = \sum_{j=1}^{d_g+K+2} e_{jj} \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \phi(D_i, D_iY_i, X_i, b, \bar{\theta}_n^j(b))$$

   where e_{jj} is the square $(d_g + K + 2) \times (d_g + K + 2)$ matrix whose (j, j)-th entry is one

[19] See Newey & McFadden (1994) footnote 25.

and all other entries are zero.[20]

Apply lemma 2.7.7 to $\frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \phi(D_i, D_i Y_i, X_i, b, \bar{\theta}_n^j(b))$ for each $j \in \{1, \ldots, d_g + K + 2\}$ to argue this is consistent for $E[\nabla_\theta \phi(D, DY, X, b, \theta_0(b))]$ uniformly over $b \in B$.

(i) $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ is i.i.d. by assumption 5.

(ii) $\sup_{b \in B} \|\bar{\theta}_n^j(b) - \theta_0(b)\| \xrightarrow{p} 0$ is shown in step 1.

(iii) $\theta_0(b) = (\nu(b), \lambda(b), p_D)$ is bounded, since B is compact by assumption 6 and $\nu(\cdot)$, $\lambda(\cdot)$ are continuous as shown by lemma 2.7.10.

(iv) $(b, \theta) \mapsto \nabla_\theta \phi(d, dy, x, b, \theta)$ is continuous at any $(b, \theta) \in \mathrm{Gr}(\theta_0)^\eta$, by examination of equations (2.21) and (2.22). Moreover, $E\left[\sup_{(b,\theta) \in \mathrm{Gr}(\theta_0)^\eta} \|\nabla_\theta \phi(D, DY, X, b, \theta)\|_o\right]$ is finite; $\nabla_\theta \phi(D, DY, X, b, \theta)$ is a submatrix of $\nabla_{(b,\theta)} \phi(D, DY, X, b, \theta)$, while assumption 7 (v) and Jensen's inequality imply

$$E\left[\sup_{(b,\theta) \in \mathrm{Gr}(\theta_0)^\eta} \|\nabla_{(b,\theta)} \phi(D, DY, X, b, \theta)\|_o\right] \leq E\left[\sup_{(b,\theta) \in \Theta^B} \|\nabla_{(b,\theta)} \phi(D, DY, X, b, \theta)\|_o\right] < \infty.$$

So by Lemma 2.7.7,

$$\sup_{b \in B} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \phi(D_i, D_i Y_i, X_i, b, \bar{\theta}_n^j(b)) - E[\nabla_\theta \phi(D, DY, X, b, \theta_0(b))] \right\|_o \xrightarrow{p} 0$$

[20] When premultiplying a square matrix A, e_{jj} "selects" the j-th row. For example,

$$e_{22} A = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & \ldots & a_{1J} \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2J} \\ a_{31} & a_{32} & a_{33} & \ldots & a_{3J} \\ \vdots & & & \ddots & \vdots \\ a_{J1} & a_{K2} & a_{K3} & \ldots & a_{JJ} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 \\ a_{21} & a_{22} & a_{23} & \ldots & a_{2J} \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

for each $j \in \{1, \ldots, d_g + K + 2\}$, from which it follows that

$$
\begin{aligned}
&\sup_{b \in B} \|\bar{\Phi}_n(b) - \Phi(b)\|_o \\
&= \sup_{b \in B} \left\| \sum_{j=1}^{d_g+K+2} e_{jj} \left(\frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \phi(D_i, D_i Y_i, X_i, b, \bar{\theta}_n^j(b)) - E[\nabla_\theta \phi(D, DY, X, b, \theta_0(b))] \right) \right\|_o \\
&\leq \sum_{j=1}^{d_g+K+2} \sup_{b \in B} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \phi(D_i, D_i Y_i, X_i, b, \bar{\theta}_n^j(b)) - E[\nabla_\theta \phi(D, DY, X, b, \theta_0(b))] \right\|_o \\
&\leq (d_g + K + 2) \\
&\qquad \times \max_j \sup_{b \in B} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \phi(D_i, D_i Y_i, X_i, b, \bar{\theta}_n^j(b)) - E[\nabla_\theta \phi(D, DY, X, b, \theta_0(b))] \right\|_o \\
&\xrightarrow{p} 0
\end{aligned}
$$

3. Uniform linearization.

   Lemma 2.7.14 shows $\sup_{b \in B} \|\Phi(b)\|_o < \infty$ and $\sup_{b \in B} \|\Phi(b)^{-1}\|_o < \infty$. Since with $\sup_{b \in B} \|\bar{\Phi}_n(b) - \Phi(b)\|_o \xrightarrow{p} 0$ is shown in step 2, lemma 2.7.8 implies that with probability approaching one, $\bar{\Phi}_n(b)^{-1}$ is well defined as a function on B. When it is, rearrange expression (2.37) to find

$$
\begin{aligned}
\sqrt{n}(\hat{\theta}_n(b) - \theta_0(b)) &= \bar{\Phi}_n(b)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(D_i, D_i Y_i, X_i, b, \theta_0(b)) \\
&= G_n(b) + R_n(b) \\
\text{where } G_n(b) &= \Phi(b)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(D_i, D_i Y_i, X_i, b, \theta_0(b)), \\
\text{and } R_n(b) &= \left[\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(D_i, D_i Y_i, X_i, b, \theta_0(b)).
\end{aligned}
$$

4. Show $G_n \xrightarrow{L} \mathbb{G}$ in $\ell^\infty(B \times \mathcal{I})$.

Define $\tilde{G}_n : B \to \mathbb{R}^{d_g+K+2}$ pointwise as

$$\tilde{G}_n(b) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(D_i, D_i Y_i, X_i, b, \theta_0(b))$$

$\{\phi(D, DY, X, b, \theta_0(b)) \; ; \; b \in B\}$ is a subset of the class considered in lemma 2.7.15, and is therefore Donsker (see van der Vaart & Wellner (1997) theorem 2.10.1). Thus, $\tilde{G}_n \xrightarrow{L} \tilde{\mathbb{G}}$ in $\ell^\infty(B)^{d_g+K+2}$, where $\tilde{\mathbb{G}}$ is a tight, mean-zero Gaussian process with covariance function

$$\text{Cov}(\tilde{\mathbb{G}}(b_1), \tilde{\mathbb{G}}(b_2)) = E\left[\phi(D, DY, X, b_1, \theta_0(b_1))\phi(D, DY, X, b, \theta_0(b_2))^\intercal\right]$$

Now define

$$L : \ell^\infty(B)^{d_g+K+2} \to \ell^\infty(B \times \mathcal{I}), \qquad L(H)(b, i) = (\Phi(b)^{-1})^{(i)} H(b)$$

and observe that $G_n = L(\tilde{G}_n)$. Note that L is a linear operator on H. Lemma 2.7.14 shows $\sup_{b \in B} \|\Phi(b)^{-1}\|_o < \infty$, which along with

$$\|LH\|_B = \sup_{b \in B} \|\Phi(b)^{-1} H(b)\| \leq \sup_{b \in B} \|\Phi(b)^{-1}\|_o \sup_{b \in B} \|H(b)\| = \left(\sup_{b \in B} \|\Phi(b)^{-1}\|_o \right) \|H\|_B$$

shows that L is bounded, hence continuous. The continuous mapping theorem then implies

$$L(\tilde{G}_n) \xrightarrow{L} L(\tilde{\mathbb{G}})$$

where $L(\tilde{\mathbb{G}})$ is a tight, mean-zero Gaussian process on $\ell^\infty(B \times \mathcal{I})$. Letting $(\Phi(b))^{(i)}$ be the i-th row of the matrix $\Phi(b)^{-1}$, the covariance function of $L(\tilde{\mathbb{G}})$ is

$$\begin{aligned}
&\text{Cov}(\mathbb{G}(b_1, i_1), \mathbb{G}(b_2, i_2)) \\
&\quad = E\left[(\Phi(b_1)^{-1})^{(i_1)} \phi(D, DY, X, b_1, \theta_0(b_1)) \left\{ (\Phi(b_2)^{-1})^{(i_2)} \phi(D, DY, X, b_2, \theta_0(b_2)) \right\} \right].
\end{aligned}$$

Notice that the marginals of $L(\tilde{\mathbb{G}})$ are equal in distribution to those of $\mathbb{G}$. By van der Vaart & Wellner (1997) lemma 1.5.3, this implies the two distributions are the same and hence $G_n = L(\tilde{G}_n) \xrightarrow{L} \mathbb{G}$.

5. Uniform linearization remainder control.

   Since $\{\phi(D, DY, X, b, \theta_0(b)) \; ; \; b \in B\}$ is Donsker and $E[\phi(D, DY, X, b, \theta_0(b))] = 0$ for all $b \in B$, $\sup_{b\in B} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(D_i, D_i Y_i, X_i, b, \theta_0(b)) \right\|_o = O_p(1)$ by the continuous mapping theorem. Lemma 2.7.14 shows $\|\Phi(b)\|_o < \infty$ and step 2 that $\|\Phi(b)^{-1}\|_p < \infty$, and $\|\bar{\Phi}_n(b) - \Phi(b)\| \xrightarrow{p} 0$, so lemma 2.7.8 implies $\sup_{b\in B} \left\|\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\right\| = o_p(1)$. Thus,

$$\begin{aligned}
\sup_{b\in B} \|R_n(b)\| &= \sup_{b\in B} \left\| \left[\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\right] \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(D_i, D_i Y_i, X_i, b, \theta_0(b)) \right\| \\
&\leq \sup_{b\in B} \left\|\bar{\Phi}_n(b)^{-1} - \Phi(b)^{-1}\right\| \sup_{b\in B} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(D_i, D_i Y_i, X_i, b, \theta_0(b)) \right\| \\
&\xrightarrow{p} 0.
\end{aligned}$$

6. Conclusion.

   As elements of $\ell^\infty(B \times \mathcal{I})$, $G_n \xrightarrow{L} \mathbb{G}$ and $R_n \xrightarrow{p} 0$, so

$$(G_n, R_n) \xrightarrow{L} (\mathbb{G}, 0) \qquad \text{in } \ell^\infty(B \times \mathcal{I})$$

   by van der Vaart (2007) theorem 18.10. The continuous mapping theorem (van der Vaart (2007) theorem 18.11) then implies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = G_n + R_n \xrightarrow{L} \mathbb{G} + 0$$

   which concludes the proof.

□

Lemma 2.7.17 (Support of $\mathbb{G}_\nu$)**.** *Suppose assumptions 5, 6, and 7 hold, let $\mathbb{G}$ be the the random element of $\ell^\infty(B \times \mathcal{I})$ from lemma 2.7.16, and let $\mathbb{G}_\nu \in \ell^\infty(B)$ be the mean-zero Gaussian process on B defined pointwise by $\mathbb{G}_\nu(b) = \mathbb{G}(b, 1)$. Then $\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu$ in $\ell^\infty(B)$ and $P(\mathbb{G}_\nu \in \mathcal{C}(B)) = 1$, where $\mathcal{C}(B)$ is the set of continuous functions defined on B.*

Proof. Lemma 2.7.16 and the continuous mapping thoerem implies $\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu$. The Portmanteau theorem (van der Vaart & Wellner (1997) theorem 1.3.4) shows that this is equivalent to

$$\limsup_{n\to\infty} P(\sqrt{n}(\hat{\nu}_n - \nu) \in F) \leq P(\mathbb{G}_\nu \in F)$$

for all closed sets $F \subseteq \ell^\infty(B)$. Since $\mathcal{C}(B)$ is closed and $\nu(\cdot)$ is continuous by lemma 2.7.10, it suffices to show that $\hat{\nu}_n$ is continuous with probability approaching one.

The argument is based on the Berge maximum theorem (Aliprantis & Border (2006) theorem 17.31). Recall $\hat{\lambda}_n(b) \equiv \arg\max_{\lambda \in \mathbb{R}^{d_g+K}} \hat{M}_n(b, \lambda)$ and $\hat{\nu}_n(b) = \hat{M}_n(b, \hat{\lambda}_n(b))$. Let $\Lambda(b) \equiv \{\lambda \,;\, \|\lambda - \lambda(b)\| \leq \eta/2\}$. Lemma 2.7.12 implies $\sup_{b\in B} \left\|\hat{\lambda}_n(b) - \lambda(b)\right\| < \eta/2$ holds with probability approaching one, and when it does,

$$\sup_{b\in B} |\hat{\nu}_n(b) - \max_{\lambda\in\Lambda(b)} \hat{M}_n(b, \lambda)| = \sup_{b\in B} \left| \sup_{\lambda\in\mathbb{R}^{d_g+K}} \hat{M}_n(b, \lambda) - \max_{\lambda\in\Lambda(b)} \hat{M}_n(b, \lambda) \right| = 0$$

It thus suffices to show that $b \mapsto \max_{\lambda\in\Lambda(b)} \hat{M}_n(b, \lambda)$ is continuous with probability approaching one. This will follow from the Berge maximum theorem, once it is shown that $\Lambda(\cdot)$ is a continuous correspondence and $\hat{M}_n$ is continuous on $\mathrm{Gr}(\Lambda)$. Since

$$\Lambda(b) \subseteq \lambda(B)^{\eta/2} \equiv \left\{\lambda \,;\, \inf_{\lambda'\in\Lambda(B)} \|\lambda - \lambda'\| \leq \eta/2\right\} = \left\{\lambda \,;\, \inf_{b'\in B} \|\lambda - \lambda(b')\| \leq \eta/2\right\}$$

we can view $\Lambda : B \rightrightarrows \lambda(B)^{\eta/2}$, and thus

$$\mathrm{Gr}(\Lambda) = \left\{(b, \lambda) \in B \times \lambda(B)^{\eta/2} \,;\, \lambda \in \Lambda(b)\right\} = \{(b, \lambda) \,;\, \|\lambda - \lambda(b)\| \leq \eta/2\}\,.$$

1. Consider continuity of the objective first.

   Assumption 7 (iv) implies $h(y, x, b)$ is continuous in b, and assumption 5 (iv) includes that $f^*(\cdot)$ is essentially smooth. It follows that

$$\hat{M}_n(b, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{\lambda^\intercal J(D_i) h(D_i Y_i, X_i, b)}{1 - \hat{p}_{D,n}} - \frac{D_i}{\hat{p}_{D,n}} f^*(\lambda^\intercal h(D_i Y_i, X_i, b))$$

   is continuous at (b, λ) if and only if $\lambda^\intercal h(D_i Y_i, X_i, b) \in (\ell^*, u^*)$ for every i, which holds if and only if $\hat{M}_n(b, \lambda) < \infty$. Notice that
   $\text{Gr}(\lambda)^{\eta/2} \equiv \left\{(b, \lambda) \; ; \; \inf_{(b', \lambda') \in \text{Gr}(\lambda)} \|(b, \lambda) - (b', \lambda')\| \leq \eta/2\right\}$ contains $\text{Gr}(\Lambda)$ because $(b', \lambda') = (b, \lambda(b))$ is an element of $\text{Gr}(\lambda) = \{(b, \lambda(b)) \; ; \; b \in B\}$. Assumption 7 (v) implies $\sup_{(b,\lambda) \in \text{Gr}(\lambda)^{\eta/2}} |M(b, \lambda)|$ is finite, and lemma 2.7.11 shows that $\hat{M}_n$ is uniformly consistent for M on $\text{Gr}(\lambda)^{\eta/2}$, thus the continuous mapping theorem implies $\sup_{(b,\lambda) \in \text{Gr}(\lambda)^{\eta/2}} |\hat{M}_n(b, \lambda)| \xrightarrow{p} \sup_{(b,\lambda) \in \text{Gr}(\lambda)^{\eta/2}} |M(b, \lambda)|$ and therefore
   $\sup_{(b,\lambda) \in \text{Gr}(\lambda)^{\eta/2}} |\hat{M}_n(b, \lambda)|$ is finite with probability approaching one. When it is,

$$\sup_{(b,\lambda) \in \text{Gr}(\Lambda)} \hat{M}_n(b, \lambda) \leq \sup_{(b,\lambda) \in \text{Gr}(\lambda)^{\eta/2}} |\hat{M}_n(b, \lambda)| < \infty.$$

   and $\hat{M}_n$ is continuous on $\text{Gr}(\Lambda)$.

2. Now consider continuity of $\Lambda : B \rightrightarrows \lambda(B)^{\eta/2}$.

   Upper hemicontinuity will follow by application of the Closed Graph Theorem, Aliprantis & Border (2006) theorem 17.11. B is compact by assumption 6 and 2.7.10 shows that $\lambda(\cdot)$ is continuous, therefore $\lambda(B) = \{\lambda(b) \; ; \; b \in B\}$ is compact, and hence $\lambda(B)^{\eta/2}$ is compact. Suppose $\{(b_n, \lambda_n)\}_{n=1}^{\infty} \subseteq \text{Gr}(\Lambda)$ converges to (b, λ). Then $b \in B$ because B is closed. Since $\lambda(\cdot)$ is continuous, $\|\lambda(b_n) - \lambda(b)\| \to 0$, and therefore

$$\|\lambda - \lambda(b)\| \leq \underbrace{\|\lambda - \lambda_n\|}_{\to 0} + \underbrace{\|\lambda_n - \lambda(b_n)\|}_{\leq \eta/2} + \underbrace{\|\lambda(b_n) - \lambda(b)\|}_{\to 0}$$

shows that $\|\lambda - \lambda(b)\| \leq \eta/2$, i.e. $\lambda \in \Lambda(b)$. Thus $(b, \lambda) \in \text{Gr}(\Lambda)$, so $\text{Gr}(\Lambda)$ is closed. Aliprantis & Border (2006) theorem 17.11 then implies $\Lambda : B \rightrightarrows \lambda(B)^{\eta/2}$ is upper hemicontinuous.

Regarding lower semicontinuity, note that $B \subseteq \mathbb{R}^{d_b}$ and $\lambda(B)^{\eta/2} \subseteq \mathbb{R}^{d_g+K}$ are both metric spaces and hence first countable. Thus Aliprantis & Border (2006) theorem 17.21 implies Λ is lower hemicontinuous at $b \in B$ if and only if for any sequence $\{b_n\} \subseteq B$ with $b_n \to b$ and any $\lambda \in \Lambda(b)$, there exists a subsequence $\{b_{n_k}\}_{k=1}^{\infty}$ and elements $\lambda_k \in \Lambda(b_{n_k})$ for each k such that $\lambda_k \to \lambda$. For the subsequence we can take the sequence itself. Notice that $\lambda_n \equiv \lambda(b_n) + \lambda - \lambda(b)$ satisfies

$$\|\lambda_n - \lambda(b_n)\| = \|\lambda(b_n) + \lambda - \lambda(b) - \lambda(b_n)\| = \|\lambda - \lambda(b)\| \leq \eta/2$$

and therefore $\lambda_n \in \lambda(b_n)$. Continuity of $\lambda(\cdot)$ and $b_n \to b$ implies $\lambda_n \to \lambda$, and thus Λ is lower semicontinuous.

To summarize,

$$\sup_{b \in B} \left\| \hat{\lambda}_n(b) - \lambda(b) \right\| < \eta/2 \qquad \text{and} \qquad \sup_{(b,\lambda) \in \text{Gr}(\Lambda)} \hat{M}_n(b, \lambda) < \infty$$

hold with probability approaching one. When both hold, $\hat{\nu}_n(b) = \max_{\lambda \in \Lambda(b)} \hat{M}_n(b, \lambda)$ is continuous by the berge Maximum theorem, implying $\sqrt{n}(\hat{\nu}_n - \nu) \in \mathcal{C}(B)$. Thus

$$P(\sqrt{n}(\hat{\nu}_n - \nu) \in \mathcal{C}(B)) \geq P\left(\sup_{b \in B} \left\| \hat{\lambda}_n(b) - \lambda(b) \right\| < \eta/2 \text{ and } \sup_{(b,\lambda) \in \text{Gr}(\Lambda)} \hat{M}_n(b, \lambda) < \infty \right) \to 1.$$

As argued above, the Portmanteau theorem implies $P(\mathbb{G}_\nu \in \mathcal{C}(B)) = 1$. □

Lemma 2.7.18 ($\sqrt{n}$-consistency and convergence in distribution)**.** *Suppose assumptions 5 and 6 hold, and 7 (i), (iii), (iv), (v) hold, but do not assume* $\boldsymbol{m}(\nu) = \arg\min_{b \in B \cap \boldsymbol{B}_0} \nu(b)$ *is*

unique. Then

$$\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}) \xrightarrow{d} \inf_{b \in \boldsymbol{m}(\nu)} \mathbb{G}_\nu(b)$$

where $\mathbb{G}_\nu$ *is the weak limit of* $\sqrt{n}(\hat{\nu}_n - \nu)$ *in* $\ell^\infty(B)$*.*

Proof. Let $\iota : \ell^\infty(B) \to \mathbb{R}$ be given by $\iota(f) = \inf_{b \in B \cap \mathbf{B}_0} f(b)$. Then

$$\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}) = \sqrt{n}(\iota(\hat{\nu}_n) - \iota(\nu))$$

suggests applying the Delta method, found in Fang & Santos (2019) as theorem 2.1. There are two assumptions to verify:

1. On the map ι:
   (i) ι maps $(\ell^\infty(B), \|\cdot\|_B)$ to $(\mathbb{R}, |\cdot|)$, which are both Banach spaces.
   (ii) Lemma 2.7.6 implies that ι is Hadamard directionally differentiable at any $f \in \mathcal{C}(B)$ tangentially to $\mathcal{C}(B)$, and lemma 2.7.10 shows that $\nu \in \mathcal{C}(B)$.
2. On the estimator $\hat{\nu}_n$:
   (i) As noted in lemma 2.7.17, $\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu$ in $\ell^\infty(B)$.
   (ii) $\mathbb{G}_\nu$ is tight. Lemma 2.7.17 shows that $P(\mathbb{G}_\nu \in \mathcal{C}(B)) = 1$, i.e. the support of $\mathbb{G}_\nu$ is included in $\mathcal{C}(B)$.

Fang & Santos (2019) theorem 2.1 then implies

$$\sqrt{n}(\iota(\hat{\nu}_n) - \iota(\nu)) = \iota'_\nu(\sqrt{n}(\hat{\nu}_n - \nu)) + o_p(1)$$

Lemma 2.7.6 shows the directional derivative of ι at ν is given by

$$\iota'_\nu : \mathcal{C}(B) \to \mathbb{R}, \qquad \iota'_\nu(h) = \inf_{b \in \mathbf{m}(\nu)} h(b)$$

and therefore

$$\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}) = \sqrt{n}(\iota(\hat{\nu}_n) - \iota(\nu)) = \inf_{b \in \mathbf{m}(\nu)} \left\{\sqrt{n}(\hat{\nu}_n(b) - \nu(b))\right\} + o_p(1) \xrightarrow{L} \inf_{b \in \mathbf{m}(\nu)} \mathbb{G}_\nu(b).$$

□

Theorem 2.4.2 (Asymptotic normality)**.** *Suppose assumptions 5, 6, and 7 hold. Let* $\hat{b}_n \equiv \arg\min_{b \in B \cap \boldsymbol{B}_0} \hat{\nu}_n(b)$ *and*

$$\hat{\sigma}_n^2 \equiv \frac{1}{n} \sum_{i=1}^n \left((\hat{\Phi}_n(\hat{b}_n)^{-1})^{(1)} \phi(D, DY, X, \hat{b}_n, \hat{\theta}_n(\hat{b}_n)) \right)^2$$

where $(\hat{\Phi}_n(\hat{b}_n)^{-1})^{(1)}$ *is the first row of the matrix* $\hat{\Phi}_n(\hat{b}_n)^{-1}$*. Then* $\frac{\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})}{\hat{\sigma}_n} \xrightarrow{d} N(0,1)$*.*

Proof. Since $\mathbf{m}(\nu)$ is a singleton, say $\mathbf{m}(\nu) = \{b_\nu\}$, lemmas 2.7.16 and 2.7.18 imply $\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}) \xrightarrow{d} N(0, \sigma^2)$ where

$$\begin{aligned}
\sigma^2 &= E\left[\left((\Phi(b_\nu)^{-1})^{(1)} \phi(D, DY, X, b_\nu, \theta_0(b_\nu)) \right)^2 \right] \\
&= e_1^\intercal \Phi(b_\nu)^{-1} E\left[\phi(D, DY, X, b_\nu, \theta_0(b_0)) \phi(D, DY, X, b_\nu, \theta_0(b_0))^\intercal \right] (\Phi(b_\nu)^{-1})^\intercal e_1
\end{aligned}$$

and $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^{d_g + K + 2}$. Now notice that $\hat{\sigma}_n^2$ is the sample analogue:

$$\begin{aligned}
\hat{\sigma}_n^2 &\equiv \frac{1}{n} \sum_{i=1}^n \left((\hat{\Phi}_n(\hat{b}_n)^{-1})^{(1)} \phi(D, DY, X, \hat{b}_n, \hat{\theta}_n(\hat{b}_n)) \right)^2 \\
&= e_1^\intercal \hat{\Phi}_n(\hat{b}_n)^{-1} \left[\frac{1}{n} \sum_{i=1}^n \phi(D_i, D_i Y_i, X_i, \hat{b}_n, \hat{\theta}_n(\hat{b}_n)) \phi(D_i, D_i Y_i, X_i, \hat{b}_n, \hat{\theta}_n(\hat{b}_n))^\intercal \right] (\hat{\Phi}_n(\hat{b}_n)^{-1})^\intercal e_1
\end{aligned}$$

It suffices to show $\hat{\Phi}_n(\hat{b}_n) \xrightarrow{p} \Phi(b_\nu)$ and

$$\frac{1}{n}\sum_{i=1}^{n} \phi(D_i, D_iY_i, X_i, \hat{b}_n, \hat{\theta}_n(\hat{b}_n))\phi(D_i, D_iY_i, X_i, \hat{b}_n, \hat{\theta}_n(\hat{b}_n))^\intercal$$
$$\xrightarrow{p} E\left[\phi(D, DY, X, b_\nu, \theta_0(b_0))\phi(D, DY, X, b_\nu, \theta_0(b_0))^\intercal\right]. \tag{2.38}$$

With these, the continuous mapping theorem will imply $\hat{\sigma}_n \xrightarrow{p} \sigma$, hence $(\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}), \hat{\sigma}) \xrightarrow{d} (N(0, \sigma^2), \sigma)$, and another application of the continuous mapping theorem gives the conclusion $\frac{\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})}{\hat{\sigma}_n} \xrightarrow{d} N(0, 1)$.

To show $\hat{\Phi}_n(\hat{b}_n) \xrightarrow{p} \Phi(b_\nu)$ and (2.38), first notice that

$$\{\phi(D, DY, X, b, \theta)\phi(D, DY, X, b, \theta)^\intercal \; ; \; (b, \theta) \in \mathrm{Gr}(\theta_0)^\eta\}$$
$$\{\nabla_\theta \phi(D, DY, X, b, \theta) \; ; \; (b, \theta) \in \mathrm{Gr}(\theta_0)^\eta\}$$

are special cases of van der Vaart (2007) example 19.8 and thus Glivenko-Cantelli. Specifically, $\mathrm{Gr}(\theta_0)^\eta$ is closed and bounded and hence compact. $(b, \theta) \mapsto \phi(D, DY, X, b, \theta) \times \phi(D, DY, X, b, \theta)^\intercal$ and $(b, \theta) \mapsto \nabla_\theta \phi(D, DY, X, b, \theta)$ are continuous by inspection of (2.19), (2.21), and (2.22). Finally, $E\left[\sup_{(b,\theta)\in \mathrm{Gr}(\theta_0)^\eta} \|\nabla_\theta \phi(D, DY, X, b, \theta)\|\right] < \infty$ and

$$E\left[\sup_{(b,\theta)\in \mathrm{Gr}(\theta_0)^\eta} \|\phi(D, DY, X, b, \theta)\phi(D, DY, X, b, \theta)^\intercal\|_o\right]$$
$$= E\left[\sup_{(b,\theta)\in \mathrm{Gr}(\theta_0)^\eta} \|\phi(D, DY, X, b, \theta)\|^2\right] < \infty$$

are implied by assumption 7 (v).

Next, observe that $(\hat{b}_n, \hat{\theta}_n(\hat{b}_n)) \xrightarrow{p} (b_\nu, \theta_0(b_\nu))$. First, $\hat{b}_n \xrightarrow{p} b_\nu$ follows from a standard extremum estimator argument. The function $\nu : B \to \mathbb{R}$ is continuous, uniquely minimized over the compact $B \cap \mathbf{B}_0$ at b_ν, and $\sup_{b\in B} |\hat{\nu}_n(b) - \nu(b)| \xrightarrow{p} 0$ by lemma 2.7.12. Thus Newey & McFadden (1994) theorem 2.1 implies $\hat{b}_n = \arg\min_{b\in B\cap \mathbf{B}_0} \hat{\nu}_n(b)$ are consistent for b_ν. Use

the triangle inequality, $\hat{b}_n \xrightarrow{p} b_\nu$, continuity of $\theta_0(b) = (\nu(b), \lambda(b), p_D)$ (lemma 2.7.10), and $\sup_{b\in B}\|\hat{\theta}_n(b) - \theta_0(b)\| = o_p(1)$ (lemma 2.7.12) to see that

$$\|\hat{\theta}_n(\hat{b}_n) - \theta_0(b_\nu)\| \leq \underbrace{\sup_{b\in B}\|\hat{\theta}_n(b) - \theta_0(b)\|}_{=o_p(1)} + \underbrace{\|\theta_0(\hat{b}_n) - \theta_0(b_\nu)\|}_{=o_p(1) \text{ by CMT}} = o_p(1)$$

Note that $(b, \theta) \mapsto E[\nabla_\theta \phi(D, DY, X, b, \theta)]$ is continuous on $\mathrm{Gr}(\theta_0)^\eta$ by the dominated convergence theorem and continuity of $(b, \theta) \mapsto \nabla_\theta \phi(D, DY, X, b, \theta)$ visible in equations (2.21) and (2.22). $(\hat{b}_n, \hat{\theta}_n(\hat{b}_n)) \xrightarrow{p} (b_\nu, \theta_0(b_\nu))$, so $(\hat{b}_n, \hat{\theta}_n(\hat{b}_n)) \in \mathrm{Gr}(\theta_0)^\eta$ holds with probability approaching one and when it does,

$$\begin{aligned}
\|\hat{\Phi}_n(\hat{b}_n) - \Phi(b_\nu)\| &= \left\| \frac{1}{n}\sum_{i=1}^n \nabla_\theta \phi(D_i, D_i Y_i, X_i, \hat{b}_n, \hat{\theta}_n(\hat{b}_n)) - E\left[\nabla_\theta \phi(D, DY, X, b_\nu, \theta_0(b_\nu))\right] \right\| \\
&\leq \left\| \frac{1}{n}\sum_{i=1}^n \nabla_\theta \phi(D_i, D_i Y_i, X_i, \hat{b}_n, \hat{\theta}_n(\hat{b}_n)) - E\left[\nabla_\theta \phi(D, DY, X, \hat{b}_n, \hat{\theta}_n(\hat{b}))\right] \right\| \\
&\quad + \left\| E\left[\nabla_\theta \phi(D, DY, X, \hat{b}_n, \hat{\theta}_n(\hat{b}))\right] - E\left[\nabla_\theta \phi(D, DY, X, b_\nu, \theta_0(b_\nu))\right] \right\| \\
&\leq \underbrace{\sup_{(b,\theta)\in \mathrm{Gr}(\theta_0)^\eta} \left\| \frac{1}{n}\sum_{i=1}^n \nabla_\theta \phi(D_i, D_i Y_i, X_i, b, \theta) - E\left[\nabla_\theta \phi(D, DY, X, b, \theta)\right] \right\|}_{=o_p(1) \text{ by Glivenko-Cantelli}} \\
&\quad + \underbrace{\left\| E\left[\nabla_\theta \phi(D, DY, X, \hat{b}_n, \hat{\theta}_n(\hat{b}))\right] - E\left[\nabla_\theta \phi(D, DY, X, b_\nu, \theta_0(b_\nu))\right] \right\|}_{=o_p(1) \text{ by CMT}} \\
&= o_p(1).
\end{aligned}$$

Essentially the same argument implies (2.38) holds, which completes the proof. □

2.7.6 Appendix: examples

2.7.6.1 Expectation

This simple example is useful primarily to illustrate the ideas in a concrete setting.

Suppose the parameter of interest is $\beta = E[Y] \in \mathbb{R}$, and the sample is $\{D_i, D_i Y_i\}_{i=1}^n$. The conclusion to be supported is that $\beta > \bar{b}$, motivating the null and alternative hypotheses

$$H_0 \ : \ \beta \leq \bar{b}, \qquad\qquad H_1 \ : \ \beta > \bar{b}$$

The model is characterized by $g(y, b) = y - b$. For the dual problem, set $h(y, b) = \begin{pmatrix} y - b & 1 \end{pmatrix}^\intercal$. The dual problem is

$$\sup_{\lambda \in \mathbb{R}^2} \lambda^\intercal c(b) - E_{P_1}[f^*(\lambda^\intercal h(Y, b))] \tag{2.39}$$

where $c(b) = \begin{pmatrix} \frac{-p_D}{1-p_D}(E_{P_1}[Y] - b) & 1 \end{pmatrix}^\intercal$.

Dual solution when d_f is Kullback-Leibler and P_1 is $\mathcal{U}[0, 1]$

Suppose that P_1, the distribution of $Y \mid D = 1$, is $\mathcal{U}[0, 1]$. Let $\mu_1 = E[Y \mid D = 1] = 1/2$. Note that, since the support of P_0 is contained within $[0, 1]$ as well, we have $\beta = E[Y] \in [p_D \mu_1, p_D \mu_1 + (1 - p_D)]$. The endpoints are only attained if P_0 concentrates degenerately at 0 or 1 respectively, distributions which violate $P_0 \ll P_1$.

For tractability, let the measure of selection be Kullback-Leibler. For this divergence we let $f(t) = t \log(t) - t + 1$, which has convex conjugate $f^*(r) = \exp(r) - 1$. The dual problem has first order condition

$$\begin{aligned} 0 &= c(b) - E_{P_1}\left[(f^*)'(\lambda^\intercal h(Y, b)) h(Y, b)\right] \\ &= \begin{pmatrix} \frac{-p_D}{1-p_D}\left(\frac{1}{2} - b\right) \\ 1 \end{pmatrix} - E\left[\exp\left(\lambda_1 (Y - b) + \lambda_2\right) \begin{pmatrix} (Y - b) \\ 1 \end{pmatrix}\right] \end{aligned}$$

From the second equation we have

$$\lambda_2 = -\log\left(E[\exp(\lambda_1 (Y - b))]\right) \tag{2.40}$$

Suppose $b = \frac{1}{2}$. Then the first equation requires

$$0 = E\left[\exp(\lambda_1(Y-b)+\lambda_2)\left(Y-\frac{1}{2}\right)\right] \tag{2.41}$$

Notice that if $\lambda_1 = 0$, then (2.40) implies $\lambda_2 = 0$, and (2.41) holds.

Now suppose $b \neq 1/2$. Consider the dual objective, and notice that

$$E_{P_1}[f^*(\lambda^\intercal h(Y,b))] = \int_0^1 \exp(\lambda^\intercal h(y,b)) - 1 dy$$

Since $b \neq 1/2$, it follows that $\frac{-p_D}{1-p_D}(1/2-b) \neq 0$ and so $\lambda_1 \neq 0$. Thus the integral above can be solved with u-substition, setting $u = \lambda_1(y-b)+\lambda_2$:

$$\begin{aligned} E_{P_1}[f^*(\lambda^\intercal h(Y,b))] &= \int_0^1 \exp(\lambda^\intercal h(y,b)) - 1 dy = \frac{1}{\lambda_1}\int_{\lambda_1(-b)+\lambda_2}^{\lambda_1(1-b)+\lambda_2} \exp(u)du - 1 \\ &= \frac{\exp(\lambda^\intercal \mathbf{b}_1) - \exp(\lambda^\intercal \mathbf{b}_0)}{\lambda^\intercal e_1} - 1 \end{aligned}$$

where $\mathbf{b}_1 = \begin{pmatrix} 1-b & 1 \end{pmatrix}^\intercal$, $\mathbf{b}_0 = \begin{pmatrix} -b & 1 \end{pmatrix}^\intercal$, and $e_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}^\intercal$. Thus (2.39) becomes

$$\sup_{\lambda \in \mathbb{R}^2} \lambda^\intercal \begin{pmatrix} \frac{-p_D}{1-p_D}(1/2-b) \\ 1 \end{pmatrix} - \frac{\exp(\lambda^\intercal \mathbf{b}_1) - \exp(\lambda^\intercal \mathbf{b}_0)}{\lambda^\intercal e_1} + 1$$

from which we can compute the first order conditions

$$0 = \begin{pmatrix} \frac{-p_D}{1-p_D}(1/2-b) \\ 1 \end{pmatrix} - \frac{\exp(\lambda^\intercal \mathbf{b}_1)\mathbf{b}_1 - \exp(\lambda^\intercal \mathbf{b}_0)\mathbf{b}_0}{\lambda^\intercal e_1} + \frac{\exp(\lambda^\intercal \mathbf{b}_1) - \exp(\lambda^\intercal \mathbf{b}_0)}{(\lambda^\intercal e_1)^2} e_1$$

Once again, the second equation can be solved for λ_2. The following form will be more

useful:

$$0 = 1 - \frac{\exp(\lambda_1(1-b)+\lambda_2) - \exp(\lambda_1(-b)+\lambda_2)}{\lambda_1}$$
$$\implies \frac{\lambda_1}{\exp(\lambda_2)} = \exp(\lambda_1(1-b)) - \exp(\lambda_1(-b)) \tag{2.42}$$

The first equation is

$$\begin{aligned}
\frac{-p_D}{1-p_D}\left(\frac{1}{2}-b\right) &= \frac{\exp(\lambda_1(1-b)+\lambda_2)(1-b) - \exp(\lambda_1(-b)+\lambda_2)(-b)}{\lambda_1} \\
&\quad - \frac{\exp(\lambda_1(1-b)+\lambda_2) - \exp(\lambda_1(-b)+\lambda_2)}{\lambda_1^2} \\
&= \frac{\exp(\lambda_2)}{\lambda_1}\Bigg[\exp(\lambda_1(1-b)) - b[\exp(\lambda_1(1-b)) - \exp(\lambda_1(-b))] \\
&\quad - \frac{\exp(\lambda_1(1-b)) - \exp(\lambda_1(-b))}{\lambda_1}\Bigg] \\
&= \frac{\exp(\lambda_1(1-b))}{\exp(\lambda_1(1-b)) - \exp(\lambda_1(-b))} - b - \frac{1}{\lambda_1} = \frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - b - \frac{1}{\lambda_1}
\end{aligned}$$

where the second to last equality uses (2.42) above. Rearranging gives

$$\frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1} = \frac{-p_D(1/2-b) + (1-p_D)b}{1-p_D} = \frac{2b-p_D}{2(1-p_D)}$$

Now notice that $\frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1}$ is well defined and continuous whenever $\lambda_1 \neq 0$, takes values between 0 and 1, with limits

$$\lim_{\lambda_1 \to \infty} \frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1} = 1, \qquad \lim_{\lambda_1 \to -\infty} \frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1} = 0$$

Repeated applications of l'Hôpital's rule shows that

$$\lim_{\lambda_1 \to 0} \frac{\exp(\lambda_1)}{\exp(\lambda_1)-1} - \frac{1}{\lambda_1} = \lim_{\lambda_1 \to 0} \frac{\lambda_1 \exp(\lambda_1) - \exp(\lambda_1) + 1}{\lambda_1(\exp(\lambda_1)-1)} = \frac{1}{2}$$

Therefore there exists a solution whenever $\frac{2b-p_D}{2(1-p_D)} \in \left(0, \frac{1}{2}\right) \cup \left(\frac{1}{2}, 1\right)$. Given this solution, (2.42) can be rearranged to obtain

$$\lambda_2 = \log\left(\frac{\lambda_1}{\exp(\lambda_1(1-b)) - \exp(\lambda_1(-b))}\right)$$

Now notice that

$$\begin{aligned}\frac{2b-p_D}{2(1-p_D)} > 0 &\implies b > \frac{p_D}{2},\\ \frac{2b-p_D}{2(1-p_D)} < 1 &\implies b < 1 - \frac{p_D}{2}\end{aligned}$$

and recall that $b = 1/2$ implies $\lambda_1 = \lambda_2 = 0$ solves the dual problem. Therefore the dual problem has a solution whenever $b \in \left(\frac{p_D}{2}, 1 - \frac{p_D}{2}\right)$.

P_1 has compact support, and $f^*(\lambda_1(y-b)+\lambda_2) = \exp(\lambda_1(y-b)+\lambda_2) - 1$ is continuous in y for any (λ_1, λ_2). Thus the extreme value theorem implies the solution is in the interior of $\{\lambda \in \mathbb{R}^2 \,;\, E[|f^*(\lambda^\intercal h(Y,b))|] < \infty\} = \left\{\lambda \in \mathbb{R}^2 \,;\, \int |\exp(\lambda_1(y-b)+\lambda_2) - 1| dy < \infty\right\}$. The implied solution to the primal, $q^b(y) = (f^*)'(\lambda^\intercal h(y,b)) = \exp(\lambda_1(y-b)+\lambda_2)$ satisfies $0 < q^b(y) < \infty$ on the support of P_1 and solves the moment conditions. Thus assumption 6 is satisfied for any convex, compact $B \subset \left(\frac{p_D}{2}, 1 - \frac{p_D}{2}\right)$.

2.7.6.2 Linear models

Lemma 2.4.1 (Convex value function, linear models)**.** *Suppose assumptions 5 and 6 hold, the sample is* $\{D_i, D_i Y_i, X_{i1}, X_{i2}\}_{i=1}^n$ *where* $Y_i \in \mathbb{R}$, $X_{i1} \in \mathbb{R}^{d_{x1}}$, *and* $X_{i2} \in \mathbb{R}^{d_{x2}}$, *and the parameter* β *is identified by*

$$E[(Y - X_1^\intercal \beta) X_2] = 0$$

Then $\hat{\nu}_n$ *and* ν *are convex. If in addition* $E[X_2 X_1^\intercal]$ *has full column rank, then* ν *is strictly convex.*

Proof. Let $b^0, b^1 \in B$, be distinct, $\alpha \in (0,1)$, and $b^\alpha = \alpha b^1 + (1-\alpha)b^0$. The proof of theorem (2.3.1) shows that the primal problem at b^0 and b^1 is attained by Q^0 and Q^1 with densities q^0, q^1. The moment conditions are $0 = E\left[X_2(Y - X_1^\intercal \beta)\right] = E\left[X_2 Y\right] - E\left[X_2 X_1^\intercal\right]\beta$, so $Q^0 \in \mathbf{P}^{b^0}$ and $Q^\in \mathbf{P}^{b^1}$ implies

$$E_{Q^1}\left[X_2 Y\right] - E_{P_{0X}}\left[X_2 X_1^\intercal\right] b^1 = \frac{-p_D}{1-p_D}\left(E_{P_1}\left[X_2 Y\right] - E_{P_1}\left[X_2 X_1^\intercal\right] b^1\right), \tag{2.43}$$

$$E_{Q^0}\left[X_2 Y\right] - E_{P_{0X}}\left[X_2 X_1^\intercal\right] b^0 = \frac{-p_D}{1-p_D}\left(E_{P_1}\left[X_2 Y\right] - E_{P_1}\left[X_2 X_1^\intercal\right] b^0\right) \tag{2.44}$$

implying that

$$E_{\alpha Q^1 + (1-\alpha)Q^0}\left[X_2 Y\right] - E_{P_{0X}}\left[X_2 X_1^\intercal\right] b^\alpha = \frac{-p_D}{1-p_D}\left(E_{P_1}\left[X_2 Y\right] - E_{P_1}\left[X_2 X_1^\intercal\right] b^\alpha\right)$$

Similarly, $E_{Q^0}[\mathbb{1}\{X = x_k\}] = E_{Q^1}[\mathbb{1}\{X = x_k\}] = E_{P_{0X}}[\mathbb{1}\{X = x_k\}]$ for all $k = 1, \ldots, K$. It follows that $Q^\alpha \equiv \alpha Q^1 + (1-\alpha)Q^0$ is feasible for $b^\alpha = \alpha b^1 + (1-\alpha)b^0$. This implies

$$d_f(Q^\alpha \| P_1) \geq \inf_{Q \in \mathbf{P}^{b^\alpha}} d_f(Q \| P_1) = \nu(b^\alpha)$$

Q^α has P_1-density $q^\alpha = \alpha q^1 + (1-\alpha)q^0$. Convexity of f implies that for any (y, x),

$$\alpha f(q^1(y,x)) + (1-\alpha) f(q^0(y,x)) \geq f(\alpha q^1(y,x) + (1-\alpha)q^0(y,x)) = f(q^\alpha(y,x))$$

integrating with respect to P_1 shows that

$$\alpha d_f(Q^1 \| P_1) + (1-\alpha) d_f(Q^0 \| P_1) \geq d_f(Q^\alpha \| P_1) \geq \nu(b^\alpha)$$

Since the left hand side equals $\alpha\nu(b^1) + (1-\alpha)\nu(b^0)$, this shows ν is convex. Notice that no properties of P_1, P_{0X} were specified in the argument above, so the same argument works to show $\hat{\nu}_n(b)$ is convex in b by replacing P_1, P_{0X} with their empirical counterparts.

Finally, to see that ν is strictly convex when $E[X_2 X_1^\intercal]$ has full column rank, use equations

(2.43) and (2.44) to see that

$$(1-p_D)\left[E_{Q^{1,n}}\left[X_2Y\right]-E_{Q^{0,n}}\left[X_2Y\right]\right]=\underbrace{\left[p_DE_{P_1}[X_2X_1^\intercal]+(1-p_D)E_{P_0}[X_2X_1^\intercal]\right]}_{=E[X_2X_1^\intercal]}(b^1-b_0)$$

Since $E[X_2X_1^\intercal]$ has full column rank and $b^1-b^0\neq 0$,

$$(1-p_D)\left[E_{Q^1}\left[X_2Y\right]-E_{Q^0}\left[X_2Y\right]\right]\neq 0$$

and thus Q^1 differs from Q^0, implying q^1 differs from q^0 on a set of positive P_1 measure. For (y,x) in that set, strict convexity of f assumed in (5) (iv) implies

$$\alpha f(q^1(y,x))+(1-\alpha)f(q^0(y,x))>f(\alpha q^1(y,x)+(1-\alpha)q^0(y,x))=f(q^\alpha(y,x))$$

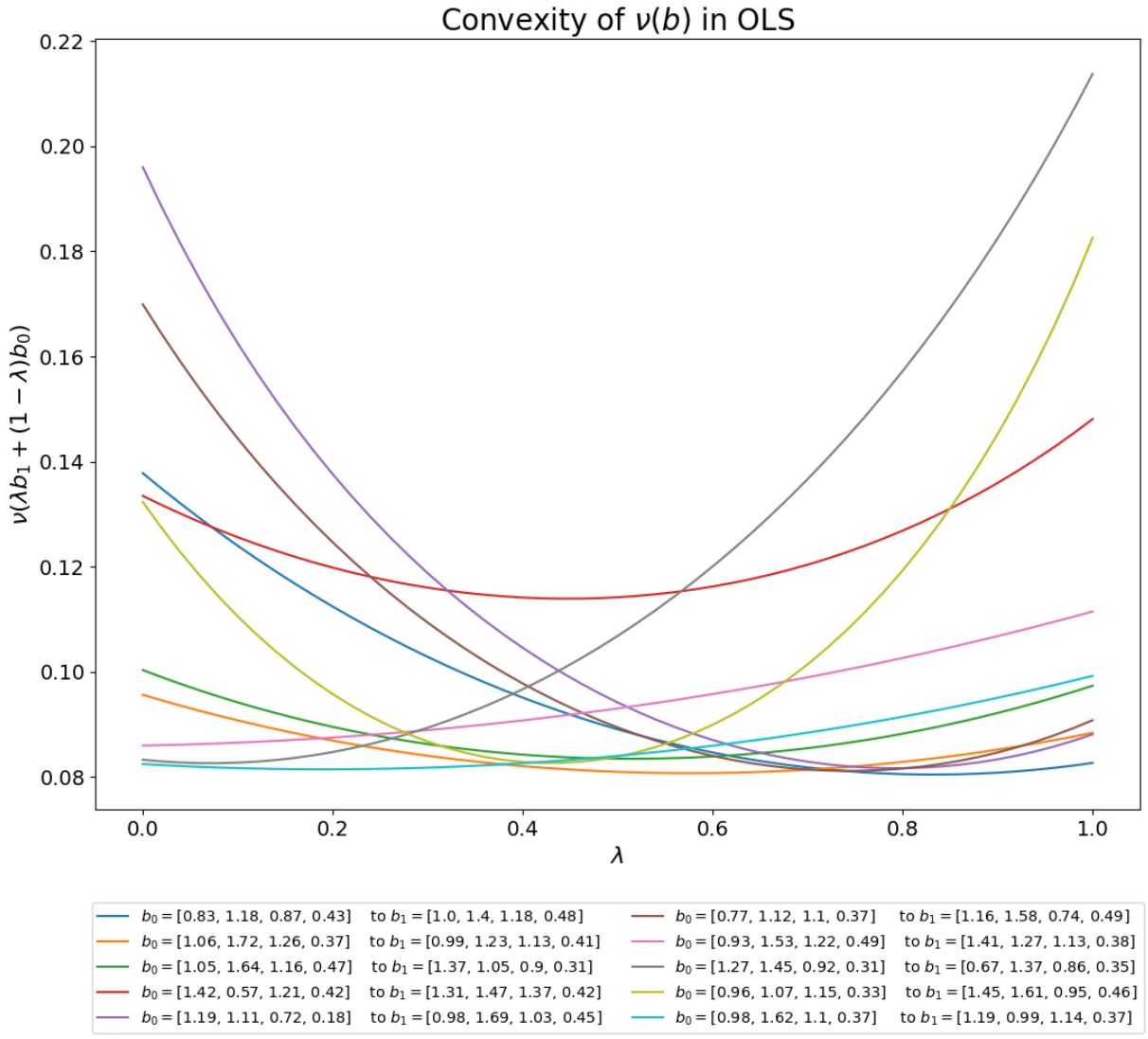
integrating with respect to P_1 implies $\alpha d_f(Q^1\|P_1)+(1-\alpha)d_f(Q^0\|P_1)>d_f(Q^\alpha\|P_1)$, and thus $\alpha\nu(b^1)+(1-\alpha)\nu(b^0)>d_f(Q^\alpha\|P_1)\geq\nu(b^\alpha)$. □

Simulations suggest that OLS more generally produces convex $\nu(b)$. Consider the data generating process described in section 2.5.2. Here the data is of the form $\{D_i,D_iY_{i1},D_iY_{i2},X_{i1},X_{i2}\}_{i=1}^n$, and the model is given by

$$Y_{i1}=\beta_0+\beta_1X_1+\beta_2Y_2+\beta_3X_2+\varepsilon,\qquad\qquad E\left[\begin{pmatrix}1\\X_1\\Y_2\\X_2\end{pmatrix}\varepsilon\right]=0$$

The following figure investigates convexity of the $\nu(b)$ (where $d_f(Q\|P)=H^2(Q,P)$) numerically, by looking for convexity along random line segments. Specifically, let b_1 and b_0 be points in the sample space and compute $\hat{\nu}_n(\lambda b_1+(1-\lambda)b_0)$ for many values of λ between 0 and 1. The following figure shows the results of this exercise for 10 randomly selected (b_0,b_1)

pairs, and shows that no deviation from convexity was detected.



2.7.6.3 Binary choice models

Let $V \in \{0, 1\}$, $W \in \mathbb{R}^d$, and suppose interest is in $P(V = 1 \mid W = w)$. A common choice of model assumes

$$P(V = 1 \mid W = w) = F(w^\intercal \beta)$$

for a known CDF $F(\cdot)$. This model can be derived from a latent variable model of the form $V = \mathbb{1}\{W'\beta \geq \xi\}$, where conditional on W, the unobserved "latent variable" ξ has distribution $F(x)$.

$$P(V = 1 \mid W = w) = P(\xi \leq W^\intercal \beta \mid W = w) = F(w^\intercal \beta)$$

For example, the logistic regression uses $F(x) = \Lambda(x) = \frac{\exp(x)}{1+\exp(x)}$, while the probit model uses $F(x) = \Phi(x) = \int_\infty^x \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)dt$.

Given i.i.d. data of the form $\{V_i, W_i\}_{i=1}^n$, the model can be estimated through maximum likelihood. The likelihood of an observation (V, W) is $F(W^\intercal b)^V (1 - F(W^\intercal b)^{1-V}$, implying a population log-likelihood of

$$\ell(b) \equiv E\left[V \ln(F(W^\intercal b)) + (1 - V) \ln(1 - F(W^\intercal b))\right]$$

Assuming $F(x)$ is differentiable with density $f(x)$ and that differentiation and expectation can be interchanged, the score is given by

$$s(b) \equiv \nabla_b \ell(b) = E\left[\frac{f(W^\intercal b)}{F(W^\intercal b)\left(1 - F(W^\intercal b)\right)} \left(V - F(W^\intercal b)\right) W\right]$$

and supposing $f(x)$ is differentiable with derivative $f'(x)$, the Hessian can be calculated and shown negative definite when $E[WW^\intercal]$ is full rank. This implies the log-likelihood is strictly concave, and hence the first order condition suffices for maximization. Therefore the model could also be viewed as a GMM model solving

$$0 = E\left[\frac{f(W^\intercal \beta)}{F(W^\intercal \beta)\left(1 - F(W^\intercal \beta)\right)} \left(V - F(W^\intercal \beta)\right) W\right]$$

Logistic model

For the logistic model, $F(x) = \Lambda(x) = \frac{\exp(x)}{1+\exp(x)}$, we can compute that

$$f(x) = \frac{\exp(x)}{(1+\exp(x))^2} = F(x)(1-F(x))$$

and thus the score simplifies to

$$s(b) = E\left[(V - \Lambda(W^\intercal b))\, W\right]$$

This makes it straightforward to compute the Hessian of the log-likelihood as

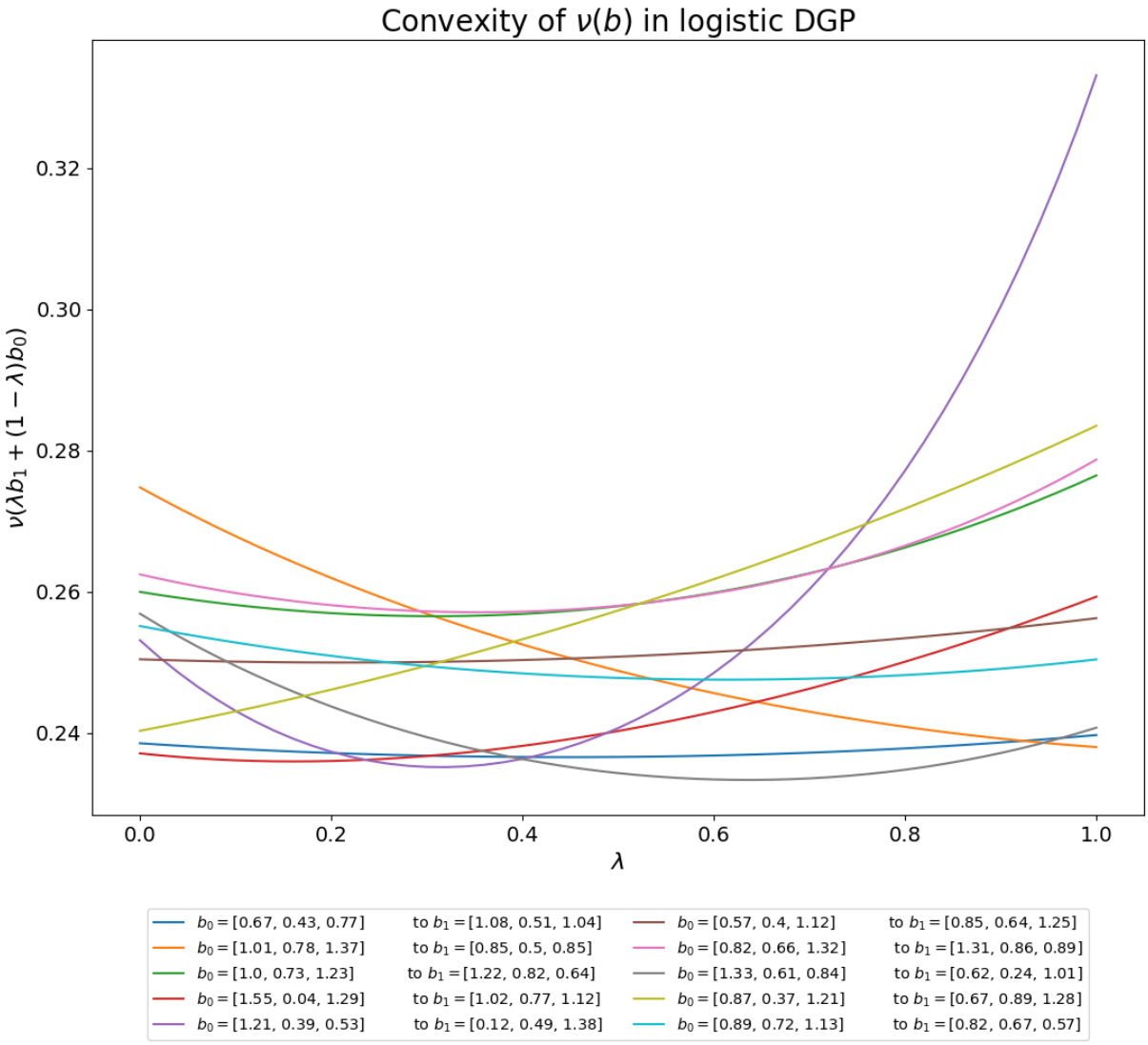
$$\nabla_b^2 \ell(b) = E\left[-\Lambda(W^\intercal b)(1-\Lambda(W^\intercal b))WW^\intercal\right]$$

Let $U \equiv \sqrt{\Lambda(W^\intercal b)(1-\Lambda(W^\intercal b))}W$ and observe that $\nabla_b^2 \ell(b) = -E[UU^\intercal]$ is negative definite if $E[WW^\intercal]$ is full rank. Thus, the logistic model can be viewed as a GMM model, where β solves

$$0 = E\left[(V - \Lambda(W^\intercal \beta))\, W\right]$$

This model can be put into the form used in assumption 5 with $Z = (Z_{(1)}, Z_{-1}) = (V, W)$, $g(z,b) = (z_1 - \Lambda(z_{-1}^\intercal b))z_{-1}$, and $\nabla_b g(z,b) = -\Lambda(z_{-1}^\intercal b)(1-\Lambda(z_{-1}^\intercal b))z_{-1}z_{-1}^\intercal$.

Simulations suggest that the logistic model may also produce a convex $\nu(b)$. Consider the data generating process described in section 2.5.3. The logistic model can also be investigated for convexity. The same numerical exercise described above results in a figure that again shows no deviation from convexity.


Convexity of $\nu(b)$ in logistic DGP
$\nu(\lambda b_1 + (1-\lambda)b_0)$
0.32
0.30
0.28
0.26
0.24
0.0
0.2
0.4
0.6
0.8
1.0
λ
$b_0 = [0.67, 0.43, 0.77]$ to $b_1 = [1.08, 0.51, 1.04]$
$b_0 = [1.01, 0.78, 1.37]$ to $b_1 = [0.85, 0.5, 0.85]$
$b_0 = [1.0, 0.73, 1.23]$ to $b_1 = [1.22, 0.82, 0.64]$
$b_0 = [1.55, 0.04, 1.29]$ to $b_1 = [1.02, 0.77, 1.12]$
$b_0 = [1.21, 0.39, 0.53]$ to $b_1 = [0.12, 0.49, 1.38]$
$b_0 = [0.57, 0.4, 1.12]$ to $b_1 = [0.85, 0.64, 1.25]$
$b_0 = [0.82, 0.66, 1.32]$ to $b_1 = [1.31, 0.86, 0.89]$
$b_0 = [1.33, 0.61, 0.84]$ to $b_1 = [0.62, 0.24, 1.01]$
$b_0 = [0.87, 0.37, 1.21]$ to $b_1 = [0.67, 0.89, 1.28]$
$b_0 = [0.89, 0.72, 1.13]$ to $b_1 = [0.82, 0.67, 0.57]$

Bibliography

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of econometrics*, *113*(2), 231–263.

Aliprantis, C. D., & Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629–676.

Antoine, B., & Dovonon, P. (2021). Robust estimation with exponentially tilted hellinger distance. *Journal of Econometrics*, *224*(2), 330–344.

Bhatia, R. (1997). *Matrix Analysis*. New York, NY: Springer New York.

Bollinger, C. R., Hirsch, B. T., Hokayem, C. M., & Ziliak, J. P. (2019). Trouble in the tails? what we know about earnings nonresponse 30 years after lillard, smith, and welch. *Journal of Political Economy*, *127*(5), 2143–2185.

Borwein, J. M., & Lewis, A. S. (1991). Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, *29*(2), 325–338.

Borwein, J. M., & Lewis, A. S. (1993). Partially-finite programming in l_1 and the existence of maximum entropy estimates. *SIAM Journal on Optimization*, *3*(2), 248–267.

Brame, R., Turner, M. G., Paternoster, R., & Bushway, S. D. (2012). Cumulative prevalence of arrest from ages 8 to 23 in a national sample. *Pediatrics*, *129*(1), 21–27.

Broniatowski, M., & Keziou, A. (2006). Minimization of φ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, *43*(4), 403–442.

Broniatowski, M., & Keziou, A. (2012). Divergences and duality for estimation and test under moment condition models. *Journal of Statistical Planning and Inference*, *142*(9), 2554–2573.

Callaway, B. (2021). Bounds on distributional treatment effect parameters using panel data with an application on job displacement. *Journal of Econometrics*, *222*(2), 861–881.

Chernozhukov, V., Lee, S., & Rosen, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica*, *81*(2), 667–737.

Coleman, R. (2012). *Calculus on normed vector spaces*. Springer Science & Business Media.

Couch, K. (1992). Long-term effects of the national supported work experiment, and parametric and nonparametric tests of model specification and the estimation of treatment effects. *Unpublished Ph. D. dissertation, University of Wisconsin-Madison. 1992b." New Evidence on the Long-Term Effects of Employment Training Programs." Journal of Labor Economics*, *10*(4), 380–88.

Csiszár, I., Gamgoa, F., & Gassiat, E. (1999). Mem pixel correlated solutions for generalized moment and interpolation problems. *IEEE Transactions on Information Theory*, *45*(7), 2253–2270.

Das, M., Newey, W. K., & Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, *70*(1), 33–58.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, *95*(3), 932–945.

Diegert, P., Masten, M. A., & Poirier, A. (2022). Assessing omitted variable bias when the controls are endogenous. *arXiv preprint arXiv:2206.02303*.

Dunipace, E. (2021). Optimal transport weights for causal inference. *arXiv preprint arXiv:2109.01991*.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife," the annals of statistics, 7, 1–26. freedman, da (1981). *Bootstrapping Regression Models," The Annals of Statistics*, *9*, 1218–1228.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Ehm, W., Gneiting, T., Jordan, A., & Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *78*(3), 505–562.

Fan, Y., & Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, *26*(3), 931–951.

Fan, Y., & Park, S. S. (2012). Confidence intervals for the quantile of treatment effects in randomized experiments. *Journal of Econometrics*, *167*(2), 330–344.

Fan, Y., Shi, X., & Tao, J. (2023). Partial identification and inference in moment models with incomplete data. *Journal of Econometrics*, *235*(2), 418–443.

Fang, Z., & Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, *86*(1), 377–412.

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

Firpo, S., & Ridder, G. (2019). Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, *213*(1), 210–234.

Frandsen, B. R., & Lefgren, L. J. (2021). Partial identification of the distribution of treatment effects with an application to the knowledge is power program (kipp). *Quantitative Economics*, *12*(1), 143–171.

Friebel, G., Heinz, M., Hoffman, M., & Zubanov, N. (2023). What do employee referral programs do? measuring the direct and overall effects of a management practice. *Journal of Political Economy*, *131*(3), 633–686.

Galichon, A. (2017). A survey of some recent applications of optimal transport methods to econometrics. *The Econometrics Journal*, *20*(2), C1–C11.

Gunsilius, F., & Xu, Y. (2021). Matching for causal effects via multimarginal unbalanced optimal transport. *arXiv preprint arXiv:2112.04398*.

Hahn, J., Kuersteiner, G., & Newey, W. (2002). Higher order properties of bootstrap and jackknife bias corrected maximum likelihood estimators. *Unpublished manuscript*.

Haile, P. A., & Tamer, E. (2003). Inference with an incomplete model of english auctions. *Journal of Political Economy*, *111*(1), 1–51.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, (pp. 153–161).

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.

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

Horowitz, J. L. (2001). The bootstrap. In *Handbook of econometrics*, vol. 5, (pp. 3159–3228). Elsevier.

Horowitz, J. L., & Manski, C. F. (1995). Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, (pp. 281–302).

Horowitz, J. L., & Manski, C. F. (2006). Identification and estimation of statistical functionals using incomplete data. *Journal of Econometrics*, *132*(2), 445–459.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, *62*(2), 467–475.

Ji, W., Lei, L., & Spector, A. (2023). Model-agnostic covariate-assisted inference on partially identified causal effects. *arXiv preprint arXiv:2310.08115*.

Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, *83*(5), 2043–2063.

Kline, P., & Santos, A. (2013). Sensitivity to missing data assumptions: Theory and an evaluation of the us wage structure. *Quantitative Economics*, *4*(2), 231–267.

Kreider, B., & Pepper, J. V. (2007). Disability and employment: Reevaluating the evidence in light of reporting errors. *Journal of the American Statistical Association*, *102*(478), 432–441.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, (pp. 604–620).

Manski, C. F. (1997). Monotone treatment response. *Econometrica: Journal of the Econometric Society*, (pp. 1311–1334).

Manski, C. F. (2005). Partial identification with missing data: concepts and findings. *International Journal of Approximate Reasoning*, *39*(2-3), 151–165.

Manski, C. F. (2013). Response to the review of 'public policy in an uncertain world'.

Masten, M. A., & Poirier, A. (2020). Inference on breakdown frontiers. *Quantitative Economics*, *11*(1), 41–111.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Pollard, D. (2002). *A user's guide to measure theoretic probability*. 8. Cambridge University Press.

Rockafellar, R. T. (1970). *Convex analysis*, vol. 18. Princeton university press.

Rubin, D. B. (1981). The bayesian bootstrap. *The annals of statistics*, (pp. 130–134).

Russell, T. M. (2021). Sharp bounds on functionals of the joint distribution in the analysis of treatment effects. *Journal of Business & Economic Statistics*, *39*(2), 532–546.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, *55*(58-63), 94.

Staudt, T., Hundrieser, S., & Munk, A. (2022). On the uniqueness of kantorovich potentials. *arXiv preprint arXiv:2201.08316*.

Torous, W., Gunsilius, F., & Rigollet, P. (2021). An optimal transport approach to causal inference. *arXiv preprint arXiv:2108.05858*.

van der Vaart, A., & Wellner, J. A. (1997). *Weak convergence and empirical processes with applications to statistics*. London: Royal Statistical Society, 1988-.

van der Vaart, A. W. (2007). *Asymptotic statistics*. Cambridge university press.

Villani, C. (2003). *Topics in optimal transportation*, vol. 58. American Mathematical Soc.

Villani, C. (2009). *Optimal transport: old and new*, vol. 338. Springer.

Zeidler, E. (1986). Nonlinear functional analysis vol. 1: Fixed-point theorems.