arXiv:submit/6296443 [econ.EM] 20 Mar 2025

Partial Identification in Moment Models with Incomplete Data via Optimal Transport*

Yanqin Fan,† Brendan Pass,‡ and Xuetao Shi§

March 20, 2025

Abstract

In this paper, we develop a unified approach to study partial identification of a finite-dimensional parameter defined by a moment equality model with incomplete data. We establish a novel characterization of the identified set for the true parameter in terms of a continuum of inequalities defined by optimal transport costs. For a special class of moment functions, we show that the identified set is convex, and its support function can be easily computed by solving an optimal transport problem. We demonstrate the generality and effectiveness of our approach through several running examples, including the linear projection model and two algorithmic fairness measures.

Keywords: Algorithmic Fairness, Causal Inference, Convex Analysis, Data Combination, Linear Projection, Partial Optimal Transport, Support Function, Time Complexity

JEL codes: C21, C61, C63

*We thank Stephane Bonhomme, Ruixuan Liu, and participants of the BIRS Workshop on Optimal Transport and Distributional Robustness in March 2024, seminar participants in the statistics department at the University of Washington and Amazon in October 2024 for useful feedback. Brendan Pass is pleased to acknowledge the support of Natural Sciences and Engineering Research Council of Canada Discovery Grant numbers 04658-2018 and 04864-2024.

†Department of Economics, University of Washington

‡Department of Mathematical and Statistical Sciences, University of Alberta

§School of Economics, University of Sydney

1 Introduction

Let $Y_1 \in \mathcal{Y}_1, Y_0 \in \mathcal{Y}_0$, and $X \in \mathcal{X}$ denote three random vectors, where $\mathcal{Y}_1 \subset \mathbb{R}^{d_1}$, $\mathcal{Y}_0 \subset \mathbb{R}^{d_0}$, and $\mathcal{X} \subset \mathbb{R}^{d_x}$ are the supports. Consider an unconditional moment model,

$$\mathbb{E}_o \left[m \left(Y_1, Y_0, X; \theta^* \right) \right] = \mathbf{0}, \tag{1}$$

where $\theta^* \in \Theta$ is the true unknown parameter, $\Theta \subset \mathbb{R}^{d_\theta}$ is the parameter space, and m is a known k dimensional vector-valued moment function. We reserve the notation $\mathbb{E}_o$ to denote the expectation taken with respect to the true but unknown joint distribution μ_o of (Y_1, Y_0, X). Many parameters of interest in applied research can be expressed as θ^* itself or its known functions, where θ^* satisfies model (1) for an appropriate choice of the moment function m. Identification, estimation, and inference on θ^* are well studied when *complete data* information on (Y_1, Y_0, X) is available to the researcher, namely, when (Y_1, Y_0, X) is observed in one dataset. See e.g., the seminal paper of Hansen (1982) on Generalized Method of Moment (GMM) and partial identification analysis of Andrews and Soares (2010). However, in many applications, complete data on (Y_1, Y_0, X) are unavailable rendering existing methods such as those in Hansen (1982) and Andrews and Soares (2010) inapplicable.

This paper considers a prevalent case of *incomplete data* in empirical research, where the relevant information on (Y_1, Y_0, X) is contained in two datasets[1] and the units of observations in different datasets cannot be matched: the first dataset contains a random sample on (Y_1, X) and the second dataset contains a random sample on (Y_0, X). Examples of such datasets include long-run returns to college attendance (PSID/NLSY+Addhealth, Fan et al. (2014, 2016)); algorithmic fairness (decision outcome and protected attribute are observed in separate datasets, Kallus et al. (2022)); long-term treatment effects (Experimental+observational data, Athey et al. (2019)); and repeated cross section data in difference-in-differences set-up (Fan and Manzanares (2017)). The sample information thus allows the identification and estimation of the distribution functions of (Y_1, X) and (Y_0, X), but does not identify the distribution of (Y_1, Y_0, X) without additional assumptions on the joint distribution of (Y_1, Y_0, X) or additional structure on the moment function. When the moment function is additively separable in Y_1 and Y_0, point identification of θ^* is possible

[1]The methods developed in this paper could in principle be extended to the case of more than two datasets. We leave extensions for future work.

under weak conditions; see e.g., Hahn (1998), Chen et al. (2008), and Graham et al. (2016), among others. However, when the parameter of interest depends on the joint distribution of Y_1 and Y_0, the moment function m will typically not be additively separable. In the special case that both Y_1 and Y_0 are *scalar random variables*, Fan et al. (2023b) study partial identification and inference in model (1) via the copula approach.[2] Their method fails to apply when at least one of Y_1 or Y_0 is multivariate. This motivates the current paper, where we allow Y_1 and Y_0 to be of *any finite dimensions*, making our framework and methodology the most general available in the current literature.

The first and main contribution of this paper is to characterize the identified set for θ^* (and known functions of θ^*) in the moment model (1) via optimal transport costs. Our characterization allows for the general moment function of any finite number of components and for Y_1 and Y_0 of any finite dimensions. We demonstrate via several running examples the flexibility and broad applicability of our methodology in diverse applications. Example 1 is the Linear Projection (LP) model in Pacini (2019); Example 2 is the demographic disparity (DD) in Kallus et al. (2022) with any finite number of protected classes; Example 3 is the true-positive rate disparity (TPRD) in Kallus et al. (2022) with any finite number of protected classes; and in Example 4, the parameter of interest is an extension of that in Fan et al. (2017) to any finite dimension and Y_1 and Y_0 of any finite dimensions.

As the second contribution of this paper, we apply our general characterization of the identified set for θ^* to an important class of moment models, where the moment function $m(\cdot;\theta)$ can be written as an affine transformation of θ with the expectation of the linear transformation matrix being identifiable but not the translation vector. We show that the identified set for θ^* is characterized by a continuum of linear inequalities and is thus convex. Furthermore, we derive the expression of the support function of the identified set and show that the value of the support function in each direction can be obtained by solving an optimal transport problem, which is easy to compute.

Third, through our running examples, we show that one can often choose the moment function m and θ^* in such a way that m is an affine transformation and the parameter of interest is a known function of θ^*. We demonstrate through Example 3

[2] Fan et al. (2023b) include a point-identified nuisance parameter γ^* in their moment function to separate the parameter of interest θ^* from γ^*. Model (1) can be accommodated to having the nuisance parameter by appending γ^* to θ^*.

that this formulation is advantageous over directly modeling the parameter of interest into the moment function because the identified set for the parameter of interest may not be convex, while the identified set for θ^* is.

Our fourth contribution is to provide a detailed analysis of each running example. For Example 1, we make use of the monotone rearrangement inequality to provide a closed-form expression of the support function of the identified set for multidimensional Y_0, extending the result in Pacini (2019) for one dimensional Y_0. We provide a numerical comparison of our identified set with the outer set in Pacini (2019) for multivariate Y_0. The result confirms that our identified set is always a subset of the outer set in Pacini (2019). We further demonstrate the importance of obtaining the identified set instead of the outer set through cases where our identified set identifies the sign of the true parameter while the outer set does not.

For Example 2 on the DD measure, again by making use of the monotone rearrangement inequality, we provide a *closed-form expression* for the support function of the identified set for *any* collection of DD measures and *any* number of protected classes. For the specific collection of DD measures studied in Kallus et al. (2022), our closed-form expression is the solution to the infinite dimensional linear programming formulation in Kallus et al. (2022). We derive the time complexity of our procedure that uses the closed-form expression of the support function to compute the identified set and demonstrate that it is much faster than solving the linear programming problem in Kallus et al. (2022). Furthermore, we show that the identified set for any single DD measure is a closed interval regardless of *the number of protected classes* and provide closed-form expressions for the interval endpoints. This extends the result in Kallus et al. (2022) established for *two protected classes*.

For Example 3, the identified set for the TPRD measures is non-convex when the number of protected classes is larger than two. As a result, Kallus et al. (2022) recommend constructing the convex hull of the identified set instead of the identified set itself and characterizes the support function of the convex hull through a complicated, computationally costly, non-convex optimization problem. Choosing the parameter vector θ^* cleverly, we express the identified set for any finite number of TPRD measures as a continuous map of the identified set for θ^*. We show that the identified set for θ^* is convex. As a continuous transformation of the convex identified set for θ^*, the identified set for the TPRD measures is connected. As a result, the identified set for any single TPRD measure is a closed interval regardless of *the number*

of protected classes. Furthermore, we provide closed-form expressions for the lower and upper bounds of the closed interval extending the result in Kallus et al. (2022) for *two protected classes*. However, for multiple TPRD measures, the identified set is more challenging to evaluate than that of multiple DD measures. In particular, the monotone rearrangement inequality is not directly applicable. Instead, we develop a simple algorithm for computing the support function of the identified set for θ^* and for any collection of TPRD measures. We call our Algorithm 1 Dual Rank Equilibration AlgorithM (DREAM). DREAM builds on a special feature of the solution to the *partial optimal transport* problem that defines the support function of the identified set for θ^*. We further analyze the time complexity of DREAM, which is much faster than solving the NP-hard non-convex optimization proposed by Kallus et al. (2022).

The rest of this paper is organized as follows. Section 2 introduces our general class of models and presents four motivating examples. In Section 3, we first characterize the identified set for θ^* in terms of a continuum of moment inequalities and then study the properties of the identified set for a special class of moment models. Sections 4 to 6 apply our general results to the motivating examples. The last section offers some concluding remarks. Technical proofs are relegated to Appendix A.

Notation We let $\mathbf{0}$ denote the zero vector and I_d denote the identity matrix of dimension $d \times d$. We denote $\mathbb{S}^d$ as the unit sphere of dimension d. For any cumulative distribution function F defined on $\mathbb{R}$, we let $F^{-1}(t) \equiv \inf\{x : F(x) \geq t\}$ denote the quantile function. For any two random variables W and V, we use $F^{-1}_{W|v}(\cdot)$ to denote the quantile function derived from the distribution of W conditioning on $V = v$.

2 Model and Motivating Examples

Denote μ_{1X} and μ_{0X} as the probability distributions of (Y_1, X) and (Y_0, X) respectively. Our model is defined by (1) and Assumption 2.1 below.

Assumption 2.1. *(i)* $d_1 \geq 1$, $d_0 \geq 1$, *and* $d_x \geq 0$. *(ii) The distributions* μ_{1X} *and* μ_{0X} *are identifiable from the sample information contained in two separate datasets, but the joint distribution* μ_o *is not identifiable. (iii) The projections of* μ_o *on* (Y_1, X) *and* (Y_0, X) *are* μ_{1X} *and* μ_{0X} *respectively.*

To avoid the trivial case, we assume that $d_1 \geq 1$ and $d_0 \geq 1$, but $d_x \geq 0$. If

$d_x = 0$, then there is no common X in both datasets. Assumption 2.1 (iii) can be interpreted as a comparability assumption. Consider, for example, the first dataset that contains only observations on (Y_1, X). Since Y_0 is missing, the distribution of (Y_0, X) is not identifiable. Assumption 2.1 (iii) requires the underlying probability distribution of (Y_0, X) in the first dataset to be the same as the one of (Y_0, X) in the second dataset. The latter probability distribution μ_{0X} is identifiable because the second dataset contains observations on (Y_0, X). Furthermore, it implies that the distribution of X in both datasets is the same.

2.1 Linear Projection Model: Short and Long

Consider the linear projection model:

$$Y_1 = \left(Y_0^\top, X^\top\right)\theta^* + \epsilon \text{ and } \mathbb{E}\left[\epsilon \begin{pmatrix} Y_0 \\ X \end{pmatrix}\right] = \mathbf{0}.$$

In many applications such as those discussed in Pacini (2019) and Hwang (2023), researchers do not observe random samples on (Y_1, Y_0, X) but observe two sets of samples on (Y_1, X) and (Y_0, X) respectively. In consequence, we can identify μ_{1X} and μ_{0X} separately but not μ_o. Let

$$m\left(Y_1, Y_0, X; \theta\right) = \left[\left(Y_0^\top, X^\top\right)\theta - Y_1\right] \begin{pmatrix} Y_0 \\ X \end{pmatrix}.$$

Then we have $\mathbb{E}_o\left[m\left(Y_1, Y_0, X; \theta^*\right)\right] = \mathbf{0}$. This is the model studied in Pacini (2019) and Hwang (2023). Both focus on scalar Y_0 but encounter difficulty when $d_0 > 1$. In Section 4, we derive the identified set and its support function for any value of d_0.

2.2 Example 2: Demographic Disparity

We illustrate the applicability of our methodology in assessing Algorithmic Fairness through data combination studied in Kallus et al. (2022). We focus on the demographic disparity (DD) measure in this section and the true-positive rate disparity (TPRD) measure in the next section. Other measures such as True-Negative Rate Disparity (TNRD) can be studied in the same way. The assumptions on the data imposed in this and next sections align with Kallus et al. (2022).

Let $Y_1 \in \{0,1\}$ denote the binary decision outcome obtained from either human decision-making or machine learning algorithms. For instance, $Y_1 = 1$ represents approval of a loan application. Let Y_0 be the protected attribute such as race or gender that takes values in $\{a_1, ..., a_J\}$. Researchers might be interested in knowing the disparity in within-class average loan approval rates. Such a measure is called demographic disparity and is defined as $\Pr(Y_1 = 1 \mid Y_0 = a_j) - \Pr\left(Y_1 = 1 \mid Y_0 = a_{j^\dagger}\right)$ between classes a_j and $a_{j^\dagger}$. Define $\theta_j^* \equiv \Pr(Y_1 = 1 \mid Y_0 = a_j)$ for $j = 1, \ldots, J$. The DD measure $\delta_{DD}\left(j, j^\dagger\right)$ for any $j \neq j^\dagger$ is defined as $\delta_{DD}\left(j, j^\dagger\right) \equiv \theta_j^* - \theta_{j^\dagger}^*$. Denote X as the set of additional observed covariates. Assume that we observe (Y_1, X) and (Y_0, X) separately.

Kallus et al. (2022) study the identified set for $[\delta_{DD}(1, J), \ldots, \delta_{DD}(J-1, J)]$, where a_J is treated as an advantage/reference group. For $J = 2$, they provide a closed-form for the identified set for $\delta_{DD}(1, 2)$. For $J > 2$, they state that the identified set is convex and characterize its support function evaluated at each direction as an infinite dimension linear program. Applying the result in the paper, in Section 5 we are able to derive a closed-form expression of the support function for $J > 2$.

Denote $\theta^* \equiv (\theta_1^*, ..., \theta_J^*)^\top$. We characterize θ^* via the moment model (1) with the following moment function:

$$m(Y_1, Y_0, X; \theta) = \begin{pmatrix} \theta_1 I\{Y_0 = a_1\} - I\{Y_1 = 1, Y_0 = a_1\} \\ \vdots \\ \theta_J I\{Y_0 = a_J\} - I\{Y_1 = 1, Y_0 = a_J\} \end{pmatrix},$$

where $\theta \equiv (\theta_1, ..., \theta_J)^\top$. It is easy to see that that $\mathbb{E}_o[m(Y_1, Y_0, X; \theta^*)] = \mathbf{0}$.

Let $e_+(j) \in \mathbb{R}^J$ ($e_-(j) \in \mathbb{R}^J$) be a row vector such that the j-th element of $e_+(j)$ ($e_-(j)$) is 1 (-1) and all the remaining elements are zero. Then any DD measure $\delta_{DD}\left(j, j^\dagger\right)$ can be expressed as $\left[e_+(j) + e_-\left(j^\dagger\right)\right]\theta^*$. Suppose we are interested in K different DD measures, where K can be smaller than, greater than, or equal to $J-1$. The vector of DD measures can be written as $E\theta^*$, where $E \in \mathbb{R}^{K \times J}$ is a matrix such that each row of E is of the form $e_+(j) + e_-\left(j^\dagger\right)$. The identified set for $E\theta^*$ follows from that of θ^*. For a particular

$$E^\star = \begin{pmatrix} 1 & \cdots & 0 & -1 \\ & \ddots & & \vdots \\ 0 & \cdots & 1 & -1 \end{pmatrix},$$

we obtain $E^{\star}\theta = [\delta_{DD}(1,J), \ldots, \delta_{DD}(J-1,J)]$ in Kallus et al. (2022).

2.3 Example 3: True-Positive Rate Disparity

Let $Y_{1s} \in \{0,1\}$ be the decision outcome and $Y_{1r} \in \{0,1\}$ be the true outcome. The true outcome is the target that justifies an optimal decision. Y_0 and X denote the protected attribute and proxy variable introduced in the previous section. True-positive rate disparity measures the disparity in the proportions of people who correctly get approved in loan applications between two classes, given their true non-default outcome. The TPRD measure between any two classes a_j and $a_{j^\dagger}$ is defined as

$$\delta_{TPRD}\left(j, j^\dagger\right) \equiv \Pr\left(Y_{1s} = 1 \mid Y_{1r} = 1, Y_0 = a_j\right) - \Pr\left(Y_{1s} = 1 \mid Y_{1r} = 1, Y_0 = a_{j^\dagger}\right).$$

Let $Y_1 \equiv (Y_{1s}, Y_{1r})$. We assume that (Y_1, X) and (Y_0, X) can be observed separately.

Kallus et al. (2022) study the identified set for $[\delta_{TPRD}(1,J), \ldots, \delta_{TPRD}(J-1,J)]$. When $J = 2$, they establish sharp bounds on $\delta_{TPRD}(1,2)$. For cases where $J > 2$, Kallus et al. (2022) state that the identified set is non-convex and provide the support function of its convex hull through rather complicated non-convex optimizations. As a result, it is difficult to directly analyze the properties of the identified set such as its connectedness. Moreover, solving the optimizations can be computationally intense.

Instead of directly analyzing the identified set for the TPRD measures, which is non-convex, we propose to formulate the TPRD measures as a nonlinear function of some carefully chosen θ^* such that the identified set Θ_I of θ^* is convex. Using this formulation, we show that the identified set for the TPRD measures is connected and can be obtained as a nonlinear map from Θ_I. The support function of Θ_I at each direction is an optimal transport problem, for which we introduce a simple algorithm to compute. We also provide a closed-form expression of the identified set for single TPRD measure that is valid for any $J \geq 2$.

Let $\theta^* \equiv (\theta^*_1, ..., \theta^*_{2J})$, where for $j = 1, ..., J$, we define

$$\theta^*_{2j-1} \equiv \Pr\left(Y_{1s} = 1, Y_{1r} = 1, Y_0 = a_j\right) \text{ and } \theta^*_{2j} \equiv \Pr\left(Y_{1s} = 0, Y_{1r} = 1, Y_0 = a_j\right).$$

For any $j \neq j^\dagger$, define a nonlinear map $g_{j,j^\dagger} : [0,1]^{2J} \to [-1,1]$ as

$$g_{j,j^\dagger}(\theta^*) = \frac{\theta^*_{2j-1}}{\theta^*_{2j-1} + \theta^*_{2j}} - \frac{\theta^*_{2j^\dagger-1}}{\theta^*_{2j^\dagger-1} + \theta^*_{2j^\dagger}}.$$

The TPRD measure $\delta_{TPRD}\left(j, j^{\dagger}\right)$ between classes a_j and $a_{j^{\dagger}}$ can be expressed as $\delta_{TPRD}\left(j, j^{\dagger}\right) = g_{j,j^{\dagger}}\left(\theta^*\right)$. And we can represent any K different TPRD measures as $G\left(\theta^*\right)$, where $G : [0,1]^{2J} \to [-1,1]^K$ is a multidimensional nonlinear map such that each row of $G\left(\cdot\right)$ takes the form of $g_{j,j^{\dagger}}\left(\cdot\right)$.

For $\theta = (\theta_1, ..., \theta_{2J})$, we let the moment function be

$$m\left(Y_1, Y_0, X; \theta\right) = \begin{pmatrix} \theta_1 - I\left\{Y_{1s} = 1 \bmod 2, Y_{1r} = 1, Y_0 = a_1\right\} \\ \theta_2 - I\left\{Y_{1s} = 2 \bmod 2, Y_{1r} = 1, Y_0 = a_1\right\} \\ \vdots \\ \theta_{2J} - I\left\{Y_{1s} = 2J \bmod 2, Y_{1r} = 1, Y_0 = a_J\right\} \end{pmatrix}.$$

It holds that that $\mathbb{E}_o\left[m\left(Y_1, Y_0, X; \theta^*\right)\right] = \mathbf{0}$.

2.4 Example 4: Causal Inference Under Strong Ignorability

Let Y_1 and Y_0 denote the potential outcomes of a binary treatment. Define $Y \equiv Y_1 D + Y_0\left(1 - D\right)$ as the realized outcome, where D is the binary treatment indicator such that an individual with $D = 1$ receives the treatment and an individual with $D = 0$ does not receive the treatment. Strong ignorability stated below is commonly adopted in the literature to identify various average treatment effect parameters.

Assumption 2.2 (Strong Ignorability)**.** (i) For all $x \in \mathcal{X}$, (Y_1, Y_0) is jointly independent of D conditional on $X = x$. (ii) For all $x \in \mathcal{X}$, $0 < p\left(x\right) < 1$, where $p\left(x\right) \equiv \Pr\left(D = 1 \mid X = x\right)$.

The strong ignorability assumption is composed of two parts. The first one is often referred to as the unconfoundedness or selection-on-observables assumption; and the second one is the overlap/common support assumption. Under strong ignorability, for all $x \in \mathcal{X}$, it holds that

$$F_j\left(y \mid x\right) = \Pr\left(Y \leq y \mid X = x, D = j\right) \text{ for } j = 1, 0.$$

Suppose that a random sample $\{(Y_i, X_i, D_i)\}_{i=1}^n$ on (Y, X, D) is available. Then $F_1\left(\cdot \mid \cdot\right)$ and $F_0\left(\cdot \mid \cdot\right)$ are point identified. Moreover, since the distribution of X is identified, the unconditional distributions of (Y_1, X) and (Y_0, X) are also point identified. As a result, if the parameter of interest depends only on the distributions of (Y_1, X) and (Y_0, X), such as the average treatment effect, then it is point identified.

Let h be a known function of dimension $k \geq 1$ and

$$m\left(Y_1, Y_0, X; \theta\right) = \theta - h\left(Y_1, Y_0\right).$$

This is an example of model (1) with $\theta^* = \mathbb{E}_o\left[h\left(Y_1, Y_0\right)\right]$. For univariate potential outcomes Y_1, Y_0 and scalar θ^*, Fan et al. (2017) characterize the identified set for θ^*. And for two special classes of functions h, they establish interval identification of θ^*. For example, when h is supermodular, an application of monotone rearrangement inequality leads to interval identification of θ^* with the identified set given by

$$\left\{\theta \in \Theta : \begin{array}{c} \int\int_0^1 h\left(F_{1|x}^{-1}\left(u \mid x\right), F_{0|x}^{-1}\left(1-u \mid x\right)\right) du d\mu_X \\ \leq \theta \leq \int\int_0^1 h\left(F_{1|x}^{-1}\left(u \mid x\right), F_{0|x}^{-1}\left(u \mid x\right)\right) du d\mu_X \end{array}\right\},$$

where $F_{1|x}^{-1}\left(u \mid x\right)$ and $F_{0|x}^{-1}\left(\cdot \mid x\right)$ are conditional quantile functions of Y_1 given $X = x$ and Y_0 given $X = x$ respectively, and μ_X is the marginal distribution of X. However, this approach is not applicable when $d_\theta > 1$ or when Y_1 and Y_0 are multivariate.

3 The Identified Set for θ^* and its Characterization

Let $\Theta_I \subseteq \Theta$ denote the identified set for θ^* defined as

$$\Theta_I \equiv \left\{\theta \in \Theta : \mathbb{E}_\mu\left[m\left(Y_1, Y_0, X; \theta\right)\right] = \mathbf{0} \text{ for some } \mu \in \mathcal{M}\left(\mu_{1X}, \mu_{0X}\right)\right\}, \tag{2}$$

where $\mathbb{E}_\mu$ denotes the expectation taken with respect to some probability measure $\mu \in \mathcal{M}\left(\mu_{1X}, \mu_{0X}\right)$, and $\mathcal{M}\left(\mu_{1X}, \mu_{0X}\right)$ is the class of probability measures whose projections on (Y_1, X) and (Y_0, X) are μ_{1X} and μ_{0X}.

When both Y_1 and Y_0 are univariate, Fan et al. (2023b) express $\mathbb{E}_\mu\left[m\left(Y_1, Y_0, X; \theta\right)\right]$ in terms of the conditional copula of (Y_1, Y_0) given X. Exploiting the result that any smooth copula can be approximated arbitrarily well by a Bernstein copula, they establish the identified set for θ^* by varying the conditional copula in the Bernstein copula space assuming that the true conditional copula is invariant with respect to the value of X. To allow for potentially multivariate Y_1 and Y_0, we establish a novel characterization of the identified set Θ_I via a continuum of inequalities defined by optimal transport costs in the next subsection. Even when both Y_1 and Y_0 are univariate, our approach is more general than that in Fan et al. (2023b), as the latter

relies on copula being smooth and invariant to X.

3.1 A Characterization via Optimal Transport

Let $\mu_{1|x}$ and $\mu_{0|x}$ be the conditional distributions of Y_1 on $X = x$ and Y_0 on $X = x$ respectively. Denote μ_X as the probability distribution of X. Recall that $\mathcal{Y}_1$, $\mathcal{Y}_0$, and $\mathcal{X}$ are the supports of random variables Y_1, Y_0, and X. For any $x \in \mathcal{X}$, let $\mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)$ denote the class of probability distributions whose projections on Y_1 and Y_0 are $\mu_{1|x}$ and $\mu_{0|x}$.[3]

Applying the method of conditioning in Rüschendorf (1991), the identified set Θ_I defined in (2) can be equivalently expressed as

$$\Theta_I = \left\{ \theta \in \Theta : \begin{array}{c} \int \left[\int \int m\left(y_1, y_0, x; \theta\right) d\mu_{10|x}\left(y_1, y_0\right) \right] d\mu_X\left(x\right) = \mathbf{0} \\ \text{for some } \mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right) \text{ for any } x \in \mathcal{X} \end{array} \right\}. \tag{3}$$

To simplify the notation, from now on, we omit the variables of integration when there is no confusion.

Let $\mathbb{S}^k$ denote the unit sphere of dimension k. Define the following set characterized by a continuum of inequalities:

$$\Theta_o \equiv \left\{ \theta \in \Theta : \int \mathcal{KT}_{t^\top m}\left(\mu_{1|x}, \mu_{0|x}; x, \theta\right) d\mu_X \leq 0 \text{ for all } t \in \mathbb{S}^k \right\}, \text{ where} \tag{4}$$

$$\mathcal{KT}_{t^\top m}\left(\mu_{1|x}, \mu_{0|x}; x, \theta\right) \equiv \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int \int t^\top m\left(y_1, y_0, x; \theta\right) d\mu_{10|x}.$$

For given $x \in \mathcal{X}$ and $\theta \in \Theta$, $\mathcal{KT}_{t^\top m}\left(\mu_{1|x}, \mu_{0|x}; x, \theta\right)$ is the (Kantorovich) optimal transport (OT) cost between measures $\mu_{1|x}$ and $\mu_{0|x}$ with (ground) cost function $t^\top m\left(y_1, y_0, x; \theta\right)$. Here we state the explicit dependence of $\mathcal{KT}_{t^\top m}\left(\mu_{1|x}, \mu_{0|x}; x, \theta\right)$ on x and θ because the cost function $t^\top m\left(y_1, y_0, x; \theta\right)$ is allowed to depend on x and θ.

It is not difficult to see that $\Theta_I \subseteq \Theta_o$. With the following continuity condition on m, we prove that these two sets are equal.

Assumption 3.1. *For every $\theta \in \Theta$ and $x \in \mathcal{X}$, $m\left(y_1, y_0, x; \theta\right)$ is continuous in $\left(y_1, y_0\right)$ on $\mathcal{Y}_1 \times \mathcal{Y}_0$.*

Assumption 3.1 holds automatically when the supports of $\mu_{0|x}$ and $\mu_{1|x}$ are finite

[3] All the statement with "for any $x \in \mathcal{X}$" can be relaxed to "for almost any $x \in \mathcal{X}$ with respect to μ_X measure". We ignore such mathematical subtlety to simplify the discussion.

for every $x \in \mathcal{X}$, i.e., $\mu_{0|x}$ and $\mu_{1|x}$ are discrete measures, because any function on a finite set is continuous. As a result, Examples 2 and 3 introduced in Sections 2.2 and 2.3 satisfy Assumption 3.1 even though they involve indicator functions.

Theorem 3.1. *If Assumption 3.1 holds, then* $\Theta_I = \Theta_o$.

The proof of the theorem is provided in the appendix. It relies on the minimax theorem derived in Vianney and Vigeral (2015) and the result on the weak compactness of $\mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)$ in Fan et al. (2023a).

Theorem 3.1 characterizes the identified set for θ^* via a continuum of inequality constraints on the integral of the optimal transport cost $\mathcal{KT}_{t^\top m}(\mu_{1|x}, \mu_{0|x}; x, \theta)$ with respect to μ_X. This novel characterization allows us to explore both theoretical and computational tools developed in the optimal transport literature to study the identified set Θ_I. See e.g., Rachev and Rüschendorf (2006), Villani et al. (2008), Santambrogio (2015), and Peyré and Cuturi (2019). For example, using Theorem 3.1, we show in Section 3.2 that Θ_I is convex when the moment function m is affine in θ and the characterization Θ_o provides a straightforward way to study the support function of Θ_I. Furthermore, we revisit Examples 1-3 in Sections 4-6 and show that for all three examples, $\mathcal{KT}_{t^\top m}(\mu_{1|x}, \mu_{0|x}; x, \theta)$ either has a closed-form expression or can be substantially simplified.

3.1.1 The Role of the Conditioning Variable X

The conditioning variable X provides information on the joint distribution of (Y_1, Y_0) by restricting the conditional distributions of Y_1 on $X = x$ and Y_0 on $X = x$ to be $\mu_{1|x}$ and $\mu_{0|x}$, which are identifiable from the datasets. The set obtained without taking X into consideration can be potentially larger than the identified set, which includes the information of X. Consider the special case that the moment function does not depend on x and denote it as $m(y_1, y_0; \theta)$. This includes Examples 2-4. Define a set using information only on the distributions of Y_1 and Y_0:

$$\Theta^O \equiv \left\{ \theta \in \Theta : \inf_{\mu_{10} \in \mathcal{M}(\mu_1, \mu_0)} \int \int t^\top m\left(y_1, y_0; \theta\right) d\mu_{10} \leq 0 \text{ for all } t \in \mathbb{S}^k \right\}. \tag{5}$$

The following proposition shows that Θ_I, which exploits the information on X, is at most as large as Θ^O, which ignores such information.

Proposition 3.1. Under Assumption 3.1, it holds that $\Theta_I \subseteq \Theta^O$.

As long as X is not independent of both Y_1 and Y_0, we expect the identified set Θ_I to be smaller than Θ^O. This is in sharp contrast to the complete data case with identified θ^*, where X would be irrelevant as far as identification of θ^* is concerned. In general, the stronger X and Y_1 (or Y_0) are dependent on each other, the smaller is the identified set for θ^*. When one of Y_1 and Y_0 is perfectly dependent on X, the identified set becomes the one for complete data, and θ^* is point identified under the standard rank condition. We show this result in the following proposition for the case where Y_0 is perfectly dependent on X. Without the common X in both datasets, this would only be possible in the extreme case that one of Y_1 and Y_0 is constant almost surely.

Proposition 3.2. Under Assumption 3.1, if the conditional measure $\mu_{0|x}$ is Dirac at $g_0(x)$, then Θ_I reduces to the identified set for the moment model with complete data:

$$\Theta_I = \{\theta \in \Theta : \mathbb{E}\left[m\left(Y_1, g_0\left(X\right), X; \theta\right)\right] = \mathbf{0}\}.$$

The expectation in the proposition is taken with respect to the joint distribution of Y_1 and X, which is identifiable from the data.

3.2 A Characterization for Affine Moment Functions

In this section, we show that if the moment function m is affine in θ, then we can exploit this structure to establish convexity of Θ_I and derive a simple expression of its support function. Throughout the discussion, we assume that Assumption 3.1 holds so that Theorem 3.1 applies.

Assumption 3.2. *Let $m_1(y_0, x)$ and $m_1(y_1, x)$ be matrix-valued functions of dimension $k \times d_\theta$ and $m_2(y_1, y_0, x)$ be a vector-valued function of dimension k. One of the following decompositions of m holds:*

(i) $m(y_1, y_0, x; \theta) = m_1(y_0, x)\,\theta + m_2(y_1, y_0, x)$;

(ii) $m(y_1, y_0, x; \theta) = m_1(y_1, x)\,\theta + m_2(y_1, y_0, x)$.

For all the examples discussed in the paper, the moment function m satisfies Assumption 3.2. We focus on the first decomposition in the following discussion because we can always rename Y_1 and Y_0.

We make the following assumption on the parameter space.

Assumption 3.3. *Θ is compact and convex with a nonempty interior.*

Theorem 3.2. *(i) Let Assumptions 3.1 and 3.2 hold. Define*

$$\mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) \equiv \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int \int t^\top m_2\left(y_1, y_0, x\right) d\mu_{10|x}. \tag{6}$$

Then, the identified set Θ_I can be rewritten as

$$\Theta_I = \left\{ \theta \in \Theta : t^\top \mathbb{E}\left[m_1\left(Y_0, X\right)\right] \theta \leq -\int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \text{ for all } t \in \mathbb{S}^k \right\}.$$

(ii) Furthermore, if Assumption 3.3 also holds, then Θ_I is closed and convex.

Note that $\mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right)$ does not depend on θ, because the cost function $t^\top m_2\left(y_1, y_0, x\right)$ does not depend on θ. Under Assumption 3.3, the convexity of Θ_I follows immediately from Theorem 3.2 (i), because the constraints in the expression of Θ_I in Theorem 3.2 (i) are affine in θ.

3.2.1 The Support Function of Θ_I

For a convex set $\Psi \subseteq \mathbb{R}^{d_\theta}$, its support function $h_\Psi\left(\cdot\right) : \mathbb{S}^{d_\theta} \to \mathbb{R}$ is defined pointwise by $h_\Psi\left(q\right) = \sup_{\psi \in \Psi} q^\top \psi$. Any non-empty, closed convex set is uniquely characterized by its support function. Moreover, the support function is compatible with many natural geometric operations, like scaling, translation, and rotation, and it provides the maximal and minimal contrasts achieved over a set. We use the latter property in Examples 2 and 3 to obtain the identified set for a single DD measure and a single TPRD measure for any number of protected classes.

Our novel characterization of the identified set in Theorem 3.2 provides a straightforward way to obtain the support function. We make the following two additional assumptions. Define

$$\Theta_R \equiv \left\{ \theta : t^\top \mathbb{E}\left[m_1\left(Y_0, X\right)\right] \theta \leq -\int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \text{ for all } t \in \mathbb{S}^k \right\},$$

where $\mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right)$ is defined in (6). The first assumption is on the parameter space Θ.

Assumption 3.4. *It holds that $\Theta_I = \Theta_R$.*

By definition, we have that $\Theta_R \cap \Theta = \Theta_I$. Assumption 3.4 requires that the parameter space Θ does not provide additional information on the identified set Θ_I

given Θ_R. It implies that the boundary of Θ_I is completely determined by the continuum of inequalities in Theorem 3.2 (i). For all the examples, we can easily choose Θ to make Assumption 3.4 hold. The second assumption is on the moment function.

Assumption 3.5. *Suppose $k = d_\theta$ and $\mathbb{E}\left[m_1\left(Y_0, X\right)\right]$ has full rank.*

Assumption 3.5 requires that the number of moments is the same as the number of parameters and that the coefficient matrix is of full rank. It implies that if the joint distribution μ_o of (Y_1, Y_0, X) is identifiable from the data, then θ^* is just identified but not overly-identified. For the examples in the paper, Assumption 3.5 is satisfied either automatically because $\mathbb{E}\left[m_1\left(Y_0, X\right)\right]$ is a diagonal matrix with positive diagonal entries or under very mild conditions.

Under Assumption 3.5, the expression of Θ_I in Theorem 3.2 (i) is equivalent to

$$\begin{aligned}
\Theta_I &= \left\{ \ \theta \in \Theta : q^\top \theta \leq s\left(q\right) \ \text{ for all } q \in \mathbb{S}^{d_\theta} \ \right\}, \text{ where} \\
s\left(q\right) &\equiv -\int \mathcal{KT}_{q^\top \mathbb{E}[m_1(Y_0,X)]^{-1} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \\
&= -\int \left[\inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int\int q^\top \mathbb{E}\left[m_1\left(Y_0, X\right)\right]^{-1} m_2\left(y_1, y_0, x\right) d\mu_{10|x} \right] d\mu_X .
\end{aligned}$$

As a result, we have that $q^\top \theta \leq s\left(q\right)$ for all $\theta \in \Theta_I$. If we can further show that for each $q \in \mathbb{S}^{d_\theta}$, $q^\top \theta = s\left(q\right)$ for at least one $\theta \in \Theta_I$, then $s\left(\cdot\right)$ is the support function of Θ_I. Define

$$\mathcal{M}_I = \left\{ \ \mu \in \mathcal{M}\left(\mu_{1X}, \mu_{0X}\right) : \mathbb{E}\left[m_1\left(Y_0, X\right)\right]\theta + \mathbb{E}_\mu\left[m_2\left(Y_1, Y_0, X\right)\right] = \mathbf{0} \text{ for some } \theta \in \Theta_I \ \right\}.$$

Assumptions 3.4 and 3.5 together imply that $\mathcal{M}_I = \mathcal{M}\left(\mu_{1X}, \mu_{0X}\right)$, which means that for any $\mu \in \mathcal{M}\left(\mu_{1X}, \mu_{0X}\right)$, there exists some $\theta \in \Theta_I$ such that the moment condition holds.[4] As a result, for any solution to the OT problem in the definition of $s\left(q\right)$, there is some $\theta \in \Theta_I$ that makes $q^\top \theta = s\left(q\right)$ hold. Thus, $s\left(q\right)$ is the support function of the identified set Θ_I.

Proposition 3.3. Under Assumptions 3.1-3.5, it holds that $h_{\Theta_I}\left(q\right) = s\left(q\right)$.

[4]Suppose $k \geq d_\theta$. Without Assumption 3.5, $\mathcal{M}_I$ can be a proper subset of $\mathcal{M}\left(\mu_{1X}, \mu_{0X}\right)$. If $\mathcal{M}_I$ is singleton, then θ^* is point identified even under Assumption 2.1. Bontemps et al. (2012) provide a detailed discussion on conditions under which point identification is restored in incomplete linear moment models when $k > d_\theta$.

Proposition 3.3 shows that we can obtain the value of the support function for any direction q by computing an optimal transport cost $\mathcal{KT}_{q^\top \mathbb{E}[m_1(Y_0,X)]^{-1} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right)$ for each x and then integrating with respect to μ_X. Moreover, if the parameter of interest is expressed as a linear map from θ^* through the linear operator L, then based on Proposition 3.3, the support function of the identified set for the parameter of interest can be simply calculated as $s\left(L^\star q\right)$, where $L^\star$ is the adjoint operator with respect to the inner product.

As a simple illustration, consider Example 4. The moment function satisfies Assumptions 3.2 and 3.5 with $m_1\left(y_0, x\right) = I_{d_\theta}$ and $m_2\left(y_1, y_0, x\right) = -h\left(y_1, y_0, x\right)$. If the parameter space Θ satisfies Assumptions 3.3 and 3.4, then the identified set for θ^* is convex with support function given by

$$
\begin{aligned}
h_{\Theta_I}(q) &= -\int \mathcal{KT}_{-q^\top h}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \\
&= -\int \left[\inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int\int -q^\top h\left(y_1, y_0, x\right) d\mu_{10|x} \right] d\mu_X .
\end{aligned}
$$

The identification results in Section 3.1 apply to the most general model defined in (1). Even if any of Assumptions 3.2-3.5 fails, one can still use the original definition of Θ_o to study the identified set Θ_I. However, if Assumptions 3.2-3.5 are satisfied, then the result in Section 3.2 provides a mathematically and computationally attractive way to obtain the support function of Θ_I. In both cases, the computational bottleneck lies in the evaluation of the optimal transport cost: $\mathcal{KT}_{t^\top m}(\mu_{1|x}, \mu_{0|x}; x, \theta)$ in the general case and $\mathcal{KT}_{t^\top m_2}(\mu_{1|x}, \mu_{0|x}; x)$ in the affine case.

In the remaining sections of this paper, we illustrate through Examples 1-3 applications of existing results in OT literature to the computation/evaluation of the above OT costs. We verify assumptions in Section 3.2 and derive the support function of Θ_I. For Examples 2 and 3, we further discuss how to connect the identified of Θ_I to the identified set for the parameter of interest, which is a function of θ^*.

For Examples 1 and 2, we show that $\mathcal{KT}_{t^\top m_2}(\cdot)$ can be evaluated by the monotone rearrangement inequality. For Example 3, $\mathcal{KT}_{t^\top m_2}(\cdot)$ is an optimal partial transport cost and we introduce a simple algorithm (DREAM) to solve it.

4 Example 1: Linear Projection Model: Short and Long

Since the moment function satisfies Assumption 3.1, Theorem 3.1 provides that $\Theta_I = \Theta_o$. Additionally, Assumption 3.2 holds with

$$m_1(y_0, x) = \begin{pmatrix} y_0 \\ x \end{pmatrix} \left(y_0^\top, x^\top\right) \text{ and } m_2(y_1, y_0, x) = -\begin{pmatrix} y_1 y_0 \\ y_1 x \end{pmatrix}.$$

As a result, Theorem 3.2 (i) applies, where

$$\mathcal{KT}_{t^\top m_2}(\mu_{1|x}, \mu_{0|x}; x) = \inf_{\mu_{10|x} \in \mathcal{M}(\mu_{1|x}, \mu_{0|x})} \int\int -\left(t_0^\top y_1 y_0 + t_X^\top y_1 x\right) d\mu_{10|x},$$

where $t \equiv \left(t_0^\top, t_X^\top\right)^\top$ with $t_0 \in \mathbb{R}^{d_0}$ and $t_X \in \mathbb{R}^{d_x}$. Although Y_1 is univariate, Y_0 has dimension d_0 with d_0 being any positive integer. Consequently, $\mathcal{KT}_{t^\top m_2}(\cdot)$ is the OT cost between a univariate measure and a measure of dimension d_0. However, the linear structure of $m_2(y_1, y_0, x)$ in y_0 simplifies the computation of the potentially challenging $\mathcal{KT}_{t^\top m_2}(\cdot)$ to a univariate OT problem amenable to the application of monotone rearrangement inequality. In particular, it holds that

$$\begin{aligned}
\mathcal{KT}_{t^\top m_2}(\mu_{1|x}, \mu_{0|x}; x) &= \inf_{\mu_{10|x} \in \mathcal{M}(\mu_{1|x}, \mu_{0|x})} \left[\int\int -\left(t_0^\top y_0\right) y_1 d\mu_{10|x} - \int\int t_X^\top y_1 x d\mu_{10|x}\right] \\
&= -\int_0^1 F^{-1}_{t_0^\top Y_0|x}(u) F^{-1}_{Y_1|x}(u) du - \int t_X^\top y_1 x d\mu_{1|x}.
\end{aligned}$$

Proposition 4.1. Define $M \equiv \mathbb{E}\left[\begin{pmatrix} Y_0 \\ X \end{pmatrix}\left(Y_0^\top, X^\top\right)\right]$. The identified set for θ^* is

$$\Theta_I = \left\{\theta \in \Theta : t^\top M \theta \leq \int\int_0^1 F^{-1}_{t_0^\top Y_0|x}(u) F^{-1}_{Y_1|x}(u) du d\mu_X + t_X^\top \mathbb{E}(Y_1 X) \text{ for all } t \in \mathbb{S}^k\right\}.$$

Proposition 4.1 holds for any value of d_0. If Assumption 3.3 holds, then Θ_I is convex. If Θ is large enough and M is invertible, then we can apply Proposition 3.3 and obtain the support function for Θ_I evaluated at q as

$$\int\int_0^1 F^{-1}_{\left(q_0^\top A + q_X^\top C\right) Y_0|x}(u) F^{-1}_{Y_1|x}(u) du d\mu_X + \left(q_0^\top B + q_X^\top D\right) \mathbb{E}(Y_1 X),$$

where $q \equiv \left(q_0^\top, q_X^\top\right)^\top \in \mathbb{S}^{d_0+d_x}$ such that $q_0 \in \mathbb{R}^{d_0}$ and $q_X \in \mathbb{R}^{d_x}$ and $\begin{pmatrix} A & B \\ C & D \end{pmatrix} = M^{-1}$ being partitioned accordingly.

When Y_0 is a scalar, the following corollary verifies that Θ_I is equal to the identified set Θ_F obtained in Pacini (2019) and the one in Proposition 1 in Hwang (2023).

Corollary 4.1. *When Y_0 is a scalar, the identified set Θ_I obtained in Proposition 4.1 is the same as Θ_F in Pacini (2019) and the identified set in Proposition 1 in Hwang (2023).*

On the other hand, when Y_0 is multivariate, either the bound in Pacini (2019) or the computationally simple bound in Proposition 2 of Hwang (2023) is tight. Let

$$\begin{aligned} \mathbb{U}^{d_0} &\equiv \left\{t_0 \in \mathbb{R}^{d_0} : \text{at most one element in } t_0 \text{ is nonzero}\right\} \text{ and} \\ \mathbb{S}_{d_0}^{d_0+d_x} &\equiv \left\{t = (t_0, t_X) \in \mathbb{S}^{d_0+d_x} : t_0 \in \mathbb{U}^{d_0} \text{ and } t_X \in \mathbb{R}^{d_x}\right\}. \end{aligned}$$

The identified set obtained in Pacini (2019) is equivalent to a set defined by the same inequality constraint as in Θ_I in Proposition 4.1 but for $t \in \mathbb{S}_{d_0}^{d_0+d_x}$ rather than for all $t \in \mathbb{S}^{d_0+d_x}$. We demonstrate the difference via the following numerical examples.

4.1 Numerical Illustration

We perform two numerical exercises. The first one is to show the difference between our identified set Θ_I and the set provided in Pacini (2019) when $d_0 > 1$. In the second exercise, we illustrate the effect of the dependence between Y_0 and X on the size of the identified set.

Simulation 1. Define $\overline{X} \equiv (1, X)$. The linear projection model is defined as

$$Y_1 = \left(Y_0^\top, \overline{X}^\top\right)\theta^* + \epsilon \text{ and } \mathbb{E}\left(\epsilon \begin{pmatrix} Y_0 \\ \overline{X} \end{pmatrix}\right) = 0. \tag{7}$$

Let $\left(Y_0^\top, X\right) \equiv (Y_{0a}, Y_{0b}, X) \sim \mathcal{N}(\mathbf{0}, \Omega)$, where

$$\Omega = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is the covariance matrix. For $\theta^* \equiv (\alpha_a^*, \alpha_b^*, \beta_1^*, \beta_2^*) = (1, 1, 0, 1)$, we let

$$\begin{aligned} Y_1 &= \alpha_a^* Y_a + \alpha_b^* Y_b + \beta_1^* + \beta_2^* X + \epsilon \\ &= Y_a + Y_b + X + \epsilon, \end{aligned}$$

where ϵ follows a standard normal distribution and is independent of (Y_{0a}, Y_{0b}, X).

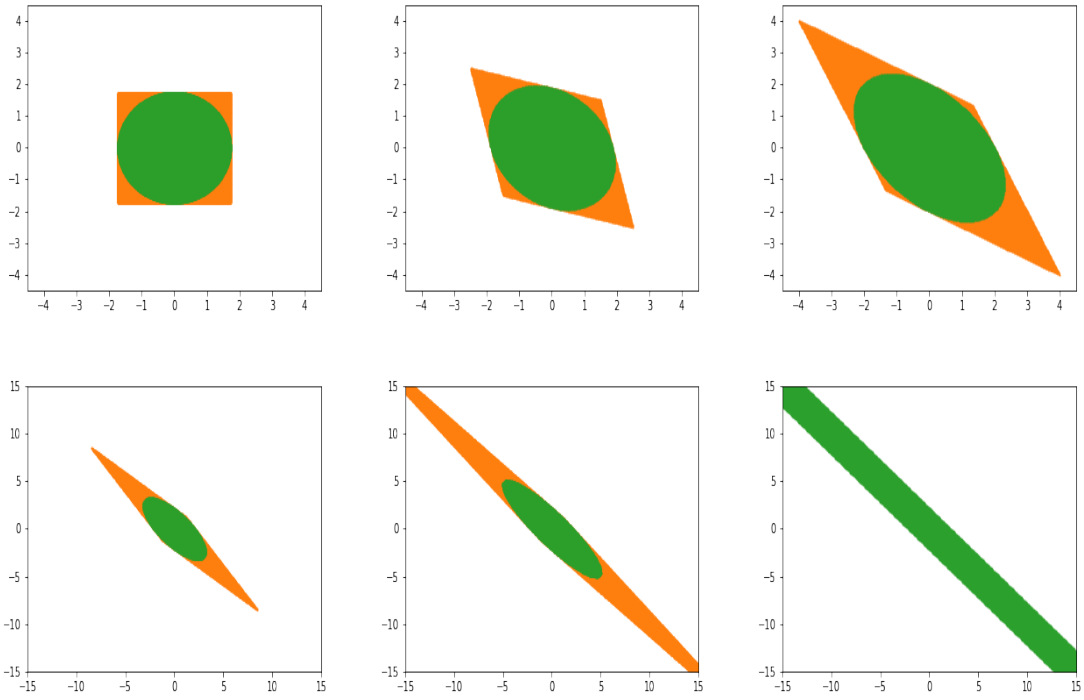

Figure 1: Top-left: $\rho = 0$; top-middle: $\rho = 0.25$; top-right: $\rho = 0.5$; bottom-left: $\rho = 0.75$; bottom-middle: $\rho = 0.9$; bottom-right: $\rho = 1$

Figure 1 compares the identified sets for (α_a^*, α_b^*) obtained from our method (green area) with the one from Pacini (2019) (orange area) for different values of ρ. When $\rho = 0$, random variables Y_{0a} and Y_{0b} are independent. However, even in this case, the approach in Pacini (2019) provides a larger identified set. This is because the method in Pacini (2019) ignores the possible connection between the dependence within (Y_{0a}, Y_1) and the dependence within (Y_{0b}, Y_1). For example, if the dependence within (Y_{0a}, Y_1) reaches the Fréchet–Hoeffding bound, then dependence within (Y_{0b}, Y_1) cannot reach the bound anymore, because that would imply a perfect correlation between Y_{0a} and Y_{0b}.

The bottom-right of Figure 1 illustrates the identified set when $\rho = 1$. When $|\rho| = 1$, the model is not identified even if we have the joint distribution of (Y_1, Y_0, X). In the case where $\rho = 1$, we can only identify $\alpha_a^* + \alpha_b^*$ but not individually. The approach in Pacini (2019) is not applicable, because it requires $\mathbb{E}\left[Y_0 Y_0^\top\right]$ be invertible. Since $\mathbb{E}\left[Y_0 Y_0^\top\right] = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ when $\rho = 1$, it is not invertible any more. On the other hand, our Proposition 4.1 still applies and provides the upper and lower bounds for $\alpha_a^* + \alpha_b^*$. Because we do not observe the distribution of $(Y_1, Y_0) = (Y_1, Y_{0a}, Y_{0b})$, we do not point identify $\alpha_a^* + \alpha_b^*$.

It is interesting to compare the identified sets between $|\rho| = 1$ and $|\rho| \neq 1$. When $\rho = \pm 1$, any t and $-t$ pair would provide the same upper and lower bound for $\alpha_a^* \pm \alpha_b^*$. On the other hand, when $|\rho| \neq 1$, the green identified set is an oval. Because for any given t, the linear inequality of α_a^*, α_b^* in Proposition 4.1 defines a half-plane, we need t to vary along $\mathbb{S}^2$ to obtain the oval-shaped identified set. On the other hand, when $\rho = \pm 1$, the identified set for (α_a^*, α_b^*) is just an interval bound for $\alpha_a^* \pm \alpha_b^*$. We only need two values to obtain such an identified set. We can just pick any t and $-t$ pair to obtain the bound.

Simulation 2. Consider the same linear projection model as defined in (7). We now let $X \sim \mathcal{N}(0, 4)$ and $Y_0 \equiv (Y_{0a}, Y_{0b})$, where

$$Y_{0a} = X^2 + \eta_a \text{ and } Y_{0b} = Y_{0a}^2 + \eta_b$$

with $\eta_a \sim \mathcal{N}(0, \sigma_a^2)$ and $\eta_b \sim \mathcal{N}(0, \sigma_b^2)$. Random variables X, η_a, and η_b are mutually independent. The value of σ_a controls the dependence between Y_{0a} and X; and the value σ_b controls the dependence between Y_{0a} and Y_{0b}. We let $Y_1 = \alpha_a^* Y_a + \alpha_b^* Y_b + \beta_1^* + \beta_2^* X + \epsilon$, where $(\alpha_a^*, \alpha_b^*, \beta_1^*, \beta_2^*) = (1, 0.2, 1, 1)$ and ϵ follows a standard normal distribution and is independent of (Y_{0a}, Y_{0b}, X).

Figure 2 shows the identified sets for (α_a^*, α_b^*) based on our method (green area) and the ones based on Pacini (2019) (orange area) for different values of σ_a and σ_b. First, we see that as σ_a and σ_b decrease, the green identified sets become smaller. This agrees with our discussion in Section 3.1 on the role of X in reducing the size of the identified set because the dependence between X and Y_{0a} and the dependence between X and Y_{0b} become stronger when σ_a and σ_b decrease. Second, the identified sets obtained from our method are always smaller than the ones from Pacini (2019).

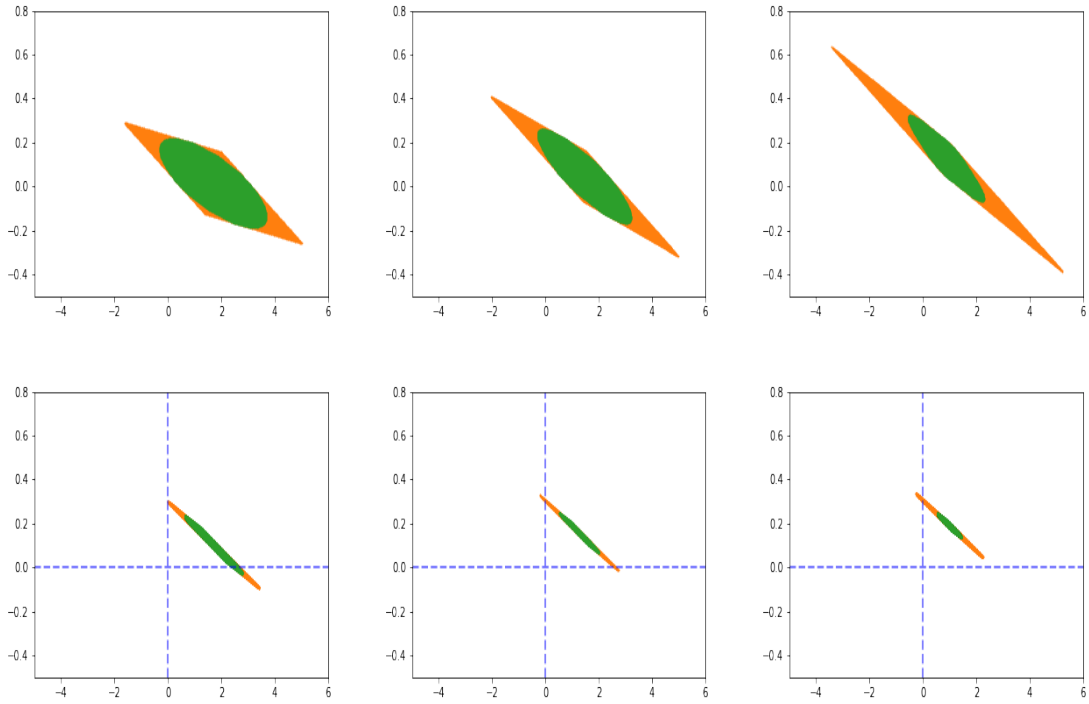

Figure 2: Top-left: $\sigma_a = 2, \sigma_b = 40$; top-middle: $\sigma_a = 2, \sigma_b = 20$ top-right: $\sigma_a = 2, \sigma_b = 2$; bottom-left: $\sigma_a = 0.5, \sigma_b = 20$; bottom-middle: $\sigma_a = 0.5, \sigma_b = 4$ bottom-right: $\sigma_a = 0.5, \sigma_b = 0.1$

Such a difference can be crucial in cases such as those described in the last two graphs. When $\sigma_a = 0.5$ and $\sigma_b = 4$, our identified set becomes small enough to lie in the first quadrant. On the other hand, the identified set from Pacini (2019) includes both $\alpha_a^* = 0$ and $\alpha_b^* = 0$. When $\sigma_a = 0.5$ and $\sigma_b = 0.1$, both Y_{0a} and Y_{0b} are strongly dependent on X. However, because the method in Pacini (2019) ignores the dependence between Y_{0a} and Y_{0b}, the orange identified set still cannot exclude $\alpha_a^* = 0$ even in this case.

5 Example 2: Demographic Disparity

5.1 Identified Set for θ^*

Because Y_1 and Y_0 are both discrete random variables, Assumption 3.1 holds. The moment function m also satisfies Assumption 3.2 with

$$m_1(y_0, x) = \text{diag}\left(I\{y_0 = a_1\}, \ldots, I\{y_0 = a_J\}\right) \text{ and}$$
$$m_2(y_0, y_1, x) = -\left(I\{y_1 = 1, y_0 = a_0\}, \ldots, I\{y_1 = 1, y_0 = a_J\}\right)^\top.$$

Therefore, the identified set Θ_I of θ^* follows from Theorem 3.2 (i), where

$$\mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) = \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int\int \sum_{j=1}^{J} -t_j I\{y_1 = 1, y_0 = a_j\} \, d\mu_{10|x}.$$

Differently from Example 1, the OT cost $\mathcal{KT}_{t^\top m_2}(\cdot)$ is between two univariate measures with a complicated ground cost function $\sum_{j=1}^{J} -t_j I\{y_1 = 1, y_0 = a_j\}$. To solve it, we let $d(y_1) = I\{y_1 = 1\}$ and $d_t(y_0) = \sum_{j=1}^{J} t_j I\{y_0 = a_j\}$. We then have

$$\begin{aligned}\mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) &= \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int\int -d(y_1) \times d_t(y_0) \, d\mu_{10|x}(y_1, y_0) \\ &= -\int_0^1 F^{-1}_{D|x}(u) \, F^{-1}_{D_t|x}(u) \, du\end{aligned}$$

with $D = I\{Y_1 = 1\}$ and $D_t = \sum_{j=1}^{J} t_j I\{Y_0 = a_j\}$. The last equality follows from monotone rearrangement inequality.

Proposition 5.1. The identified set for θ^* is

$$\Theta_I = \left\{\theta \in \Theta : \sum_{j=1}^{J} t_j \theta_j \Pr(Y_0 = a_j) \leq \int\int_0^1 F^{-1}_{D|x}(u) \, F^{-1}_{D_t|x}(u) \, du d\mu_X \text{ for all } t \in \mathbb{S}^J\right\}.$$

Because the elements of θ^* are probabilities, we have $\Theta = [0,1]^J$. Both Assumptions 3.3 and 3.4 hold. Assumption 3.5 is satisfied with

$$\mathbb{E}\left[m_1(Y_0, X)\right] = \text{diag}\left(\Pr(Y_0 = a_1), \ldots, \Pr(Y_0 = a_J)\right).$$

We apply Proposition 3.3 to obtain the support function of Θ_I as

$$h_{\Theta_I}(q) = \int \int_0^1 F^{-1}_{D|x}(u) F^{-1}_{D_q|x}(u) du d\mu_X,$$

where $D_q = \sum_{j=1}^J q_j I\{Y_0 = a_j\} \Pr(Y_0 = a_j)^{-1}$.

5.2 Identified Set for any Collection of DD Measures

The vector of K different DD measures $E\theta^*$ is a known function of θ^*. In consequence, we can express the identified set for $E\theta^*$, which is denoted as Δ_{DD}, by mapping each element θ in Θ_I through $E\theta$: $\Delta_{DD} = \{E\theta : \theta \in \Theta_I\}$. Meanwhile, because Θ_I is convex and E is a linear map, Δ_{DD} is convex. We can characterize Δ_{DD} through its support function:

$$\Delta_{DD} = \left\{\delta : p^\top \delta \leq h_{\Delta_{DD}}(p) \text{ for all } p \in \mathbb{S}^K\right\}, \tag{8}$$

where for any $p \in \mathbb{S}^K$, the support function of Δ_{DD} can be immediately obtained as

$$h_{\Delta_{DD}}(p) = h_{\Theta_I}\left(E^\top p\right) = \int \int_0^1 F^{-1}_{D|x}(u) F^{-1}_{D_{E^\top p}|x}(u) du d\mu_X. \tag{9}$$

Given p, both D and $D_{E^\top p}$ are discrete random variables. As a result, we can easily obtain their quantile functions. For any given x, in principle, we can compute $\int_0^1 F^{-1}_{D|x}(u) F^{-1}_{D_{E^\top p}|x}(u) du$ completely by hand.

Our closed-form expression of the support function of Δ_{DD} in (9) provides an easy way to obtain the identified set for any single DD measure $\delta_{DD}\left(j, j^\dagger\right)$ for $J \geq 2$. For instance, if we are interested in $\delta_{DD}(1,2)$, then we can let $K = 1$ and $E = (1, -1, 0, \ldots, 0)$. Then p can be either 1 or -1. We first let $p = 1$. Plugging p in the support function $h_{\Delta_{DD}}(\cdot)$, we obtain the following equalities:

$$h_{\Delta_{DD}}(1) = h_{\Theta_I}\left(E^\top\right) = \sup_{\theta \in \Theta_I} E^\top \theta = \sup_{\theta \in \Theta_I} (\theta_1 - \theta_2) = \sup_{\theta \in \Theta_I} \delta_{DD}(1,2).$$

As a result, it suffices to compute $h_{\Theta_I}\left(E^\top\right)$ to obtain the tight upper bound of $\delta_{DD}(1,2)$. The tight lower bound can be obtained similarly with $p = -1$. Since any one-dimensional convex set is either an interval or a degenerated interval, i.e. a point, the tight upper and lower bounds characterize the identified set for $\delta_{DD}(1,2)$. The following corollary provides the expression for the identified set for $\delta_{DD}\left(j, j^\dagger\right)$ for $J \geq 2$.

Corollary 5.1. *For any $J \geq 2$, the identified set for any single DD measure: $\delta_{DD}\left(j, j^{\dagger}\right)$ is given by $\left[\delta_{DD}^{L}\left(j, j^{\dagger}\right), \delta_{DD}^{U}\left(j, j^{\dagger}\right)\right]$, where*

$$\begin{aligned}
\delta_{DD}^{U}\left(j, j^{\dagger}\right) \equiv & \frac{\mathbb{E}\left[\min\left\{P_A(X), P_B(X)\right\}\right]}{\Pr\left(Y_0 = a_j\right)} - \frac{\mathbb{E}\left[\max\left\{P_A(X) + P_C(X) - 1, 0\right\}\right]}{\Pr\left(Y_0 = a_{j^{\dagger}}\right)} \\
\delta_{DD}^{L}\left(j, j^{\dagger}\right) \equiv & \frac{\mathbb{E}\left[\max\left\{P_A(X) + P_B(X) - 1, 0\right\}\right]}{\Pr\left(Y_0 = a_j\right)} - \frac{\mathbb{E}\left[\min\left\{P_A(X), P_C(X)\right\}\right]}{\Pr\left(Y_0 = a_{j^{\dagger}}\right)},
\end{aligned}$$

in which $P_A(x) \equiv \Pr\left(Y_1 = 1 \mid X = x\right)$, $P_B(x) \equiv \Pr\left(Y_0 = a_j \mid X = x\right)$, and $P_C(x) \equiv \Pr\left(Y_0 = a_{j^{\dagger}} \mid X = x\right)$.

Because Y_1 and Y_0 are both one dimensional, the Fréchet-Hoeffding inequalities provide the identified set for θ_j^* as a closed interval $\left[\theta_j^L, \theta_j^U\right]$, where

$$\begin{aligned}
\theta_j^L = & \frac{\mathbb{E}\left[\max\left\{\Pr\left(Y_1 = 1 \mid X\right) + \Pr\left(Y_0 = a_j \mid X\right) - 1, 0\right\}\right]}{\Pr\left(Y_0 = a_j\right)} \text{ and} \\
\theta_j^U = & \frac{\mathbb{E}\left[\min\left\{\Pr\left(Y_1 = 1 \mid X\right), \Pr\left(Y_0 = a_j \mid X\right)\right\}\right]}{\Pr\left(Y_0 = a_j\right)}.
\end{aligned}$$

Since $\delta_{DD}\left(j, j^{\dagger}\right) = \theta_j^* - \theta_{j^{\dagger}}^*$, we have that $\theta_j^L - \theta_{j^{\dagger}}^U \leq \delta_{DD}\left(j, j^{\dagger}\right) \leq \theta_j^U - \theta_{j^{\dagger}}^L$. When $J = 2$, Kallus et al. (2022) prove that there is a joint probability on (Y_1, Y_0, X) that allows θ_j^* to achieve its lower (upper) bound and $\theta_{j^{\dagger}}^*$ to achieve its upper (lower) bound simultaneously. In consequence, interval $\left[\theta_j^L - \theta_{j^{\dagger}}^U, \theta_j^U - \theta_{j^{\dagger}}^L\right]$ is in fact the identified set for $\delta_{DD}\left(j, j^{\dagger}\right)$ when $J = 2$. Corollary 5.1 shows that the interval $\left[\theta_j^L - \theta_{j^{\dagger}}^U, \theta_j^U - \theta_{j^{\dagger}}^L\right]$ is the identified set for $\delta_{DD}\left(j, j^{\dagger}\right)$ even for $J > 2$.

5.3 Time Complexity

To construct the identified set Δ_{DD} via (8), we first sample N_p vectors $p_1, \ldots, p_{N_p}$ uniformly from the K-dimensional unit sphere and N_δ vectors $\delta_1, \ldots, \delta_{N_\delta}$ uniformly from $[-1, 1]^K$. Then, we construct the set:

$$\widehat{\Delta}_{DD} \equiv \left\{\delta = \delta_1, \ldots, \delta_{N_\delta} : p^{\top}\delta \leq h_{\Delta_{DD}}(p) \text{ for all } p = p_1, \ldots, p_{N_p}\right\}$$

as an approximate to Δ_{DD}.

Given any p, we compute $h_{\Delta_{DD}}(p)$ based on (9). Suppose a grid-based method is employed for the numerical integration with respect to measure μ_X. It remains to study the time complexity for computing $\int_0^1 F_{D|x}^{-1}(u) F_{D_{E^{\top}p}|x}^{-1}(u)\, du$. The following

lemma shows the result.

Lemma 5.1. *The time complexity for computing* $\int_0^1 F_{D|x}^{-1}(u) F_{D_{E^\top p}|x}^{-1}(u) du$ *is approximately* $\frac{1}{2}J^2 + \frac{3}{2}J + 2K$.[5]

Let N_x denote the number of grid points used in the numerical integration. Then for any given p, computing $h_{\Delta_{DD}}(p)$ requires $\left(\frac{1}{2}J^2 + \frac{3}{2}J + 2K\right) N_x$ number of operations. We obtain the overall time complexity as $\left(\frac{1}{2}J^2 + \frac{3}{2}J + 2K\right) N_x N_p + N_\delta N_p$.

5.4 Comparison with Kallus et al. (2022)

Kallus et al. (2022) study the specific collection of DD measures

$$E^\star \theta^* = \left[\delta_{DD}(1, J), \ldots, \delta_{DD}(J-1, J)\right].$$

Denote $\Delta_{DD}^\star$ as the identified set for $E^\star\theta^*$. They propose to approximate $\Delta_{DD}^\star$ by

$$\widehat{\Delta}_{DD}^\star \equiv \left\{\delta = \delta_1, \ldots, \delta_{N_\delta} : p^\top \delta \leq \int \Phi_K(p, x) \, d\mu_X(x) \text{ for all } p = p_1, \ldots, p_{N_p}\right\},$$

where $\int \Phi_K(p, x) \, d\mu_X(x)$ is the value of the support function evaluated at $p \in \mathbb{S}^{J-1}$. Kallus et al. (2022) develop the following result to compute $\Phi_K(p, x)$.

Lemma 5.2 (Proposition 10 in Kallus et al. (2022))**.** *Given* $p \in \mathbb{S}^{J-1}$ *and* $x \in \mathcal{X}$, $\Phi_K(p, x)$ *is the value of the following maximization problem:*

$$\max_{P_1(\cdot), \ldots, P_J(\cdot)} \sum_{j=1}^{J-1} \sum_{y_1 \in \{0,1\}} p_j \left(\frac{y_1 P_j(y_1)}{\Pr(Y_0 = a_j)} - \frac{y_1 P_J(y_1)}{\Pr(Y_0 = a_J)} \right) \Pr(Y_1 = y_1 \mid X = x) \quad (10)$$

s.t. for any $j = \{1, \ldots, J\}$ *and* $y_1 \in \{0, 1\}$:

(i) $0 \leq P_j(y_1) \leq 1$; *(ii)* $\sum_{j=1}^{J} P_j(y_1) = 1$; *and*

(iii) $\sum_{y_1 \in \{0,1\}} P_j(y_1) \Pr(Y_1 = y_1 \mid X = x) = \Pr(Y_0 = a_j \mid X = x)$.

For any given p and x, (10) is a linear program with $2J$ variables and about $5J$ (equality and inequality) constraints. The time complexity of solving (10) varies

[5] We ignore the effect of the bit-length, which refers to the number of bits required to represent the input data of the problem.

depending on the algorithm used to solve the linear programming program. For relatively small J (e.g. $J < 5$), the simplex method is more efficient and requires roughly $(5J)^2 = 25J^2$ number of operations (Dantzig (2016)); while for moderate J, interior-point methods are often computationally cheaper and require about $33\,(5J)^3 \approx 4000J^3$ number of operations (Karmarkar (1984); Wright (1997)).[6] When applied to $E^\star\theta^*$, Lemma 5.1 shows that our procedure computes $\Phi_K\,(p, x)$ for any given $p \in \mathbb{S}^{J-1}$ and $x \in \mathcal{X}$ with $\frac{1}{2}\,(J^2 + 7J)$ number of operations. This is advantageous even compared with the fastest theoretical running time for solving (10), which is roughly $\widetilde{O}\left((2J)^{2.37}\right)$, where the notation $\widetilde{O}$ hides polylogarithmic factors (van den Brand (2020); Cohen et al. (2021); Williams et al. (2024)).

6 Example 3: True-Positive Rate Disparity

6.1 Identified Sets of θ^* and TPRD Measures

Similar to Section 5, we obtain the identified set Θ_I of θ^* by applying Theorem 3.2 (i). Assumption 3.2 is satisfied with $m_1\,(y_0, x)$ being the identify matrix of dimension $2J \times 2J$ and

$$m_2\,(y_1, y_0, x) = -\begin{pmatrix} I\,\{y_{1s} = 1 \bmod 2, y_{1r} = 1, y_0 = a_1\} \\ I\,\{y_{1s} = 2 \bmod 2, y_{1r} = 1, y_0 = a_1\} \\ \vdots \\ I\,\{y_{1s} = 2J \bmod 2, y_{1r} = 1, y_0 = a_J\} \end{pmatrix},$$

where $y_1 \equiv (y_{1s}, y_{1r})$. Consequently, the identified set for θ^* is

$$\Theta_I = \left\{\theta \in \Theta : \sum_{j=1}^{2J} t_j\theta_j \leq -\int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \text{ for all } t \in \mathbb{S}^{2J}\right\}.$$

Furthermore, Assumptions 3.3, 3.4, and 3.5 are satisfied, because $\Theta = [0, 1]^{2J}$ and $\mathbb{E}\,[m_1\,(Y_0, X)]$ is the identity matrix. Thus, the support function of Θ_I for any $q \in \mathbb{S}^{2J}$

[6] The algorithms used in modern solvers, such as Mosek and Gurobi, share the ideas of the simplex method and interior-point methods. The information on the exact time complexity of these solvers is often not public.

is given by

$$h_{\Theta_I}(q) = -\int \mathcal{KT}_{q^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X. \tag{11}$$

Let Δ_{TPRD} denote the identified set for K different TPRD measures represented by $G(\theta^*)$. We can compute Δ_{TPRD} as a non-linear map from Θ_I:

$$\Delta_{TPRD} = \{\delta = G(\theta) : \theta \in \Theta_I\} = \left\{\delta = G(\theta) : q^\top \theta \leq h_{\Theta_I}(q) \text{ for all } q \in \mathbb{S}^{2J}\right\},$$

where $h_{\Theta_I}(\cdot)$ is defined in (11). Because Θ_I is convex and the map G is continuous, we know that Δ_{TPRD} is connected, which implies that the identified set for any single TPRD measure is a closed interval. Using the simple expression for the support function (11) of Θ_I, we can derive the closed-form expressions for the lower and upper endpoints of the interval, extending the result in Kallus et al. (2022) established for $J = 2$.

Corollary 6.1. *For any $J \geq 2$, the identified set for any single TPRD measure: $\delta_{TPRD}\left(j, j^\dagger\right)$ is given by $\left[\delta^L_{TPRD}\left(j, j^\dagger\right), \delta^U_{TPRD}\left(j, j^\dagger\right)\right]$, where*

$$\delta^U_{TPRD}\left(j, j^\dagger\right) = \frac{\theta^U_{2j-1}}{\theta^U_{2j-1} + \theta^L_{2j}} - \frac{\theta^L_{2j^\dagger-1}}{\theta^L_{2j^\dagger-1} + \theta^U_{2j^\dagger}} \text{ and}$$
$$\delta^L_{TPRD}\left(j, j^\dagger\right) = \frac{\theta^L_{2j-1}}{\theta^L_{2j-1} + \theta^U_{2j}} - \frac{\theta^U_{2j^\dagger-1}}{\theta^U_{2j^\dagger-1} + \theta^L_{2j^\dagger}},$$

where the expressions for each term in $\delta^U_{TPRD}\left(j, j^\dagger\right)$ and $\delta^L_{TPRD}\left(j, j^\dagger\right)$ are provided in the proof of the corollary.

For any $j \in \{1, \ldots, J\}$, θ^U_{2j-1} and θ^L_{2j-1} are the Fréchet-Hoeffding upper and lower bounds of $\Pr(Y_1 = (1,1), Y_0 = a_j)$; and θ^U_{2j} and θ^L_{2j} are the Fréchet-Hoeffding upper and lower bounds of $\Pr(Y_1 = (0,1), Y_0 = a_j)$. The proof reduces to computing $h_{\Theta_I}(q)$ for two carefully chosen values of q, which essentially solves two optimal transport problems.

For the general case of multiple measures, we need to compute $\mathcal{KT}_{q^\top m_2}(\cdot)$. In contrast to the OT costs in Examples 1 and 2, evaluating $\mathcal{KT}_{q^\top m_2}(\cdot)$ for Example 3 is more complicated because the ground cost function $t^\top m_2$ is a complex nonlinear function of $y_1 = (y_{1s}, y_{1r})$ and y_0. To understand the challenge and the idea underlying our algorithm, in the following, we rewrite the OT problem for $\mathcal{KT}_{q^\top m_2}(\cdot)$.

Given $t \equiv (t_1, \ldots, t_{2J})$, define $\pi(i,j) \equiv -t_{2j-(i \bmod 2)}$ for $i = 0, 1$ and $j = 1, \ldots, J$. Because Y_1 and Y_0 are both discrete random variables, we let

$$\Gamma_{0|x}(j) \equiv \int I\{y_0 = a_j\} \, d\mu_{0|x}(y_0) = \Pr(Y_0 = a_j \mid X = x) \text{ and}$$
$$\Gamma_{1|x}(i,1) \equiv \int I\{y_{1s} = i, y_{1r} = 1\} \, \mu_{1|x}(y_{1s}, y_{1r}) = \Pr(Y_{1s} = i, Y_{1r} = 1 \mid X = x).$$

Given any $\mu_{10|x} \in \mathcal{M}(\mu_{1|x}, \mu_{0|x})$, define

$$\Gamma_{10|x}(i,1,j) \equiv \int \int I\{y_1 = (i,1), y_0 = a_j\} \, d\mu_{10|x}(y_1, y_0).$$

The OT cost $\mathcal{KT}_{t^\top m_2}(\cdot)$ can then be rewritten as the following optimization problem with the argument $\Gamma_{10|x}(i,1,j)$:

$$\mathcal{KT}_{q^\top m_2}(\mu_{1|x}, \mu_{0|x}; x) = \min_{\Gamma_{10|x}} \sum_{j=1}^{J} \sum_{i=0}^{1} \pi(i,j) \, \Gamma_{10|x}(i,1,j) \tag{12}$$

$$\text{s.t. (i) } \Gamma_{10|x}(i,1,j) \geq 0 \text{ for } i = 0,1 \text{ and } j = 1, \ldots, J,$$
$$\text{(ii) } \sum_{i=0}^{1} \Gamma_{10|x}(i,1,j) \leq \Gamma_{0|x}(j) \text{ for } j = 1, \ldots, J, \text{ and}$$
$$\text{(iii) } \sum_{j=1}^{J} \Gamma_{10|x}(i,1,j) = \Gamma_{1|x}(i,1) \text{ for } i = 0,1.$$

Note that the second constraint is an inequality rather than an equality because the OT cost $\mathcal{KT}_{t^\top m_2}(\cdot)$ does not involve $\mu_{10|x}(y_1, y_0)$ for $y_1 = (y_{1s}, 0)$. In consequence, (12) is actually an optimal *partial* transport problem, where $\Gamma_{0|x}(j)$ and $\Gamma_{1|x}(i,1)$ are identified from the sample information.

The following lemma shows that there exists a solution to this reformulation of the optimal partial transport problem whose support is monotone.

Lemma 6.1. *Suppose that $\pi(i,j)$ is submodular for $i = 0,1$ and $j = 1, \ldots, J$, i.e., for any $j > j^\dagger$, it holds that*

$$\pi(1,j) - \pi(0,j) - \pi(1,j^\dagger) + \pi(0,j^\dagger) \leq 0.$$

Then there is a solution $\left\{\Gamma^*_{10|x}(i,1,j) : i = 0,1 \text{ and } j = 1, \ldots, J\right\}$ to (12) with monotone support which implies that for some $J^* \in \{1, \ldots, J\}$, $\Gamma^*_{10|x}(1,1,j) = 0$ for all

$j < J^*$ *and* $\Gamma^*_{10|x}(0,1,j) = 0$ *for all* $j > J^*$.

Lemma 6.1 is analogous to monotone rearrangement inequality for the classical OT problem with submodular cost. Since i in the optimization problem (12) takes two values only, we can always make π submodular and apply Lemma 6.1.

6.2 Dual Rank Equilibration Algorithm

Exploiting the structure of the solution $\Gamma^*_{10|x}(i,1,j)$ described in Lemma 6.1, we propose our Algorithm 1 for solving (12) and call it Dual Rank Equilibration AlgorithM (DREAM). DREAM involves only basic arithmetic, comparison, and logical operations, making it straightforward to implement directly in code. In fact, it can be executed by hand. DREAM is built on an equivalent characterization of (12) which allows it to use the ranks of $\pi(i,j)$ to equilibrate between minimizing the partial transport cost and respecting the constraints in (12). The equilibration is done twice by first separately adjusting $\Gamma^*_{10|x}(0,1,j)$ and $\Gamma^*_{10|x}(1,1,j)$ across index j and then jointly adjusting $\Gamma^*_{10|x}(0,1,j)$ and $\Gamma^*_{10|x}(1,1,j)$ for a fixed j—Dual Rank Equilibration AlgorithM.

In the rest of this section, we describe each of the three steps in DREAM and the equivalent optimization problem that it solves.

Algorithm 1: Dual Rank Equilibration Algorithm (Part 1)

Input : $\pi(i,j)$, $\Gamma_{1|x}(i,1)$, and $\Gamma_{0|x}(j)$ for $i = 0,1$ and $j = 1,\ldots,J$
Output: $\mathcal{KT}_{t^\top m_2}(\mu_{1|x}, \mu_{0|x}; x)$
`/* Initialization */`
$d(j) \leftarrow \pi(1,j) - \pi(0,j)$ for $j = 1,\ldots,J$;
Sort the indices j in ascending order of $d(j)$;
Relabel the indices in $\pi(i,j)$ and $\Gamma_{0|x}(j)$ accordingly;
$\Gamma_{10|x}(i,1,j) \leftarrow 0$ for $i = 0,1$ and $j = 1,\ldots,J$;
$JL \leftarrow 1$;
$JU \leftarrow J$;
`/* Step 1: Determine` JL `and` JU `*/`
while $\sum_{j=1}^{JL} \Gamma_{0|x}(j) < \Gamma_{1|x}(0,1)$ **do**
  $JL \leftarrow JL + 1$;
end
while $\sum_{j=JU}^{J} \Gamma_{0|x}(j) < \Gamma_{1|x}(1,1)$ **do**
  $JU \leftarrow JU - 1$;
end

Algorithm 1: Dual Rank Equilibration Algorithm (Part 2)

for $jj = JL$ **to** JU **do**

  `/* Step 2: Ranking and assignment based on` $\pi(0,j)$ `and` $\pi(1,j)$ `*/`

  Compute ranking $r_0(1), \ldots, r_0(jj)$ for $\pi(0,1), \ldots, \pi(0,jj)$;

  $m \leftarrow \Gamma_{1|x}(0,1)$;

  for $r \leftarrow 1$ **to** jj **do**

    Let w be the index where $r_0(w) = r$;

    $\Gamma_{10|x}(0,1,w) \leftarrow \min\{\Gamma_{0|x}(w),\, m\}$;

    $m \leftarrow m - \Gamma_{10|x}(0,1,w)$;

  end

  Compute ranking $r_1(1), \ldots, r_1(J - jj + 1)$ for $\pi(1,jj), \ldots, \pi(1,J)$;

  $m \leftarrow \Gamma_{1|x}(1,1)$;

  for $r \leftarrow 1$ **to** $J - jj + 1$ **do**

    Let w be the index where $r_1(w) = r$;

    $\Gamma_{10|x}(1,1,w+jj-1) \leftarrow \min\{\Gamma_{0|x}(w+jj-1),\, m\}$;

    $m \leftarrow m - \Gamma_{10|x}(1,1,w+jj-1)$;

  end

  `/* Step 3: Adjustment for feasibility */`

  while $\Gamma_{10|x}(0,1,jj) + \Gamma_{10|x}(1,1,jj) > \Gamma_{0|x}(jj)$ **do**

    $dd, llb \leftarrow []$;

    for $r \leftarrow r_0(jj) + 1$ **to** jj **do**

      Let w be the index where $r_0(w) = r$;

      Append $\pi(0,w) - \pi(0,jj)$ to dd;

      Append w to llb;

    end

    $jdd_0 \leftarrow$ length of dd;

    for $r \leftarrow r_1(1) + 1$ **to** $J - jj + 1$ **do**

      Let w be the index where $r_1(w) = r$;

      Append $\pi(1,w+jj-1) - \pi(1,jj)$ to dd;

      Append $w + jj - 1$ to llb;

    end

    $jdd \leftarrow$ length of dd;

    Compute ranking $rdd(1), \ldots, rdd(jdd)$ for dd;

    for $r \leftarrow 1$ **to** jdd **do**

      Let w be the index where $rdd(w) = r$;

      if $w \leq jdd_0$ **then**

        $i \leftarrow 0$;

      else

        $i \leftarrow 1$;

      end

      $mm_1 \leftarrow \Gamma_{10|x}(0,1,jj) + \Gamma_{10|x}(1,1,jj) - \Gamma_{0|x}(jj)$;

      $mm_2 \leftarrow \Gamma_{0|x}(lld(w)) - \Gamma_{10|x}(i,1,lld(w))$;

      $mm \leftarrow \min\{mm_1,\, mm_2\}$;

      $\Gamma_{10|x}(i,1,lld(w)) \leftarrow \Gamma_{10|x}(i,1,lld(w)) + mm$;

      $\Gamma_{10|x}(i,1,jj) \leftarrow \Gamma_{10|x}(i,1,jj) - mm$;

    end

  end

  $c(jj) \leftarrow \sum_{j=1}^{J} \sum_{i=0}^{1} \pi(i,j)\, \Gamma_{10|x}(i,1,j)$;

end

return

$\mathcal{KT}_{t^\top m_2}(\mu_{1|x}, \mu_{0|x}; x) \leftarrow \min\{c(JL), c(JL+1), \ldots, c(JU)\}$

Part 1 of DREAM includes Initialization and Step 1. During Initialization, we relabel j so that $\pi(i,j)$ is submodular.[7] Lemma 6.1 shows that there is a solution $\Gamma^*_{10|x}(i,1,j)$ with the following structure: there exists some $J^* \in \{1,\ldots,J\}$ such that $\Gamma^*_{10|x}(1,1,j) = 0$ for all $j < J^*$ and $\Gamma^*_{10|x}(0,1,j) = 0$ for all $j > J^*$. In Step 1, we narrow down the range of J^* to $\{JL, JL+1, \ldots, JU\}$.

The idea behind Step 1 is that the values of $\Gamma_{1|x}(0,1)$ and $\Gamma_{1|x}(1,1)$ only allow certain values of $J^* \in \{1,\ldots,J\}$ to satisfy the structure of the solution. Consider the lower bound JL. For any value of J^*, it must hold that

$$\sum_{j=1}^{J^*} \Gamma_{0|x}(j) \geq \sum_{j=1}^{J^*} \sum_{i=0}^{1} \Gamma_{10|x}(i,1,j) \geq \sum_{j=1}^{J^*} \Gamma_{10|x}(0,1,j) = \sum_{j=1}^{J} \Gamma_{10|x}(0,1,j) = \Gamma_{1|x}(0,1),$$

where the first inequality follows from the second constraint in (12), the second inequality holds by the first constraint in (12), the first equality follows from the structure of the solution, and the last equality holds by the third constraint in (12). Consequently, $\sum_{j=1}^{J^*} \Gamma_{0|x}(j) \geq \Gamma_{1|x}(0,1)$. Because $\sum_{j=1}^{JL} \Gamma_{0|x}(j)$ is an increasing function of JL, we start with $JL = 1$ and gradually increase JL until inequality $\sum_{j=1}^{JL} \Gamma_{0|x}(j) \geq \Gamma_{1|x}(0,1)$ holds. This provides us with the lower bound JL. Similarly, it must also be true that $\sum_{j=J^*}^{J} \Gamma_{0|x}(j) \geq \Gamma_{1|x}(1,1)$. Because $\sum_{j=JU}^{J} \Gamma_{0|x}(j)$ is a decreasing function of JU, we start with $JU = J$ and gradually decrease JU until $\sum_{j=JU}^{J} \Gamma_{0|x}(j) \geq \Gamma_{1|x}(1,1)$ holds. The resulting JU is the upper bound.

For any $jj \in \{JL, JL+1, ..., JU\}$, after imposing the structure that $\Gamma_{10|x}(1,1,j) = 0$ for all $j < jj$ and $\Gamma_{10|x}(0,1,j) = 0$ for all $j > jj$, the optimization problem (12)

[7] By reordering the label j, we can always make $\pi(i,j)$ submodular.

reduces to (13) below:

$$\min_{jj\in\{JL,JL+1,...,JU\}} \min_{\Gamma_{10|x}} \left[\sum_{j=1}^{jj} \pi(0,j)\,\Gamma_{10|x}(0,1,j) + \sum_{j=jj}^{J} \pi(1,j)\,\Gamma_{10|x}(1,1,j)\right] \quad (13)$$

s.t. (i) $\Gamma_{10|x}(0,1,j) \geq 0$ for $j = 1,\ldots,jj$ and $\Gamma_{10|x}(1,1,j) \geq 0$ for $j = jj,\ldots,J$;

(ii) $\Gamma_{10|x}(0,1,j) \leq \Gamma_{0|x}(j)$ for $j = 1,\ldots,jj$;

(iii) $\Gamma_{10|x}(0,1,jj) + \Gamma_{10|x}(1,1,jj) \leq \Gamma_{0|x}(jj)$;

(iv) $\Gamma_{10|x}(1,1,j) \leq \Gamma_{0|x}(j)$ for $j = jj,\ldots,J$;

(v) $\sum_{j=1}^{jj} \Gamma_{10|x}(0,1,j) = \Gamma_{1|x}(0,1)$ and $\sum_{j=jj}^{J} \Gamma_{10|x}(1,1,j) = \Gamma_{1|x}(1,1)$.

Steps 2 and 3 of DREAM solve the inner and outer minimization problems in (13) sequentially. Once we obtain the solution to the inner optimization problem for any given jj, solving for the outer one is trivial. In the following, we give a succinct discussion of Steps 2 and 3 for solving the inner optimization problem for a given $jj \in \{JL, JL+1, ..., JU\}$ and refer interested reader to Appendix B for detailed explanations.

Given jj, directly solving the inner minimization problem in (13) can still be complicated due to its third constraint (iii). It is the only constraint in (13) that involves both $\Gamma_{10|x}(0,1,j)$ and $\Gamma_{10|x}(1,1,j)$ for the same j. So we first ignore this constraint and solve the following simpler optimization problem:

$$\min_{\Gamma_{10|x}} \left[\sum_{j=1}^{jj} \pi(0,j)\,\Gamma_{10|x}(0,1,j) + \sum_{j=jj}^{J} \pi(1,j)\,\Gamma_{10|x}(1,1,j)\right] \quad (14)$$

s.t. (i), (ii), (iv), and (v) in (13) hold.

Since the objective function and the constraints in (14) can be separated in two parts,

(14) is equivalent to the following two minimization problems:

$$\textbf{Problem 1:} \quad \min_{\Gamma_{10|x}} \sum_{j=1}^{jj} \pi(0,j)\,\Gamma_{10|x}(0,1,j)$$

$$\text{s.t.} \quad \text{(i)} \ \Gamma_{10|x}(0,1,j) \geq 0 \text{ for } j = 1,\ldots,jj,$$

$$\text{(ii)} \ \Gamma_{10|x}(0,1,j) \leq \Gamma_{0|x}(j) \text{ for } j = 1,\ldots,jj, \text{ and}$$

$$\text{(iii)} \ \sum_{j=1}^{jj} \Gamma_{10|x}(0,1,j) = \Gamma_{1|x}(0,1);$$

$$\textbf{Problem 2:} \quad \min_{\Gamma_{10|x}} \sum_{j=jj}^{J} \pi(1,j)\,\Gamma_{10|x}(1,1,j)$$

$$\text{s.t.} \quad \text{(i)} \ \Gamma_{10|x}(1,1,j) \geq 0 \text{ for } j = jj,\ldots,J,$$

$$\text{(ii)} \ \Gamma_{10|x}(1,1,j) \leq \Gamma_{0|x}(j) \text{ for } j = jj,\ldots,J, \text{ and}$$

$$\text{(iii)} \ \sum_{j=jj}^{J} \Gamma_{10|x}(1,1,j) = \Gamma_{1|x}(1,1).$$

Problem 1 and **Problem 2** are solved during Step 2 of DREAM. First we assign mass to $\Gamma_{10|x}(0,1,j)$ based on the rank of $\pi(0,j)$ to solve **Problem 1** and then assign mass to $\Gamma_{10|x}(1,1,j)$ based on the rank of $\pi(1,j)$ to solve **Problem 2**. If the third constraint $\sum_{i=0}^{1} \Gamma_{10|x}(i,1,jj) \leq \Gamma_{0|x}(jj)$ in (13) is satisfied automatically by the solution to (14), then we obtain the solution to (13). If it is violated, then we proceed to Step 3 of DREAM.

In Step 3, we move mass from $\Gamma_{10|x}(0,1,jj)$ to other $\Gamma_{10|x}(0,1,j)$ and/or mass from $\Gamma_{10|x}(1,1,jj)$ to other $\Gamma_{10|x}(1,1,j)$ until the third constraint in (13) is satisfied. Given the values of $\Gamma_{10|x}(i,1,j)$, we can compute

$$c(jj) = \sum_{j=1}^{J} \sum_{i=0}^{1} \pi(i,j)\,\Gamma_{10|x}(i,1,j)$$

for the given jj.

Finally, we repeat Steps 2 and 3 for each $jj \in \{JL, JL+1, \ldots, JU\}$. Then the optimal value of the outer minimization problem in (13) or equivalently of (12) equals to $\min\{c(JL), c(JL+1), \ldots, c(JU)\}$.

6.3 Time Complexity

In this section, we compute the time complexity of our algorithm. In practice, we construct Δ_{TPRD} via three steps. First, we sample N_q vectors $q_1, \ldots, q_{N_q}$ uniformly from the $2J$-dimensional unit sphere and N_θ vectors $\theta_1, \ldots, \theta_{N_\theta}$ uniformly from $[0,1]^{2J}$. Then, for each θ, we compute a set

$$\widehat{\Theta}_I = \left\{\theta = \theta_1, \ldots, \theta_{N_\theta} : q^\top \theta \leq h_{\Theta_I}(q) \text{ for all } q = q_1, \ldots, q_{N_q}\right\}.$$

Finally, we map from each element in $\widehat{\Theta}_I$ to obtain

$$\widehat{\Delta}_{TPRD} = \left\{\delta = G(\theta) : \theta \in \widehat{\Theta}_I\right\}.$$

Assume that a grid-based method is used for the numerical integration when computing $h_{\Theta_I}(q)$ for any given q. During the construction of $\widehat{\Delta}_{TPRD}$, all the operations are basic except computing the optimal partial transport cost $\mathcal{KT}_{q^\top m_2}(\mu_{1|x}, \mu_{0|x}; x)$. The following lemma establishes that DREAM computes $\mathcal{KT}_{q^\top m_2}$ with remarkable efficiency.

Lemma 6.2. *The time complexity for computing $\mathcal{KT}_{q^\top m_2}(\mu_{1|x}, \mu_{0|x}; x)$ is approximately $\frac{1}{2}(3J^3 + 38J^2 + 9J)$.*

Let N_x denote the total number of grid points used during the numerical integration. Then we need $\frac{1}{2}(3J^3 + 38J^2 + 9J) N_x$ basic operations for computing $h_{\Theta_I}(q)$ for one given q. As a result, constructing $\widehat{\Theta}_I$ requires $\frac{1}{2}(3J^3 + 38J^2 + 9J) N_x N_q + N_q N_\theta$ number of operations. By definition, for any θ, $G(\theta)$ takes $5K$ operations. Thus, in total, it requires at most

$$\frac{1}{2}\left(3J^3 + 38J^2 + 9J\right) N_x N_q + N_q N_\theta + 5K N_\theta$$

basic operations to construct $\widehat{\Delta}_{TPRD}$. It can be seen that our efficient DREAM is essential for such a low time complexity.

6.4 Comparison with Kallus et al. (2022)

Differently from our approach, Kallus et al. (2022) study the convex hull of the identified set for a collection of TPRD measures instead of the identified set itself. They focus on $\delta_{TPRD} \equiv [\delta_{TPRD}(1, J), \ldots, \delta_{TPRD}(J-1, J)]$, which is one example of

our TPRD measures with $G^{\star}(\theta^{*}) \equiv [g_{1,J}(\theta^{*}), \ldots, g_{J-1,J}(\theta^{*})]$. Kallus et al. (2022) propose to compute the support function $h_{\text{conv}}(p)$ of the convex hull at direction $p \equiv (p_1, \ldots, p_{J-1})$ from the maximization problem (15) below.

Lemma 6.3 (Proposition 11 in Kallus et al. (2022)). *For any $p \in \mathbb{S}^{J-1}$, $h_{conv}(p)$ is the value of the following maximization problem*

$$\max_{\mathfrak{p} \in \mathbb{R}^{J}: \mathfrak{p} \geq 1} \max_{P_1(\cdot), \ldots, P_J(\cdot)} \sum_{j=1}^{J-1} p_j \left[\mathbb{E}\left[P_j\left(Y_{1s}, Y_{1r}, X\right) Y_{1s} Y_{1r} \right] - \mathbb{E}\left[P_J\left(Y_{1s}, Y_{1r}, X\right) Y_{1s} Y_{1r} \right] \right] \tag{15}$$

s.t. for any $j = \{1, \ldots, J\}$, $y_{1s} \in \{0,1\}$, $y_{1r} \in \{0,1\}$ and $x \in \mathcal{X}$:

(i) $\mathbb{E}\left[P_j\left(Y_{1s}, Y_{1r}, X\right) Y_{1r} \right] = 1$; *(ii)* $\sum_{j=1}^{J} \dfrac{P_j\left(y_{1s}, y_{1r}, x\right)}{\mathfrak{p}_j} = 1$; *(iii)* $P_j\left(y_{1s}, y_{1r}, x\right) \geq 0$; *and*

(iv) $\sum_{y_{1s}, y_{1r} \in \{0,1\}} P_j\left(y_{1s}, y_{1r}, x\right) \Pr\left(Y_{1s} = y_{1s}, Y_{1r} = y_{1r} \mid X = x\right) = \Pr\left(Y_0 = a_j \mid X = x\right) \mathfrak{p}_j$.

Algorithm 2 in Kallus et al. (2022) and the following-up discussion provide that the approximate of the convex hull of $\Delta^{\star}_{TPRD}$ is computed as

$$\widehat{\text{Conv}}\left(\Delta^{\star}_{TPRD}\right) = \left\{ \delta = \delta_1, \ldots, \delta_{N_\delta} : p^{\top} \delta \leq h_{\text{conv}}(p) \text{ for all } p = p_1, \ldots, p_{N_p} \right\}.$$

Assume that a grid-based method is used for computing the expectation in (15), then the inner maximization in (15) is a linear programming with $4JN_x$ number of variables, where N_x is the number of grid points. As a result, the inner maximization itself takes at least $\widetilde{O}\left((4JN_x)^{2.37}\right)$ number of operations. Moreover, the outer maximization is a non-convex optimization problem which is known to be NP-hard. Thus, we expect our procedure to be computationally much more efficient than the one in Kallus et al. (2022).

7 Concluding Remarks

In this paper, we have developed a unified approach via optimal transport to characterize the identified set for a finite-dimensional parameter in a moment equality model with incomplete data. Using three running examples, we have demonstrated the advantages and simplicity of our approach compared with works in the current

literature. First, for Example 1, we constructed the identified set allowing Y_0 to be of any finite dimension while existing work established the identified set for univariate Y_0 only; Second, for Example 2, we solved the existing linear programming characterization in closed form; Third, for Example 3, we developed an effective algorithm, DREAM, for evaluating the identified set for any collection of TPRD measures and established its time complexity. In contrast, the existing work requires solving a non-convex optimization problem and only constructs an outer set of the identified set for more than two TPRD measures.

In a follow-up paper, we will study estimation and inference for θ^* and its known transformations. Given the inequality characterization of the identified set for θ^* in (16), plug-in estimation of its identified set and that of a known function of θ^* is straightforward. Inference is complicated by the presence of unknown functions/marginal distributions in the continuum of moment inequalities. However, methods developed for partially identified/incomplete models in the literature can be extended to account for the first step estimation of the unknown nuisance functions such as Andrews and Shi (2013) and Chernozhukov et al. (2013) for inference on the *entire vector* θ^* and Bei (2024) for *subvector* inference. For models satisfying Assumption 3.2, more computationally efficient methods may be developed by taking into account the convexity of the identified set especially the closed-form expression of its support function, see Beresteanu and Molinari (2008), Beresteanu et al. (2011), Bontemps et al. (2012), and Kaido (2016) for confidence sets of the identified set and Kaido et al. (2019) for *subvector* inference.

Finally, we focused on parametric moment condition models in this paper. However, our general identification result also applies to semiparametric moment condition models with both finite-dimensional and infinite-dimensional parameters. It would be interesting to work out some specific examples. This is left for future work.

A Technical Proofs

Lemma A.1. *Consider Θ_I defined in (2). Define*

$$\Theta_o \equiv \left\{ \theta \in \Theta : \inf_{\mu \in \mathcal{M}(\mu_{1X}, \mu_{0X})} \mathbb{E}_\mu \left[t^\top m \left(Y_1, Y_0, X; \theta \right) \right] \leq 0 \text{ for all } t \in \mathbb{S}^k \right\}. \quad (16)$$

(i) It holds that $\Theta_I \subseteq \Theta_o$. (ii) Conversely, if $\theta_o \in \Theta_o$, then there exists a sequence of measures $\mu^{(s)} \in \mathcal{M}\left(\mu_{1X}, \mu_{0X}\right)$ such that $\mathbb{E}_{\mu^{(s)}} \left[m \left(Y_1, Y_0, X; \theta_o \right) \right] \to \mathbf{0}$ as $s \to \infty$.

Proof of Lemma A.1: Part (i). By the definition of Θ_I in (2), for any $\theta_I \in \Theta_I$, $\mathbb{E}_\mu \left[t^\top m \left(Y_1, Y_0, X; \theta_I \right) \right] = \mathbf{0}$ for some $\mu \in \mathcal{M}\left(\mu_{1X}, \mu_{0X}\right)$ and for all $t \in \mathbb{S}^k$. Then it must be true that

$$\inf_{\mu \in \mathcal{M}(\mu_{1X}, \mu_{0X})} \mathbb{E}_\mu \left[t^\top m \left(Y_1, Y_0, X; \theta_I \right) \right] \leq 0$$

for all $t \in \mathbb{S}^k$. Hence we have $\theta_I \in \Theta_o$ and $\Theta_I \subseteq \Theta_o$.

Part (ii). Because the left-hand side of the inequality in the definition (16) is positive homogeneous in t, Θ_o is equivalent to

$$\Theta_o = \left\{ \theta \in \Theta : \begin{array}{c} \inf_{\mu \in \mathcal{M}(\mu_{1X}, \mu_{0X})} \mathbb{E}_\mu \left[t^\top m \left(Y_1, Y_0, X; \theta \right) \right] \leq 0 \\ \text{for all } t \in \mathbb{R}^k \text{ s.t. } \|t\| \leq 1 \end{array} \right\}.$$

If $\theta_o \in \Theta_o$, then we have that for all $t \in \mathbb{R}^k$ such that $\|t\| \leq 1$,

$$\inf_{\mu \in \mathcal{M}(\mu_{1X}, \mu_{0X})} \mathbb{E}_\mu \left[t^\top m \left(Y_1, Y_0, X; \theta_o \right) \right] \leq 0.$$

Taking the supremum over all t with $\|t\| \leq 1$, we get

$$\sup_{t \in \mathbb{R}^k, \|t\| \leq 1} \inf_{\mu \in \mathcal{M}(\mu_{1X}, \mu_{0X})} \mathbb{E}_\mu \left[t^\top m \left(Y_1, Y_0, X; \theta_o \right) \right] \leq 0.$$

The idea is now to apply a minimax theorem to the left-hand side of the above inequality. The result in Vianney and Vigeral (2015) is sufficiently general. Because the function $(t, \mu) \mapsto \mathbb{E}_\mu \left[t^\top m \left(Y_1, Y_0, X; \theta_o \right) \right]$ is bilinear, it is concave in t and convex in μ. The domains of both t and μ are convex, and t is finite-dimensional and bounded. For $t = \mathbf{0}$, the function is identically 0 and therefore bounded below.

Therefore, Theorem 1 in Vianney and Vigeral (2015) applies and we obtain that

$$\begin{aligned}&\sup_{t\in\mathbb{R}^k,\|t\|\leq 1}\inf_{\mu\in\mathcal{M}(\mu_{1X},\mu_{0X})}\mathbb{E}_\mu\left[t^\top m\left(Y_1,Y_0,X;\theta_o\right)\right]\\=&\inf_{\mu\in\mathcal{M}(\mu_{1X},\mu_{0X})}\sup_{t\in\mathbb{R}^k,\|t\|\leq 1}\mathbb{E}_\mu\left[t^\top m\left(Y_1,Y_0,X;\theta_o\right)\right]\leq 0.\end{aligned}$$

For any $\mu\in\mathcal{M}\left(\mu_{1X},\mu_{0X}\right)$, if $\left\|\mathbb{E}_\mu\left[m\left(Y_1,Y_0,X;\theta_o\right)\right]\right\|\neq 0$, then the supremum of the optimization problem $\sup_{t\in\mathbb{R}^k,\|t\|\leq 1}\mathbb{E}_\mu\left[t^\top m\left(Y_1,Y_0,X;\theta_o\right)\right]$ is attained at

$$t=\left\|\mathbb{E}_\mu\left[m\left(Y_1,Y_0,X;\theta_o\right)\right]\right\|^{-1}\mathbb{E}_\mu\left[m\left(Y_1,Y_0,X;\theta_o\right)\right].$$

If $\left\|\mathbb{E}_\mu\left[m\left(Y_1,Y_0,X;\theta_o\right)\right]\right\|=0$, then the supremum is achieved at any $t\in\mathbb{R}^k$ for $\|t\|\leq 1$. In both cases, the maximal value equals to $\left\|\mathbb{E}_\mu\left[m\left(Y_1,Y_0,X;\theta_o\right)\right]\right\|$. Thus, we have

$$\begin{aligned}0\geq&\inf_{\mu\in\mathcal{M}(\mu_{1X},\mu_{0X})}\sup_{t\in\mathbb{R}^k,\|t\|\leq 1}\mathbb{E}_\mu\left[t^\top m\left(Y_1,Y_0,X;\theta_o\right)\right]\\=&\inf_{\mu\in\mathcal{M}(\mu_{1X},\mu_{0X})}\left\|\mathbb{E}_\mu\left[m\left(Y_1,Y_0,X;\theta_o\right)\right]\right\|\geq 0,\end{aligned}$$

which implies that $\inf_{\mu\in\mathcal{M}(\mu_{1X},\mu_{0X})}\left\|\mathbb{E}_\mu\left[m\left(Y_1,Y_0,X;\theta_o\right)\right]\right\|=0$. This completes the proof for the second part of the lemma. □

Lemma A.2. *Consider Θ_o defined in (4). If $\theta_o\in\Theta_o$, then for any $x\in\mathcal{X}$, there exists a sequence of measures $\mu_{10|x}^{(s)}\in\mathcal{M}\left(\mu_{1|x},\mu_{0|x}\right)$ such that as $s\to\infty$,*

$$\int\left[\int\int m\left(y_1,y_0,x;\theta_o\right)d\mu_{10|x}^{(s)}\right]d\mu_X\to\mathbf{0}.$$

Proof of Lemma A.2: By Lemma A.1 (ii), there exists a sequence of measures $\mu^{(s)}\in\mathcal{M}\left(\mu_{1X},\mu_{0X}\right)$ such that $\mathbb{E}_{\mu^{(s)}}\left[m\left(Y_1,Y_0,X;\theta_o\right)\right]\to\mathbf{0}$ as $s\to\infty$. For any s, the measure $\mu^{(s)}$ represents a joint distribution of (Y_1,Y_0,X) with μ_{1X} and μ_{0X} being the distributions of (Y_1,X) and (Y_0,X) that do not depend on s. From such a joint distribution of (Y_1,Y_0,X), by the disintegration theorem, we can obtain μ_X as the probability measure of X and $\mu_{10|x}^{(s)}$ for each $x\in\mathcal{X}$ whose projections on Y_1 and Y_0 are $\mu_{1|x}$ and $\mu_{0|x}$. None of the measures μ_X, $\mu_{1|x}$ for all $x\in\mathcal{X}$, or $\mu_{0|x}$ for all $x\in\mathcal{X}$ change with s. Therefore, we obtain the sequence of measures $\mu_{10|x}^{(s)}\in\mathcal{M}\left(\mu_{1|x},\mu_{0|x}\right)$ for each $x\in\mathcal{X}$ that satisfies the condition in the lemma. □

Proof of Theorem 3.1: By Lemma A.1 (i), it remains to show that for any $\theta_o \in \Theta_o$, it holds that $\theta_o \in \Theta_I$. By Assumption 3.1, given every $\theta \in \Theta$ and almost every $x \in \mathcal{X}$ with respect to μ_X measure, the continuity of $m(y_1, y_0, x; \theta_o)$ with respect to (y_1, y_0) makes the functional

$$\mu_{10|x} \mapsto \int \int m(y_1, y_0, x; \theta_o) \, d\mu_{10|x}(y_1, y_0)$$

continuous with respect to weak convergence of measures. By Proposition 9 in Fan et al. (2023a), $\mathcal{M}(\mu_{1|x}, \mu_{0|x})$ is compact with respect to the weak convergence. In consequence, for almost every $x \in \mathcal{X}$ with respect to μ_X measure, the sequence of measures $\mu^{(k)}_{10|x}$ in Lemma A.2 has a subsequence converging to some $\mu^{\star}_{10|x} \in \mathcal{M}(\mu_{1|x}, \mu_{0|x})$. Thus, there exists $\mu^{\star}_{10|x} \in \mathcal{M}(\mu_{1|x}, \mu_{0|x})$ for almost every $x \in \mathcal{X}$ such that

$$\int \left[\int \int m(y_1, y_0, x; \theta_o) \, d\mu^{\star}_{10|x} \right] d\mu_X = \mathbf{0}.$$

Hence, $\theta_o \in \Theta_I$ and the theorem holds. □

Proof of Proposition 3.1: Under Assumption 3.1, Theorem 3.1 provides that $\Theta_I = \Theta_o$. It suffices to show that for any $\theta \in \Theta_o$, it holds that $\theta \in \Theta^O$. The weak compactness of $\mathcal{M}(\mu_{1|x}, \mu_{0|x})$ shown in Fan et al. (2023a) provides that for any $t \in \mathbb{S}^k$ and $\theta \in \Theta$,

$$\int \mathcal{KT}_{t^\top m}(\mu_{1|x}, \mu_{0|x}; x, \theta) \, d\mu_X = \int \left[\int \int t^\top m(y_1, y_0; \theta) \, d\mu^{\dagger}_{10|x} \right] d\mu_X$$

for some $\mu^{\dagger}_{10|x} \in \mathcal{M}(\mu_{1|x}, \mu_{0|x})$ for each $x \in \mathcal{X}$. By the definition of conditional probability, we have that

$$\int \left[\int \int t^\top m(y_1, y_0; \theta) \, d\mu^{\dagger}_{10|x} \right] d\mu_X = \int \int t^\top m(y_1, y_0; \theta) \, d\mu^{\dagger}_{10},$$

where $\mu^{\dagger}_{10}(y_1, y_0) \equiv \int_{x \in \mathcal{X}} \mu^{\dagger}_{10|x}(y_1, y_0) \, d\mu_X(x)$. Because $\mu_{1|x}$ and $\mu_{0|x}$ are the conditional distributions of Y_1 on $X = x$ and Y_0 on $X = x$ respectively by definition and μ_X is the probability measure of X, it holds that $\mu^{\dagger}_{10}(y_1, y_0) \in \mathcal{M}(\mu_1, \mu_0)$. Thus, we

have that for any $t \in \mathbb{S}^k$ and $\theta \in \Theta$,

$$\begin{aligned}\int \mathcal{KT}_{t^\top m}\left(\mu_{1|x}, \mu_{0|x}; x, \theta\right) d\mu_X &= \int\int t^\top m\left(y_1, y_0; \theta\right) d\mu_{10}^\dagger \\ &\geq \inf_{\mu_{10} \in \mathcal{M}(\mu_1, \mu_0)} \int\int t^\top m\left(y_1, y_0; \theta\right) d\mu_{10}.\end{aligned}$$

In consequence, if θ satisfies the inequality conditions in the definition of Θ_o for each $t \in \mathbb{S}^k$, then it would also satisfy the inequality conditions in the definition of Θ^O. Hence, $\Theta_I \subseteq \Theta^O$. □

Proof of Proposition 3.2: Because $\mu_{0|x}$ is Dirac at $g_0(x)$, we can compute the optimal transport cost in the definition of Θ_o as

$$\begin{aligned}&\int \left[\inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int\int t^\top m\left(y_1, y_0, x; \theta\right) d\mu_{10|x}\right] d\mu_X \\ &= \int \left[\int t^\top m\left(y_1, g_0(x), x; \theta\right) d\mu_{1|x}(y_1)\right] d\mu_X(x) \\ &= \int\int t^\top m\left(y_1, g_0(x), x; \theta\right) d\mu_{1X}(y_1, x) \\ &= \mathbb{E}\left[t^\top m\left(Y_1, g_0(X), X; \theta\right)\right],\end{aligned}$$

where the second equality follows from the definition of $\mu_{1|x}$ and μ_X. Denote the set in the proposition as Θ^P. Since $\Theta_I = \Theta_o$ by Theorem 3.1, in the following, we show that $\Theta_o \subseteq \Theta^P$ and $\Theta^P \subseteq \Theta_o$.

If $\theta_o \in \Theta_o$, then for any $t \in \mathbb{S}^k$ we have that

$$\mathbb{E}\left[t^\top m\left(Y_1, g_0(X), X; \theta_o\right)\right] \leq 0.$$

If $t \in \mathbb{S}^k$, then $-t \in \mathbb{S}^k$ holds as well. In consequence, we have that for any $t \in \mathbb{S}^k$,

$$0 \geq \mathbb{E}\left[t^\top m\left(Y_1, g_0(X), X; \theta_o\right)\right] = -\mathbb{E}\left\{-t^\top m\left(Y_1, g_0(X), X; \theta_o\right)\right\} \geq 0.$$

This shows that $\mathbb{E}\left[t^\top m\left(Y_1, g_0(X), X; \theta_o\right)\right] = 0$ for any $t \in \mathbb{S}^k$, which implies that

$$\mathbb{E}\left[m\left(Y_1, g_0(X), X; \theta_o\right)\right] = \mathbf{0}.$$

Thus, we have that $\theta_o \in \Theta^P$. The other side $\Theta^P \subseteq \Theta_o$ is straightforward. Hence, the proposition holds. □

Proof of Theorem 3.2: Theorem 3.1 implies that $\Theta_o = \Theta_I$. Plugging the first decomposition of m provided by Assumption 3.2, we obtain that

$$\begin{aligned}\int\int t^\top m\left(y_1, y_0, x; \theta\right) d\mu_{10|x} &= \int\int t^\top m_1\left(y_0, x\right) \theta d\mu_{10|x} + \int\int t^\top m_2\left(y_1, y_0, x\right) d\mu_{10|x} \\ &= t^\top \int\int m_1\left(y_0, x\right) \theta d\mu_{0|x} + \int\int t^\top m_2\left(y_1, y_0, x\right) d\mu_{10|x} \\ &= t^\top \mathbb{E}\left[m_1\left(Y_0, X\right) \mid X\right] \theta + \int\int t^\top m_2\left(y_1, y_0, x\right) d\mu_{10|x},\end{aligned}$$

where the second equality follows from Assumption 2.1. Thus, it holds that for any $t \in \mathbb{S}^k$,

$$\begin{aligned}&\int \mathcal{KT}_{t^\top m}\left(\mu_{1|x}, \mu_{0|x}; x, \theta\right) d\mu_X \leq 0 \\ \iff &\int t^\top \mathbb{E}\left[m_1\left(Y_0, X\right) \mid X\right] \theta d\mu_X + \int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \leq 0 \\ \iff & t^\top \mathbb{E}\left[m_1\left(Y_0, X\right)\right] \theta + \int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \leq 0 \\ \iff & t^\top \mathbb{E}\left[m_1\left(Y_0, X\right)\right] \theta \leq -\int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X.\end{aligned}$$

Thus, we show that each inequality in (4) can be alternatively expressed as the inequality in part (i) of the theorem.

To prove that Θ_I is closed, assume the contrary. Then there exists a sequence $\{\theta_n\}_{n=1}^\infty$ and its limit $\theta^\dagger$ such that $\theta_n \in \Theta_I$ for all n, but $\theta^\dagger \notin \Theta_I$. Then, there is at least one $t^\dagger \in \mathbb{S}^k$ and some $\delta > 0$ such that

$$t^{\dagger\top} \mathbb{E}\left[m_1\left(Y_0, X\right)\right] \theta^\dagger > -\int \mathcal{KT}_{t^{\dagger\top} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X.$$

Because $\left\|\theta_n - \theta^\dagger\right\| \to 0$ as $n \to \infty$, there must be some N and some $\delta > 0$ such that

$$\begin{aligned}t^{\dagger\top} \mathbb{E}\left[m_1\left(Y_0, X\right)\right] \theta_N &> t^{\dagger\top} \mathbb{E}\left[m_1\left(Y_0, X\right)\right] \theta^\dagger - \delta \\ &> -\int \mathcal{KT}_{t^{\dagger\top} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X.\end{aligned}$$

This implies that $\theta_N \notin \Theta_I$, which contradicts that $\theta_n \in \Theta_I$ for all n. Therefore, Θ_I is closed. The convexity of Θ_I holds because the constraints in the expression of Θ_I are affine in θ and Θ is convex by Assumption 3.3. Thus, part (ii) of the theorem holds. □

Lemma A.3. *Under Assumptions 3.1, 3.2, and 3.5, the identified set Θ_I in Theorem 3.1 (i) can be rewritten as* $\left\{ \theta \in \Theta : q^\top \theta \leq s(q) \text{ for all } q \in \mathbb{S}^{d_\theta} \right\}$, *where*

$$s(q) \equiv -\int \mathcal{KT}_{q^\top \mathbb{E}[m_1(Y_0,X)]^{-1} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X.$$

Proof of Lemma A.3: Denote the set in the lemma as $\Theta_I^\dagger$. We aim to show that $\Theta_I^\dagger \subseteq \Theta_I$ and $\Theta_I \subseteq \Theta_I^\dagger$. Under Assumption 3.5, matrix $\mathbb{E}[m_1(Y_0, X)]^{-1}$ exists. For any $t \in \mathbb{S}^k$, let $q = \mathbb{E}[m_1(Y_0, X)]^\top t$. Then for any $t \in \mathbb{S}^k$, we have

$$\begin{aligned}\int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X &= \int \mathcal{KT}_{t^\top \mathbb{E}[m_1(Y_0,X)]\mathbb{E}[m_1(Y_0,X)]^{-1} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \\ &= \int \mathcal{KT}_{q^\top \mathbb{E}[m_1(Y_0,X)]^{-1} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X.\end{aligned}$$

In consequence, for any $t \in \mathbb{S}^k$, if $\theta \in \Theta$ satisfies the inequality constraint

$$t^\top \mathbb{E}[m_1(Y_0, X)]\theta \leq -\int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X,$$

then it would also satisfy

$$q^\top \theta \leq -\int \mathcal{KT}_{q^\top \mathbb{E}[m_1(Y_0,X)]^{-1} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X,$$

where q does not necessarily belong to $\mathbb{S}^{d_\theta}$. On the other hand, because the constraint is positive homogeneous in q, it is equivalent to letting $\|q\| = 1$. Thus, we have shown that for any constraint in Θ_I, there is a corresponding constraint in $\Theta_I^\dagger$. It holds that $\Theta_I^\dagger \subseteq \Theta_I$. The other direction $\Theta_I \subseteq \Theta_I^\dagger$ follows from a similar argument by letting $t = \left[\mathbb{E}[m_1(Y_0, X)]^{-1}\right]^\top q$. Hence, $\Theta_I = \Theta_I^\dagger$. □

Lemma A.4. *Under Assumptions 3.1-3.5, for any $q \in \mathbb{S}^{d_\theta}$ there exists some $\theta \in \Theta_I$ such that*

$$-\int \mathcal{KT}_{q^\top \mathbb{E}[m_1(Y_0,X)]^{-1} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X = q^\top \theta. \tag{17}$$

Proof of Lemma A.4: Under Assumption 3.5, $\mathbb{E}\left[m_1\left(Y_0, X\right)\right]^{-1}$ exists. By Assumption 3.1, the proof of Theorem 3.1 shows that for any $q \in \mathbb{S}^{d_\theta}$,

$$\begin{aligned}
&\int \mathcal{KT}_{q^\top \mathbb{E}[m_1(Y_0,X)]^{-1} m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \\
&= \int \left[\int\int q^\top \mathbb{E}\left[m_1\left(Y_0, X\right)\right]^{-1} m_2\left(y_1, y_0, x\right) d\mu^\dagger_{10|x}\right] d\mu_X \\
&= q^\top \mathbb{E}\left[m_1\left(Y_0, X\right)\right]^{-1} \int \left[\int\int m_2\left(y_1, y_0, x\right) d\mu^\dagger_{10|x}\right] d\mu_X,
\end{aligned}$$

for some $\mu^\dagger_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)$ for every $x \in \mathcal{X}$. Let

$$\theta^\dagger = -\mathbb{E}\left[m_1\left(Y_0, X\right)\right]^{-1} \int \left[\int\int m_2\left(y_1, y_0, x\right) d\mu^\dagger_{10|x}\right] d\mu_X.$$

Then for such $\theta^\dagger$, Equality (17) holds. It remains to show that $\theta^\dagger \in \Theta_I$.

Note that $\theta^\dagger$ satisfies the following system of linear equations

$$\mathbb{E}\left[m_1\left(Y_0, X\right)\right] \theta^\dagger + \int \left[\int\int m_2\left(y_1, y_0, x\right) d\mu^\dagger_{10|x}\right] d\mu_X = \mathbf{0}.$$

By Assumptions 3.2, the proof of Theorem 3.1 implies that

$$\begin{aligned}
&\mathbb{E}\left[m_1\left(Y_0, X\right)\right] \theta^\dagger + \int \left[\int\int m_2\left(y_1, y_0, x\right) d\mu^\dagger_{10|x}\right] d\mu_X \\
&= \int \left[\int\int m\left(y_1, y_0, x; \theta^\dagger\right) d\mu^\dagger_{10|x}\right] d\mu_X = \mathbf{0}.
\end{aligned}$$

Since Θ_I is in the interior of Θ by Assumption 3.4, Θ_I is equivalent to

$$\Theta_I = \left\{ \theta \in \mathbb{R}^{d_\theta} : \begin{array}{c} \int \left[\int\int m\left(y_1, y_0, x; \theta\right) d\mu_{10|x}\left(y_1, y_0\right)\right] d\mu_X\left(x\right) = \mathbf{0} \\ \text{for some } \mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right) \text{ for every } x \in \mathcal{X} \end{array} \right\}.$$

Because for every $x \in \mathcal{X}$, $\mu^\dagger_{10|x}$ belongs to $\mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)$, it holds that $\theta^\dagger \in \Theta_I$, which completes the proof. □

Proof of Proposition 3.3: The definition of the support function states that $h_{\Theta_I}\left(q\right) = \sup_{\theta \in \Theta} q^\top \theta$. For any given $q \in \mathbb{S}^{d_\theta}$, the expression of Θ_I in Lemma A.3 implies that $q^\top \theta \leq s\left(q\right)$ for all $\theta \in \Theta_I$. Therefore, $h_{\Theta_I}\left(q\right) \leq s\left(q\right)$. As the same time, Lemma A.4 states that there exists some $\theta^\dagger \in \Theta_I$, such that $q^\top \theta^\dagger = s\left(q\right)$. This implies

that $h_{\Theta_I}(q) \geq s(q)$. Combining the two results, we have that $h_{\Theta_I}(q) = s(q)$. □

Lemma A.5. *Let V and W be two univariate random variables with probability measures μ_V and μ_W respectively. It holds that*

$$\inf_{\mu_{V,W} \in \mathcal{M}(\mu_V, \mu_W)} \int \int -vw d\mu_{V,W}(v,w) = -\int_0^1 F_V^{-1}(u) F_W^{-1}(u) du.$$

Proof of Lemma A.5: The proof follows directly from Proposition 2.17 in Santambrogio (2015) by letting the cost function $h(v-w)$ be $(v-w)^2$. □

Proof of Proposition 4.1: Both Assumptions 3.1 and 3.2 hold by the nature of the moment function. Thus Theorem 3.2 implies that

$$\Theta_I = \left\{ \theta \in \Theta : t^\top M\theta \leq -\int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X \text{ for all } t \in \mathbb{S}^k \right\}.$$

It remains to compute $\int \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) d\mu_X$. By definition, we have

$$\begin{aligned}
\mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) &= \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int \int -t^\top \begin{pmatrix} y_0 \\ x \end{pmatrix} y_1 d\mu_{10|x} \\
&= \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \left[\int \int -\left(t_0^\top y_0\right) y_1 d\mu_{10|x} - \int \int t_X^\top y_1 x d\mu_{10|x} \right] \\
&= \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int \int -\left(t_0^\top y_0\right) y_1 d\mu_{10|x} - \int \int t_X^\top y_1 x d\mu_{1|x}.
\end{aligned}$$

Apply Lemma A.5, we obtain that

$$\inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int \int -\left(t_0^\top y_0\right) y_1 d\mu_{10|x} = -\int F^{-1}_{t_0^\top Y_0|x}(u) F^{-1}_{Y_1|x}(u) du.$$

Moreover, since

$$\int \int t_X^\top y_1 x d\mu_{1|x} d\mu_X = t_X^\top \mathbb{E}\left(Y_1 X\right),$$

we obtain the expression for the right-hand side of the inequality in the proposition. □

Proof of Corollary 4.1: For an easy comparison with the result in Pacini (2019), we adopt the notation in Pacini (2019) in this proof. The support function can be

rewritten as

$$\int \int_0^1 F^{-1}_{\left(q_\alpha A + q_\beta^\top C\right) X|z}(u) F^{-1}_{Y|z}(u) \, du d\mu_Z + \left(q_\alpha B + q_\beta^\top D\right) s_{zy}$$

Using the definitions in Pacini (2019), we have

$$\begin{aligned} q_\alpha A + q_\beta^\top C = q_\alpha b_o + q_\beta^\top d_o = e_{oq} \text{ and} \\ \left(q_\alpha B + q_\beta^\top D\right) s_{zy} = q_\alpha a_o + q_\beta^\top c_o = v_{oq}. \end{aligned}$$

As a result, the support function can be further written as

$$\int_0^1 F^{-1}_{e_{oq} X|z}(u) F^{-1}_{Y|z}(u) \, du d\mu_Z + v_{oq}. \tag{18}$$

Depending on the sign of e_{oq}, it holds that

$$\begin{aligned} F^{-1}_{e_{oq} X|z}(u) = e_{oq} F^{-1}_{X|z}(u) \text{ for } e_{oq} > 0 \text{ and} \\ F^{-1}_{e_{oq} X|z}(u) = e_{oq} F^{-1}_{X|z}(1-u) \text{ for } e_{oq} \leq 0. \end{aligned}$$

Therefore, the left hand side of (18) is equivalent to

$$\begin{aligned} & e_{oq} I(e_{oq} > 0) \int_0^1 F^{-1}_{X|z}(u) F^{-1}_{Y|z}(u) \, du d\mu_Z \\ & + e_{oq} I(e_{oq} \leq 0) \int_0^1 F^{-1}_{X|z}(1-u) F^{-1}_{Y|z}(u) \, du d\mu_Z + v_{oq}. \end{aligned}$$

Applying the change-of-variable, we have that

$$\begin{aligned} \int_0^1 F^{-1}_{X|z}(u) F^{-1}_{Y|z}(u) \, du d\mu_Z = \mathbb{E}\left[Y F^{-1}_{X|Z}\left(F_{Y|Z}(Y)\right)\right] \equiv \lambda^o_{Fu} \text{ and} \\ \int_0^1 F^{-1}_{X|z}(1-u) F^{-1}_{Y|z}(u) \, du d\mu_Z = \mathbb{E}\left[Y F^{-1}_{X|Z}\left(1 - F_{Y|Z}(Y)\right)\right] \equiv \lambda^o_{Fl} \end{aligned}$$

using the definitions of λ^o_{Fu} and λ^o_{Fl} in Pacini (2019). In addition, by the last statement in the proof of Lemma 3 in Pacini (2019), it holds that

$$e_{oq} I(e_{oq} > 0) \lambda^o_{Fu} + e_{oq} I(e_{oq} \leq 0) \lambda^o_{Fl} = \max\left\{ e_{oq} \lambda^o_{Fl}, e_{oq} \lambda^o_{Fu} \right\}.$$

Thus, the support function (18) can be expressed as

$$\max\left\{ e_{oq} \lambda^o_{Fl}, e_{oq} \lambda^o_{Fu} \right\} + v_{oq}.$$

This is the same as the support function for Θ_F in Pacini (2019).

Simple algebra would show that the support function of Θ_F defined in Lemma 1 in Pacini (2019) is equal to that of the identified set in Proposition 1 in Hwang (2023). Since any non-empty closed convex set is uniquely determined by its support function, the two sets are equal, which in turn shows that Θ_I is equal to the identified set in Hwang (2023). □

Proof of Proposition 5.1: Since Y_1 and Y_0 are both discrete, Assumption 3.1 is satisfied. Additionally, Assumption 3.2 holds with $m_1(y_0, x)$ and $m_2(y_0, y_1, x)$ discussed in Section 5. Thus, Theorem 3.2 (i) applies. We have that

$$\mathbb{E}\left[m_1\left(Y_0, X\right)\right] = \operatorname{diag}\left(\Pr\left(Y_0 = a_1\right), \ldots, \Pr\left(Y_0 = a_J\right)\right).$$

It remains to compute $\mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right)$.

Plugging in the functional form of $m_2(y_0, y_1, x)$, we have that

$$\begin{aligned}
& \mathcal{KT}_{t^\top m_2}\left(\mu_{1|x}, \mu_{0|x}; x\right) \\
= & \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int\int \sum_{j=1}^{J} -t_j I\left\{y_1 = 1, y_0 = a_j\right\} d\mu_{10|x} \\
= & \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int\int -I\left\{y_1 = 1\right\} \left[\sum_{j=1}^{J} t_j I\left\{y_0 = a_j\right\}\right] d\mu_{10|x}.
\end{aligned}$$

Let $d(y_1) = I\{y_1 = 1\}$ and $d_t(y_0) = \sum_{j=1}^{J} t_j I\{y_0 = a_j\}$. Apply Lemma A.5, we have that

$$\inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int\int -d(y_1) \times d_t(y_0)\, d\mu_{10|x}(y_1, y_0) = -\int_0^1 F^{-1}_{D|x}(u)\, F^{-1}_{D_t|x}(u)\, du,$$

where the random variables D and D_t are defined in the proposition. We have obtained each term in Theorem 3.2 (i). The proposition hence follows. □

Proof of Lemma 5.1: We count the number of elementary operations when computing $\int_0^1 F^{-1}_{D|x}(u)\, F^{-1}_{D_{E^\top p}|x}(u)\, du$ for given p and x. First, we compute $E^\top p$. Because $E^\star$ has only $2K$ non-zero elements and p is one dimensional, it takes $2K$ operations to compute. Random variable D takes two values. As a result, the quantile function $F^{-1}_{D|x}(u)$ is a step function with two jumps. Similarly, discrete random variable

$D_{E^\top p}$ has J values, where each value requires 1 multiplication/operation to compute. The function $F^{-1}_{D_{E^\top p}|x}(u)$ is a step function with J jumps. Sorting its J jumps needs $\frac{1}{2}J(J-1)$ number of operations. The product $F^{-1}_{D|x}(u) F^{-1}_{D_{E^\top p}|x}(u)$ is also a step function of u with $J+1$ jumps. The integration in $\int_0^1 F^{-1}_{D|x}(u) F^{-1}_{D_{E^\top p}|x}(u)\,du$ is essentially a summation of $J+1$ terms, where each term requires 1 multiplication. In total, we need

$$2K + J + \frac{1}{2}J(J-1) + J + 1 + J + 1 \simeq \frac{1}{2}\left(J^2 + 3J\right) + 2K$$

operations to compute $\int_0^1 F^{-1}_{D|x}(u) F^{-1}_{D_{E^\top p}|x}(u)\,du$. □

Proof of Corollary 5.1: Following the discussion in Section 5, we know that the identified set for $\delta_{DD}\left(j, j^\dagger\right) = \theta_j - \theta_{j^\dagger}$ is an interval or a degenerated interval. We aim to find the two endpoints of the interval. Let $E = e_+(j) + e_-\left(j^\dagger\right)$.

Part 1. We first derive the upper bound. Let $p = 1$. The discussion in Section 5 shows that

$$h_{\Delta_{DD}}(1) = h_{\Theta_I}\left(E^\top\right) = \sup_{\theta \in \Theta_I} \delta_{DD}\left(j, j^\dagger\right).$$

Thus, it suffices to compute

$$h_{\Theta_I}\left(E^\top\right) = \int \int_0^1 F^{-1}_{D|x}(u) F^{-1}_{D_E|x}(u)\,du d\mu_X,$$

where $D = I\{Y_1 = 1\}$ and

$$D_E = I\{Y_0 = a_j\} \Pr(Y_0 = a_j)^{-1} - I\left\{Y_0 = a_{j^\dagger}\right\} \Pr\left(Y_0 = a_{j^\dagger}\right)^{-1}.$$

Define

$$\delta^U_{DD}\left(j, j^\dagger \mid x\right) \equiv \int_0^1 F^{-1}_{D|x}(u) F^{-1}_{D_E|x}(u)\,du$$

We now compute $\delta^U_{DD}\left(j, j^\dagger \mid x\right)$ for each x. To simply the notation, we use $\Pr(\cdot \mid x)$ to denote $\Pr(\cdot \mid X = x)$. We also let $p_j \equiv \Pr(Y_0 = a_j)^{-1}$ and $p_{j^\dagger} \equiv \Pr\left(Y_0 = a_{j^\dagger}\right)^{-1}$.

A simple calculation would show that

$$F_{D|x}(d) = \begin{cases} 0 & d < 0 \\ \Pr(Y_1 = 0 \mid x) & 0 \le d < 1 \\ 1 & 1 \le d \end{cases} \text{ and}$$

$$F^{-1}_{D|x}(u) = \begin{cases} 0 & 0 \le u \le \Pr(Y_1 = 0 \mid x) \\ 1 & \Pr(Y_1 = 0 \mid x) < u \le 1 \end{cases}.$$

Similarly, we obtain that

$$F_{D_E|x}(d) = \begin{cases} 0 & d < -p_{j^\dagger} \\ \Pr\left(Y_0 = a_{j^\dagger} \mid x\right) & -p_{j^\dagger} \le d < 0 \\ \Pr(Y_0 \neq a_j \mid x) & 0 \le d < p_j \\ 1 & p_j \le d \end{cases}.$$

This implies that

$$F^{-1}_{D_E|x}(u) = \begin{cases} -p_{j^\dagger} & 0 \le u \le \Pr\left(Y_0 = a_{j^\dagger} \mid x\right) \\ 0 & \Pr\left(Y_0 = a_{j^\dagger} \mid x\right) < u \le \Pr(Y_0 \neq a_j \mid x) \\ p_j & \Pr(Y_0 \neq a_j \mid x) < u \le 1 \end{cases}.$$

Therefore, for any given x, it holds that $\delta^U_{DD}\left(j, j^\dagger \mid x\right)$

$$= \begin{cases} -p_{j^\dagger}\left[P_A(x) + P_C(x) - 1\right] + p_j P_B(x) & \text{if } 1 - P_A(x) < P_C(x) \\ p_j P_B(x) & \text{if } P_C(x) \le 1 - P_A(x) < 1 - P_B(x), \\ p_j P_A(x) & \text{if } 1 - P_B(x) \le 1 - P_A(x) \end{cases}$$

where the conditional probabilities $P_A(x)$, $P_B(x)$, and $P_C(x)$ are defined in the corollary. By plugging the values of p_j and $p_{j^\dagger}$ into the above expression and combining terms, we obtain the expression of $\delta^U_{DD}\left(j, j^\dagger \mid x\right)$:

$$\delta^U_{DD}\left(j, j^\dagger \mid x\right) = \frac{\min\left\{P_A(x), P_B(x)\right\}}{\Pr(Y_0 = a_j)} - \frac{\max\left\{P_A(x) + P_C(x) - 1, 0\right\}}{\Pr\left(Y_0 = a_{j^\dagger}\right)}.$$

Therefore, we obtain that the upper bound for $\delta_{DD}\left(j, j^{\dagger}\right)$ as

$$\int \delta_{DD}^{U}\left(j, j^{\dagger} \mid X\right) d\mu_X = \frac{\mathbb{E}\left[\min\left\{P_A\left(X\right), P_B\left(X\right)\right\}\right]}{\Pr\left(Y_0 = a_j\right)} - \frac{\mathbb{E}\left[\max\left\{P_A\left(X\right) + P_C\left(X\right) - 1, 0\right\}\right]}{\Pr\left(Y_0 = a_{j^{\dagger}}\right)}.$$

Part 2. By the definition of the support function, we have that

$$-\inf_{\theta \in \Theta_I}\left(\theta_j - \theta_{j^{\dagger}}\right) = \sup_{\theta \in \Theta_I} -\left(\theta_j - \theta_{j^{\dagger}}\right) = h_{\Delta_{DD}}\left(-1\right).$$

In consequence, the lower bound is obtain by computing $-h_{\Delta_{DD}}\left(-1\right) = -h_{\Theta_I}\left(-E^{\top}\right)$, where

$$-h_{\Theta_I}\left(-E^{\top}\right) = \int \int_0^1 -F_{D|x}^{-1}\left(u\right) F_{D_{-E}|x}^{-1}\left(u\right) du d\mu_X \equiv \int \delta_{DD}^{L}\left(j, j^{\dagger} \mid x\right) d\mu_X,$$

where $D = I\left\{Y_1 = 1\right\}$ and

$$D_{-E} = -I\left\{Y_0 = a_j\right\} \Pr\left(Y_0 = a_j\right)^{-1} + I\left\{Y_0 = a_{j^{\dagger}}\right\} \Pr\left(Y_0 = a_{j^{\dagger}}\right)^{-1}.$$

We can obtain that

$$F_{D_{-E}|x}\left(d\right) = \begin{cases} 0 & d < -p_j \\ \Pr\left(Y_0 = a_j \mid x\right) & -p_j \leq d < 0 \\ \Pr\left(Y_0 \neq a_{j^{\dagger}} \mid x\right) & 0 \leq d < p_{j^{\dagger}} \\ 1 & p_{j^{\dagger}} \leq d \end{cases}.$$

This implies that

$$F_{D_{-E}|x}^{-1}\left(u\right) = \begin{cases} -p_j & 0 \leq u \leq \Pr\left(Y_0 = a_j \mid x\right) \\ 0 & \Pr\left(Y_0 = a_j \mid x\right) < u \leq \Pr\left(Y_0 \neq a_{j^{\dagger}} \mid x\right) \\ p_{j^{\dagger}} & \Pr\left(Y_0 \neq a_{j^{\dagger}} \mid x\right) < u \leq 1 \end{cases}.$$

Therefore, for any given x, it holds that $\delta_{DD}^{L}\left(j, j^{\dagger} \mid x\right)$

$$= \begin{cases} p_j \left[P_B\left(x\right) + P_A\left(x\right) - 1\right] - p_{j^{\dagger}} P_C\left(x\right) & \text{if } 1 - P_A\left(x\right) \leq P_B\left(x\right) \\ -p_{j^{\dagger}} P_C\left(x\right) & \text{if } P_B\left(x\right) \leq 1 - P_A\left(x\right) < 1 - P_C\left(x\right) \\ -p_{j^{\dagger}} P_A\left(x\right) & \text{if } 1 - P_C\left(x\right) \leq 1 - P_A\left(x\right) \end{cases}.$$

We obtain the expression of $\delta_{DD}^{L}\left(j, j^{\dagger} \mid x\right)$ by replacing p_j and $p_{j^{\dagger}}$ with their values and combining terms:

$$\delta_{DD}^{L}\left(j, j^{\dagger} \mid x\right) = \frac{\max\left\{P_B\left(x\right) + P_A\left(x\right) - 1, 0\right\}}{\Pr\left(Y_0 = a_j\right)} - \frac{\min\left\{P_A\left(x\right), P_C\left(x\right)\right\}}{\Pr\left(Y_0 = a_{j^{\dagger}}\right)}.$$

Therefore, we obtain that the lower bound for $\delta_{DD}\left(j, j^{\dagger}\right)$ as

$$\begin{aligned} \int \delta_{DD}^{L}\left(j, j^{\dagger} \mid X\right) d\mu_X = & \frac{\mathbb{E}\left[\max\left\{P_B\left(X\right) + P_A\left(X\right) - 1, 0\right\}\right]}{\Pr\left(Y_0 = a_j\right)} \\ & - \frac{\mathbb{E}\left[\min\left\{P_A\left(X\right), P_C\left(X\right)\right\}\right]}{\Pr\left(Y_0 = a_{j^{\dagger}}\right)}. \end{aligned}$$

This completes the proof. □

Proof of Lemma 6.1: Let $\Gamma_{10|x}^{\triangle}\left(\cdot\right)$ solve the partial transport problem (12). We aim to show that there exists $\Gamma_{10|x}^{*}\left(\cdot\right)$ with monotone support such that

$$\sum_{j=1}^{J}\sum_{i=0}^{1} \pi\left(i, j\right) \Gamma_{10|x}^{*}\left(i, 1, j\right) = \sum_{j=1}^{J}\sum_{i=0}^{1} \pi\left(i, j\right) \Gamma_{10|x}^{\triangle}\left(i, 1, j\right). \tag{19}$$

Set $\Gamma_{0|x}^{\triangle}\left(j\right) \equiv \sum_{i=0}^{1} \Gamma_{10|x}^{\triangle}\left(i, 1, j\right)$. Consider the following optimization problem:

$$\begin{aligned} & \min_{\Gamma_{10|x}} \sum_{j=1}^{J}\sum_{i=0}^{1} \pi\left(i, j\right) \Gamma_{10|x}\left(i, 1, j\right) \qquad (20) \\ \text{s.t. } & \text{(i) } \Gamma_{10|x}\left(i, 1, j\right) \geq 0 \text{ for } i = 0, 1 \text{ and } j = 1, \ldots, J, \\ & \text{(ii) } \sum_{i=0}^{1} \Gamma_{10|x}\left(i, 1, j\right) = \Gamma_{0|x}^{\triangle}\left(j\right) \text{ for } j = 1, \ldots, J, \text{ and} \\ & \text{(iii) } \sum_{j=1}^{J} \Gamma_{10|x}\left(i, 1, j\right) = \Gamma_{1|x}\left(i, 1\right) \text{ for } i = 0, 1. \end{aligned}$$

Because $\sum_{i=0}^{1}\Gamma_{1|x}(i,1) = \sum_{j=1}^{J}\Gamma_{0|x}^{\triangle}(j)$, (20) is a full optimal transport problem between $\Gamma_{0|x}^{\triangle}(\cdot)$ and $\Gamma_{1|x}(\cdot)$ by ignoring the normalizing constant $1/\sum_{i=0}^{1}\Gamma_{1|x}(i,1)$. We claim that $\Gamma_{10|x}^{\triangle}(\cdot)$ then must solve (20). This holds because if there is another $\Gamma_{10|x}^{\triangledown}(\cdot)$ that satisfies the above constraints and

$$\sum_{j=1}^{J}\sum_{i=0}^{1}\pi(i,j)\,\Gamma_{10|x}^{\triangledown}(i,1,j) < \sum_{j=1}^{J}\sum_{i=0}^{1}\pi(i,j)\,\Gamma_{10|x}^{\triangle}(i,1,j),$$

then it clearly satisfies the constraints in (12), and therefore $\Gamma_{10|x}^{\triangle}(\cdot)$ cannot be optimal in (12). This establishes the claim.

Second, it is well known that the full optimal transport problem (20) has a solution with monotone support. Therefore, if the support of $\Gamma_{10|x}^{\triangle}(\cdot)$ is not monotone, there exists another solution $\Gamma_{10|x}^{*}(\cdot)$ to (20) whose support is monotone. Since $\Gamma_{10|x}^{*}(\cdot)$ satisfies the constraints in (20), it must also satisfy the less stringent constraints in (12). By the optimality in (20), Equality (19) holds. Hence, $\Gamma_{10|x}^{*}(\cdot)$ is optimal in (12), which proves the desired result. □

Proof of Lemma 6.2: We compute the number of basic operations in DREAM. We often overestimate the required operations to simplify the calculation.

During initialization, we first compute J differences and then sort them. This requires in total $\frac{1}{2}J(J+1)$ operations. In the worst case where $JL = JU$, Step 1 requires $J(J+1)$ operations.

In the beginning of Step 2, we rank $\pi(0,1),\ldots,\pi(0,jj)$, which takes $\frac{1}{2}jj(jj-1)$ operations. During the first for-loop, finding the index w requires jj operations; assigning values takes in total 4 operations. Thus, the first for-loop needs $jj(jj+4)$ operations. Similarly, ranking $\pi(1,jj),\ldots,\pi(1,J)$ and the subsequence for-loop takes $\frac{3}{2}(J-jj)^2+\frac{7}{2}(J-jj)$ operations. We can obtain that the time complexity of Step 2 is at most $\left(\frac{3}{4}J^2+\frac{7}{2}J\right)$.

Step 3 requires first constructing two arrays: *dd* and *llb*. During the first for-loop, finding the index and appending values to *dd* and *lld* need $(jj+3)$ operations in total. So the first for-loop requires $jj(jj+3)$ operations. The second loop takes maximum $(J-jj)(J-jj+3)$ operations. Thus, we need at most about $\left(\frac{1}{2}J^2+3J\right)$ operations to obtain *dd* and *llb*. The length of *dd* is at most J. Next, we rank *dd*, which takes J operations. The third for-loop contains $J+13$ basic operations for

each loop and $J(J+13)$ operations in total.

We repeat Steps 2 and 3 at most J times. As a result, the total number of operations is $\frac{1}{2}(3J^3+38J^2+9J)$. Note that this is a greatly amplified upper bound because we only consider worst cases during the calculation and ignore results that can be reused across steps. □

Proof of Corollary 6.1: We first provide the definitions of each term in the corollary. For any $j \in \{1,\ldots,J\}$, define

$$\begin{aligned}
\theta_{2j-1}^U &\equiv \mathbb{E}\left[\min\left\{\Pr\left(Y_{1s}=1, Y_{1r}=1 \mid X\right), \Pr\left(Y_0=a_j \mid X\right)\right\}\right], \\
\theta_{2j-1}^L &\equiv \mathbb{E}\left[\max\left\{\Pr\left(Y_{1s}=1, Y_{1r}=1 \mid X\right)+\Pr\left(Y_0=a_j \mid X\right)-1, 0\right\}\right], \\
\theta_{2j}^U &\equiv \mathbb{E}\left[\min\left\{\Pr\left(Y_{1s}=0, Y_{1r}=1 \mid X\right), \Pr\left(Y_0=a_j \mid X\right)\right\}\right], \text{ and} \\
\theta_{2j}^L &\equiv \mathbb{E}\left[\max\left\{\Pr\left(Y_{1s}=0, Y_{1r}=1 \mid X\right)+\Pr\left(Y_0=a_j \mid X\right)-1, 0\right\}\right].
\end{aligned}$$

Without loss of generality, we prove the result for $\delta_{TPRD}(1,2)$ with any $J \geq 2$. Since any one-dimensional connected set is an interval, it suffices to derive the tight upper and lower bounds of the identified set. We focus on the tight upper bound. The tight lower bound can be obtained in the same way.

By definition, we have that

$$\delta_{TPRD}(1,2) \equiv \frac{\theta_1^*}{\theta_1^*+\theta_2^*} - \frac{\theta_3^*}{\theta_3^*+\theta_4^*}.$$

It is easy to see that $\delta_{TPRD}(1,2)$ is an increasing function of θ_1^* and θ_4^* and a decreasing function of θ_2^* and θ_3^*. Let θ_1^U, θ_4^U, θ_3^L, and θ_4^L denote the tight upper bounds of θ_1^* and θ_4^* and tight lower bounds of θ_2^* and θ_3^* respectively. Let $\Theta_I^{(1-4)}$ be the set obtained from projecting Θ_I onto its first four elements $(\theta_1,\theta_2,\theta_3,\theta_4)$. In the following, we first provide the expressions of θ_1^U, θ_4^U, θ_3^L, and θ_4^L. Then we show that the bounds can be achieved simultaneously: $\left(\theta_1^U, \theta_2^L, \theta_3^L, \theta_4^U\right) \in \Theta_I^{(1-4)}$.

By the support function expressed in (11), we can let $q = [1,0,\ldots,0]^\top$ to obtain the tight upper bound of θ_1^*. For any $\theta \in \Theta_I$, its first element θ_1 satisfies that

$$\int \left\{ \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int\int I\left\{y_{1s}=1, y_{1r}=1, y_0=a_1\right\} d\mu_{10|x} \right\} d\mu_X \geq \theta_1.$$

In addition, the bound is tight by Lemma A.4. We obtain that

$$\begin{aligned}
&\int \left\{ \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int \int I\left\{y_{1s}=1, y_{1r}=1, y_0=a_1\right\} d\mu_{10|x} \right\} d\mu_X \\
&= \int \left\{ \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \int \int I\left\{y_0=a_1\right\} I\left\{y_{1s}=1, y_{1r}=1\right\} d\mu_{10|x} \right\} d\mu_X \\
&= \int \left\{ \int_0^1 F^{-1}_{D_0|x}(u)\, F^{-1}_{D_1|x}(1-u)\, du \right\} d\mu_X,
\end{aligned}$$

where $D_0 = I\left\{Y_0 = a_1\right\}$ and $D_1 = I\left\{Y_{1s}=1, Y_{1r}=1\right\}$. The last equality follows from the monotone rearrangement inequality. To simply the notation, we use $\Pr\left(\cdot \mid x\right)$ to denote $\Pr\left(\cdot \mid X = x\right)$. We have that

$$F^{-1}_{D_0|x}(u) = \begin{cases} 1 & 0 \le u \le \Pr\left(Y_0 = a_1 \mid x\right) \\ 0 & \Pr\left(Y_0 = a_1 \mid x\right) < u \le 1 \end{cases}, \text{ and}$$

$$F^{-1}_{D_1|x}(1-u) = \begin{cases} 1 & 0 \le u \le \Pr\left(Y_{1s}=1, Y_{1r}=1 \mid x\right) \\ 0 & \Pr\left(Y_{1s}=1, Y_{1r}=1 \mid x\right) < u \le 1 \end{cases}.$$

We thus obtain the tight upper bound of θ_1^* as

$$\theta_1^U = \mathbb{E}\left[\min\left\{\Pr\left(Y_{1s}=1, Y_{1r}=1 \mid X\right), \Pr\left(Y_0 = a_1 \mid X\right)\right\}\right].$$

Applying the same technique, we can obtain that

$$\begin{aligned}
\theta_2^L &= \mathbb{E}\left[\max\left\{\Pr\left(Y_{1s}=0, Y_{1r}=1 \mid X\right) + \Pr\left(Y_0 = a_1 \mid X\right) - 1, 0\right\}\right], \\
\theta_3^L &= \mathbb{E}\left[\max\left\{\Pr\left(Y_{1s}=1, Y_{1r}=1 \mid X\right) + \Pr\left(Y_0 = a_2 \mid X\right) - 1, 0\right\}\right], \text{ and} \\
\theta_4^U &= \mathbb{E}\left[\min\left\{\Pr\left(Y_{1s}=0, Y_{1r}=1 \mid X\right), \Pr\left(Y_0 = a_2 \mid X\right)\right\}\right].
\end{aligned}$$

Next, we show that $\left(\theta_1^U, \theta_2^L, \theta_3^L, \theta_4^U\right) \in \Theta_I^{(1-4)}$. Let $q = [1, -1, -1, 1, 0, \ldots, 0]^\top$. We ignore the normalizing factor that makes $q \in \mathbb{S}^{2J}$ hold to simplify the derivation. Then, by the definition of support function, we have that

$$h_{\Theta_I}\left([1, -1, -1, 1, 0, \ldots, 0]^\top\right) \ge \theta_1 - \theta_2 - \theta_3 + \theta_4.$$

By Lemma A.4, the left-hand-side of the inequality is the tight upper bound of $\theta_1 - \theta_2 - \theta_3 + \theta_4$. Because $\theta_1 - \theta_2 - \theta_3 + \theta_4$ is an increasing function of θ_1 and θ_4 and a

decreasing function of θ_3 and θ_4, the following equality

$$h_{\Theta_I}\left([1,-1,-1,1,0,\ldots,0]^\top\right) = \theta_1^U - \theta_2^L - \theta_3^L + \theta_4^U$$

holds if and only if $\left(\theta_1^U, \theta_2^L, \theta_3^L, \theta_4^U\right) \in \Theta_I^{(1-4)}$. In the following, we solve the optimal transport problem on the left-hand-side and verify that the equality holds.

Plugging in the expression of m_2, we have that

$$\begin{aligned}
&h_{\Theta_I}\left([1,-1,-1,1,0,\ldots,0]^\top\right) \\
&= \int \left\{ \inf_{\mu_{10|x} \in \mathcal{M}\left(\mu_{1|x}, \mu_{0|x}\right)} \left[\int\int I\left\{y_{1s}=1, y_{1r}=1, y_0=a_1\right\} d\mu_{10|x} \right.\right. \\
&\qquad - \int\int I\left\{y_{1s}=0, y_{1r}=1, y_0=a_1\right\} d\mu_{10|x} \\
&\qquad - \int\int I\left\{y_{1s}=1, y_{1r}=1, y_0=a_2\right\} d\mu_{10|x} \\
&\qquad \left.\left. + \int\int I\left\{y_{1s}=0, y_{1r}=1, y_0=a_2\right\} d\mu_{10|x} \right] \right\} d\mu_X .
\end{aligned}$$

Let $\mu_{10|x}^\star$ denotes the optimal coupling conditional on $X = x$. We have that

$$\begin{aligned}
&h_{\Theta_I}\left([1,-1,-1,1,0,\ldots,0]^\top\right) \\
&= \int \left\{ \left[\int\int I\left\{y_{1s}=1, y_{1r}=1, y_0=a_1\right\} d\mu_{10|x}^\star - \int\int I\left\{y_{1s}=0, y_{1r}=1, y_0=a_1\right\} d\mu_{10|x}^\star \right.\right. \\
&\quad \left.\left. \int\int -I\left\{y_{1s}=1, y_{1r}=1, y_0=a_2\right\} d\mu_{10|x}^\star + \int\int I\left\{y_{1s}=0, y_{1r}=1, y_0=a_2\right\} d\mu_{10|x}^\star \right] \right\} d\mu_X .
\end{aligned} \tag{21}$$

Next, we solve for $\mu_{10|x}^\star$.

Note that the cost function assigns only the values of ± 1. We must couple all of the mass of the points $(y_{1s}, y_{1r}) \in \{(1,1),(0,1)\}$ with the points $y_0 \in \{a_1, a_2\}$. Note that the mass satisfies

$$\mu_{1|x}\left((1,1)\right) + \mu_{1|x}\left((0,1)\right) \leq 1 = \mu_{0|x}\left(a_1\right) + \mu_{0|x}\left(a_2\right). \tag{22}$$

The extra mass on the $\mu_{1|x}$ side (at points $(1,0)$ and $(0,0)$) is assigned with cost 0, regardless of where it couples to.

The cost of coupling $(1,1)$ with a_1 and $(0,1)$ with a_2 is -1, while the cost of $(1,1)$

with a_2 and $(0,1)$ with a_1 is 1. It is therefore optimal to assign as much mass to the former configurations as possible. As a result, the optimizer $\mu^\star_{10|x}$ will satisfy

$$\mu^\star_{10|x}\big((1,1),a_1\big) = \min\left\{\mu_{1|x}\left((1,1)\right),\mu_{0|x}\left(a_1\right)\right\} \text{ and}$$
$$\mu^\star_{10|x}\big((0,1),a_2\big) = \min\left\{\mu_{1|x}\left((0,1)\right),\mu_{0|x}\left(a_2\right)\right\}.$$

Now, if either $\mu_{1|x}\left((1,1)\right) > \mu_{0|x}\left(a_1\right)$ or $\mu_{1|x}\left((0,1)\right) > \mu_{0|x}\left(a_2\right)$ (note that by (22), at most one of these can be true) then these configurations do not use up all of the mass from $(1,1)$ and $(0,1)$ and we will have to couple either the remaining mass from $(1,1)$ with a_2 or the remaining mass from $(0,1)$ with a_1, at cost 1. The optimal coupling therefore satisfies

$$\mu^\star_{10|x}\big((1,1),a_2\big) = \max\left\{\mu_{1|x}\left((1,1)\right)-\mu_{0|x}\left(a_1\right),0\right\} = \max\left\{\mu_{1|x}\left((1,1)\right)+\mu_{0|x}\left(a_2\right)-1,0\right\} \text{ and}$$
$$\mu^\star_{10|x}\big((0,1),a_1\big) = \max\left\{\mu_{1|x}\left((0,1)\right)-\mu_{0|x}\left(a_2\right),0\right\} = \max\left\{\mu_{1|x}\left((0,1)\right)+\mu_{0|x}\left(a_1\right)-1,0\right\}.$$

Computing each term in (21) with the optimal coupling $\mu^\star_{10|x}$, we have that the optimal cost conditioning on $X = x$ is

$$\mu^\star_{10|x}\big((1,1),a_1\big) + \mu^\star_{10|x}\big((0,1),a_2\big) - \mu^\star_{10|x}\big((1,1),a_2\big) - \mu^\star_{10|x}\big((0,1),a_1\big).$$

Integrating over x and rearranging terms, we obtain that

$$h_{\Theta_I}\left(\left[1,-1,-1,1,0,\ldots,0\right]^\top\right) = \theta_1^U - \theta_2^L - \theta_3^L + \theta_4^U.$$

This completes the proof of the corollary. □

B Additional Detail on DREAM

In this appendix, we provide additional details on Steps 2 and 3 of DREAM.

In Step 2, we solve **Problem 1** and **Problem 2**. The first for-loop solves **Problem 1** during which we assign values to $\Gamma_{10|x}\left(0,1,j\right)$ for $j \leq jj$ according to the following rule. We first order $\pi\left(0,j\right)$ for $j \leq jj$ from the smallest to the largest. Then we assign the value of $\Gamma_{0|x}\left(j\right)$ to $\Gamma_{10|x}\left(0,1,j\right)$ following this order until we use all the mass in $\Gamma_{1|x}\left(0,1\right)$. If we exhaust $\Gamma_{1|x}\left(0,1\right)$ before the assignment of $\Gamma_{10|x}\left(0,1,j\right)$ for every j, then we let all the $\Gamma_{10|x}\left(0,1,j\right)$ without assignment equal to zero. We will eventually exhaust $\Gamma_{1|x}\left(0,1\right)$ after assigning mass to all j because

$\sum_{j=1}^{jj} \Gamma_{0|x}(j) \geq \Gamma_{1|x}(0,1)$. The procedure is the same for the second for-loop which solves **Problem 2**. We assign the value of $\Gamma_{0|x}(j)$ to $\Gamma_{10|x}(1,1,j)$ starting from j corresponding to the smallest $\pi(1,j)$ until we use all the mass in $\Gamma_{1|x}(1,1)$. Next, we check if the third constraint $\sum_{i=0}^{1} \Gamma_{10|x}(i,1,jj) \leq \Gamma_{0|x}(jj)$ in (13) is satisfied. If not, we execute Step 3.

During Step 3, we move the masses from $\Gamma_{10|x}(0,1,jj)$ to other $\Gamma_{10|x}(0,1,j)$ and/or mass from $\Gamma_{10|x}(1,1,jj)$ to other $\Gamma_{10|x}(1,1,j)$ until the constraint is satisfied. We shall not look for j such that $\pi(0,j) < \pi(0,jj)$ because we have already assigned the largest possible mass to $\Gamma_{10|x}(0,1,j)$, otherwise there will be no mass left for $\Gamma_{10|x}(0,1,jj)$. For a similar reason, we do not need to look for j such that $\pi(1,j) < \pi(1,jj)$. Therefore, we only work on $0,j$ pair such that $\pi(0,j) \geq \pi(0,jj)$ and $1,j$ pair such that $\pi(1,j) \geq \pi(1,jj)$. Denote the two sets of j's as $\mathcal{J}_0$ and $\mathcal{J}_1$. Rank $\pi(0,j) - \pi(0,jj)$ for all $j \in \mathcal{J}_0$ together with $\pi(1,j) - \pi(1,jj)$ for all $j \in \mathcal{J}_1$ from the smallest to the largest. Then move mass from $\Gamma_{10|x}(0,1,jj)$ and/or $\Gamma_{10|x}(1,1,jj)$ to $\Gamma_{10|x}(0,1,j)$ and/or to $\Gamma_{10|x}(1,1,j)$ following this order until $\sum_{i=0}^{1} \Gamma_{10|x}(i,1,jj) \leq \Gamma_{0|x}(jj)$. The inequality will eventually be satisfied because $\sum_{j=1}^{jj} \Gamma_{0|x}(j) \geq \Gamma_{1|x}(0,1)$ and $\sum_{j=jj}^{J} \Gamma_{0|x}(j) \geq \Gamma_{1|x}(1,1)$, which implies that there is enough space to move away the extra mass at $\Gamma_{10|x}(0,1,jj)$ and/or $\Gamma_{10|x}(1,1,jj)$.

References

Andrews, D. W. and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666.

Andrews, D. W. and G. Soares (2010): "Inference for parameters defined by moment inequalities using generalized moment selection," *Econometrica*, 78, 119–157.

Athey, S., R. Chetty, G. W. Imbens, and H. Kang (2019): "The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely," Tech. rep., National Bureau of Economic Research.

Bei, X. (2024): "Local linearization based subvector inference in moment inequality models," *Journal of Econometrics*, 238, 105549.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp identification regions in models with convex moment predictions," *Econometrica*, 79, 1785–1821.

Beresteanu, A. and F. Molinari (2008): "Asymptotic properties for a class of partially identified models," *Econometrica*, 76, 763–814.

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set identified linear models," *Econometrica*, 80, 1129–1155.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric efficiency in GMM models with auxiliary data," *The Annals of Statistics*, 36, 808–843.

Chernozhukov, V., S. Lee, and A. M. Rosen (2013): "Intersection bounds: Estimation and inference," *Econometrica*, 81, 667–737.

Cohen, M. B., Y. T. Lee, and Z. Song (2021): "Solving linear programs in the current matrix multiplication time," *Journal of the ACM (JACM)*, 68, 1–39.

Dantzig, G. B. (2016): "Linear programming and extensions," in *Linear programming and extensions*, Princeton university press.

Fan, Y., E. Guerre, and D. Zhu (2017): "Partial identification of functionals of the joint distribution of "potential outcomes"," *Journal of econometrics*, 197, 42–59.

Fan, Y. and C. A. Manzanares (2017): "Partial identification of average treatment effects on the treated through difference-in-differences," *Econometric Reviews*, 36, 1057–1080.

Fan, Y., H. Park, and G. Xu (2023a): "Quantifying Distributional Model Risk in Marginal Problems via Optimal Transport," *arXiv preprint arXiv:2307.00779*.

Fan, Y., R. Sherman, and M. Shum (2014): "Identifying treatment effects under data combination," *Econometrica*, 82, 811–822.

——— (2016): "Estimation and inference in an ecological inference model," *Journal of Econometric Methods*, 5, 17–48.

Fan, Y., X. Shi, and J. Tao (2023b): "Partial identification and inference in moment models with incomplete data," *Journal of Econometrics*, 235, 418–443.

Graham, B. S., C. C. d. X. Pinto, and D. Egel (2016): "Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST)," *Journal of Business & Economic Statistics*, 34, 288–301.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 315–331.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the econometric society*, 1029–1054.

Hwang, Y. (2023): "Bounding Omitted Variable Bias Using Auxiliary Data: With an Application to Estimate Neighborhood Effects," *Available at SSRN 3866876*.

Kaido, H. (2016): "A dual approach to inference for partially identified econometric models," *Journal of econometrics*, 192, 269–290.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence intervals for projections of partially identified parameters," *Econometrica*, 87, 1397–1432.

Kallus, N., X. Mao, and A. Zhou (2022): "Assessing algorithmic fairness with unobserved protected class using data combination," *Management Science*, 68, 1959–1981.

Karmarkar, N. (1984): "A new polynomial-time algorithm for linear programming," in *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, 302–311.

Pacini, D. (2019): "Two-sample least squares projection," *Econometric Reviews*, 38, 95–123.

Peyré, G. and M. Cuturi (2019): "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, 11, 355–607.

Rachev, S. T. and L. Rüschendorf (2006): *Mass Transportation Problems: Volume 1: Theory*, Springer Science & Business Media.

Rüschendorf, L. (1991): "Bounds for distributions with multivariate marginals," *Lecture Notes-Monograph Series*, 285–310.

Santambrogio, F. (2015): *Optimal transport for applied mathematicians*, vol. 87, Springer.

van den Brand, J. (2020): "A deterministic linear program solver in current matrix multiplication time," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SIAM, 259–278.

Vianney, P. and G. Vigeral (2015): "A minmax theorem for concave-convex mappings with no regularity assumptions." *Journal of Convex Analysis*, 22, 537–540.

Villani, C. et al. (2008): *Optimal transport: old and new*, vol. 338, Springer.

Williams, V. V., Y. Xu, Z. Xu, and R. Zhou (2024): "New bounds for matrix multiplication: from alpha to omega," in *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, SIAM, 3792–3835.

Wright, S. J. (1997): *Primal-dual interior-point methods*, SIAM.