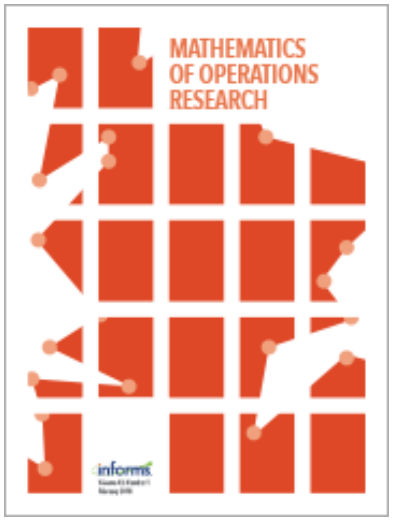

Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Quantifying Distributional Model Risk in Marginal Problems via Optimal Transport

Yanqin Fan, Hyeonseok Park, Gaoqian Xu

To cite this article:
Yanqin Fan, Hyeonseok Park, Gaoqian Xu (2025) Quantifying Distributional Model Risk in Marginal Problems via Optimal Transport. Mathematics of Operations Research

Published online in Articles in Advance 06 Aug 2025

. https://doi.org/10.1287/moor.2024.0557

https://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–44
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Quantifying Distributional Model Risk in Marginal Problems via Optimal Transport

Yanqin Fan,[a] Hyeonseok Park,[b,*] Gaoqian Xu[a]

[a] Department of Economics, University of Washington, Seattle, Washington 98195; [b] Center for Industrial and Business Organization and Institute for Advanced Economic Research, Dongbei University of Finance and Economics, Dalian, Liaoning 116025, China
*Corresponding author
Contact: fany88@uw.edu, https://orcid.org/0009-0009-6929-4127 (YF); hynskpark21@dufe.edu.cn, https://orcid.org/0009-0000-2556-3967 (HP); gx8@uw.edu, https://orcid.org/0009-0001-5084-4389 (GX)

Received: June 10, 2024
Revised: March 20, 2025
Accepted: May 18, 2025
Published Online in Articles in Advance: August 6, 2025

MSC2020 Subject Classifications: Primary: 90C15; secondary: 90C46

https://doi.org/10.1287/moor.2024.0557

Abstract. This paper studies distributional model risk in marginal problems, where each marginal measure is assumed to lie in a Wasserstein ball. We establish fundamental results including strong duality, finiteness of the proposed Wasserstein distributional model risk, and the existence of an optimizer at each radius. We also show continuity of the Wasserstein distributional model risk as a function of the radius. Using strong duality, we extend the well-known Makarov bounds for the distribution function of the sum of two random variables with given marginals to Wasserstein distributionally robust Makarov bounds. We illustrate our results on four distinct applications when the sample information comes from multiple data sources and only some marginal reference measures are identified: partial identification of treatment effects, externally valid treatment choice via robust welfare functions, Wasserstein distributionally robust estimation under data combination, and evaluation of the worst aggregate risk measures.

Funding: This work was supported by the National Science Foundation [Infrastructure Grant (PIHOT) DMS-2133244], acknowledged by Y. Fan.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/moor.2024.0557.

Keywords: data combination • distributionally robust optimization • marginal problems • Makarov bounds • partial identification • treatment choice

1. Introduction

Distributionally robust optimization (DRO) has emerged as a powerful tool for hedging against model misspecification and distributional shifts. It minimizes *distributional model risk* (DMR), defined as the worst risk over a class of distributions lying in a *distributional uncertainty set*; see Blanchet and Murthy [5]. Among many different choices of uncertainty sets, Wasserstein DRO (W-DRO) with distributional uncertainty sets based on optimal transport costs has gained much popularity; see Kuhn et al. [35] and Blanchet et al. [6] for recent reviews. W-DRO has found successful applications in robust decision making in all disciplines including economics, finance, machine learning, and operations research. Its success is largely credited to the strong duality and other nice properties of the Wasserstein DMR (W-DMR). The objective of this paper is to propose and study W-DMR in *marginal problems where only some marginal measures of a reference measure are given*; see, for example, Kellerer [31], Rachev and Rüschendorf [45], Villani [52], Villani [53], and Rüschendorf [48].

In practice, *marginal problems* arise from either the lack of complete data or an incomplete model. In insurance and risk management, computing model-free measures of aggregate risks such as Value-at-Risk (VaR) and Expected Short-Fall is of utmost importance and routinely done. When the exact dependence structure between individual risks is lacking, researchers and policy makers rely on the worst risk measures, defined as the maximum value of aggregate risk measures over all joint measures of the individual risks with some fixed marginal measures; see Embrechts and Puccetti [12] and Embrechts et al. [15]. In causal inference, distributional treatment effects such as the variance and the proportion of participants who benefit from the treatment depend on the joint distribution of the potential outcomes. Even with ideal randomized experiments such as double-blind clinical trials, the joint distribution of potential outcomes is not identified, and as a result, only the lower and upper bounds on distributional treatment effects are identified from the sample information; see Fan and Wu [21], Fan and Park [19], Fan and Park [20], Fan et al. [22], Ridder and Moffitt [46], and Firpo and Ridder [23]. In algorithmic fairness when the sensitive group variable is not observed in the main data set, assessment of unfairness measures must be done using multiple data sets; see Kallus et al. [29]. Abstracting away from estimation, all these

problems involve optimizing the expected value of a functional of multiple random variables with fixed marginals and thus belong to the class of marginal problems for which optimal transport-related tools are important.[1]

The marginal measures in the aforementioned applications and general marginal problems are typically empirical measures computed from multiple data sets such as in the evaluation of worst aggregate risk measures or identified under specific assumptions such as randomization or strong ignorability in causal inference. Developing a unified framework for hedging against model misspecification and/or distributional shifts in marginal measures motivates the current paper.

Theoretically, this paper makes several contributions to the literature on distributional robustness and the literature on marginal problems. First, it introduces *Wasserstein distributional model risk in marginal problems* (W-DMR-MP), where each marginal measure is assumed to lie in a Wasserstein ball centered at a fixed reference measure with a given radius. We focus on the important case with two marginals and consider both nonoverlapping and overlapping marginals. For *nonoverlapping marginal measures*, when the radius is zero, the W-DMR-MP reduces to the marginal problems or optimal transport problems studied in Kellerer [31], Rachev and Rüschendorf [45], Villani [52], and Villani [53]. For *overlapping marginals*, when the radius is zero, the W-DMR-MP reduces to the overlapping marginals problem studied in Rüschendorf [48]. Second, we establish strong duality for our W-DMR with both nonoverlapping and overlapping marginals under conditions similar to those for W-DMR; see Zhang et al. [57], Blanchet and Murthy [5], and Gao and Kleywegt [25]. As a first application of our strong duality result for nonoverlapping marginals, we extend the well-known Makarov bounds for the distribution function of the sum of two random variables to Wasserstein distributionally robust Makarov bounds. Third, we prove finiteness of the W-DMR-MP and existence of an optimizer at each radius. Based on both results, we show that the identified set of the expected value of a smooth functional of random variables with fixed marginals is a closed interval. Fourth, we show continuity of the W-DMR in marginal problems as a function of the radius. Together these results extend those for W-DMR in Blanchet and Murthy [5], Zhang et al. [57], and Yue et al. [56]. Lastly, we extend our formulations and theory to W-DMR with multimarginals. On a technical note, our proofs build on existing work on W-DMR such as Blanchet and Murthy [5], Zhang et al. [57], and Yue et al. [56]. However, an additional challenge due to the presence of multiple marginal measures in our Wasserstein uncertain sets is the verification of the existence of a joint measure with overlapping marginals. We make use of existing results for a given consistent product marginal system in Vorob'ev [55], Kellerer [30], and Shortt [50] to address this issue.

Practically, we demonstrate the flexibility and broad applicability of our W-DMR-MP via four distinct applications when the sample information comes from multiple data sources. First, we consider partial identification of treatment effects when the marginal measures of the potential outcomes lie in their respective Wasserstein balls centered at the measures identified under strong ignorability. The validity of strong ignorability is often questionable when unobservable confounders may be present. We apply our W-DMR-MP to establishing the identified sets of treatment effects which can be used to conduct stability/robustness checks to the selection-on-observables assumption. For average treatment effects, we show that when the cost functions are separable, incorporating covariate information does not help shrink the identified set; on the other hand, for nonseparable cost functions such as the Mahalanobis distance, incorporating covariate information may help shrink the identified set. Second, in causal inference when the optimal treatment choice is to be applied to a target population different from the training population, Adjaho and Christensen [1] introduce robust welfare functions defined by W-DMR to study externally valid treatment choice. The W-DMR-MP we propose allows us to dispense with the assumption of a *known dependence structure* for the reference measure in Adjaho and Christensen [1]. When shifts in the covariate distribution are allowed, we show that our robust welfare function is upper bounded by the worst robust welfare function of Adjaho and Christensen [1]. Third, one important application of W-DMR is in distributionally robust estimation and classification. However, as Awasthi et al. [2] point out,[2] some sensitive variables may not be observed in the same data set as the response variable, rendering W-DRO inapplicable. We apply W-DMR-MP to distributionally robust estimation under data combination.[3] Fourth, applying our W-DMR-MP to the evaluation of the worst aggregate risk measures allows us to dispense with the known marginals assumption in Embrechts and Puccetti [12] and Embrechts et al. [15].

The rest of this paper is organized as follows. Section 2 reviews the W-DMR and strong duality, introduces our W-DMR-MP, and then presents four motivating examples. Section 3 establishes strong duality and Wasserstein distributionally robust Makarov bounds. Section 4 studies finiteness of W-DMR-MP and existence of optimal solutions. Moreover, we show that the identified set of the expected value of a smooth functional of random variables with fixed marginals is a closed interval. Section 5 establishes continuity of W-DMR-MP as a function of the radius. Section 6 revisits the motivating examples in Section 2. Section 7 extends our W-DMR-MP to more than two marginals. The last section offers some concluding remarks. Technical proofs are relegated to an appendix. Additional materials and technical lemmas can be found in the Online Supplement.

We close this section by introducing the notation used in the rest of this paper. For two sets A and B, the relative complement is denoted by $A \setminus B$. Let $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, $[d] = \{1, 2, \dots, d\}$, $\mathbb{R}^d_+ = \{x \in \mathbb{R}^d : x_i \geq 0, \ \forall i \in [d]\}$, and $\mathbb{R}^d_{++} = \{x \in \mathbb{R}^d : x_i > 0, \ \forall i \in [d]\}$. For any real numbers $x, y \in \mathbb{R}$, we define $x \wedge y := \min\{x, y\}$ and $x \vee y := \max\{x, y\}$. The Euclidean inner product of x and y in $\mathbb{R}^d$ is denoted by $\langle x, y \rangle$. For any real matrix $W \in \mathbb{R}^{m \times n}$, let $A^\top$ denote the transpose of W. For an extended real function f on $\mathcal{X}$, the positive part f^+ and the negative part f^- are defined as $f^+(x) = \max\{f(x), 0\}$ and $f^-(x) = \max\{-f(x), 0\}$, respectively.

For any Polish space $\mathcal{S}$, let $\mathcal{B}_\mathcal{S}$ be the associated Borel σ-algebra and $\mathcal{P}(\mathcal{S})$ be the collection of probability measures on $\mathcal{S}$. Given a Polish probability space $(\mathcal{S}, \mathcal{B}_\mathcal{S}, \nu)$, let $\mathcal{B}^\nu_\mathcal{S}$ denote the ν-completion of $\mathcal{B}_\mathcal{S}$. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a map $T : \Omega \to \mathcal{S}$, let $T\mu$ denote the push forward of $\mathbb{P}$ by T, that is, $(T\mathbb{P})(A) = \mathbb{P}(T^{-1}(A))$ for all $A \in \mathcal{B}_\mathcal{S}$, where $T^{-1}(A) = \{\omega \in \Omega : T(\omega) \in A\}$. The law of a random variable $S : \Omega \to \mathbb{R}$ is denoted by $\mathrm{Law}(S)$ which is the same as $S\mathbb{P}$. For any $\mu, \nu \in \mathcal{P}(\mathcal{S})$, let $\Pi(\mu, \nu)$ denote the set of all couplings (or joint measures) with marginals μ and ν.

For any $\mathcal{B}^\nu_\mathcal{S}$-measurable function f, let $\int_\mathcal{S} f d\nu$ denote the integral of f in the completion of $(\mathcal{S}, \mathcal{B}_\mathcal{S}, \nu)$. For a random element $S : \Omega \to \mathcal{S}$ with $\mathrm{Law}(S) = \nu$, we write $\mathbb{E}_\nu[f(S)] = \int_\mathcal{S} f d\nu$. Given $p \in (0, \infty)$ and a Borel measure ν on $\mathcal{S}$, let $L^p(\nu) := L^p(\mathcal{S}, \mathcal{B}_\mathcal{S}, \nu)$ denote the set of all the $\mathcal{B}^\nu_\mathcal{S}$-measurable functions $f : \mathcal{S} \to \mathbb{R}$ such that $\|f\|_{L^p(\nu)} := (\int_\mathcal{S} |f|^p d\nu)^{1/p} < \infty$.

2. W-DMR and Motivating Examples

In this section, we first review W-DMR and then introduce W-DMR in marginal problems. Lastly, we present four motivating examples of marginal problems which will be used to illustrate our results in the rest of this paper.

2.1. A Review of W-DMR and Strong Duality

W-DMR is defined as the worst model risk over a class of distributions lying in a Wasserstein uncertainty set composed of all probability measures that are a fixed Wasserstein distance away from a given reference measure; see Blanchet and Murthy [5].

Before presenting W-DMR, we review some basic definitions. Let $\mathcal{X}$ be a Polish (metric) space with a metric $\boldsymbol{d}$.

Definition 1 (Optimal Transport Cost). Let $\mu, \nu \in \mathcal{P}(\mathcal{X})$ be given probability measures. The *optimal transport cost* between μ and ν associated with a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \cup \{\infty\}$ is defined as

$$\boldsymbol{K}_c(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} c \, d\pi.$$

When the cost function c is lower semicontinuous, there exists an optimal coupling corresponding to $\boldsymbol{K}_c(\mu, \nu)$. In other words, there exists $\pi^* \in \Pi(\mu, \nu)$ such that $\boldsymbol{K}_c(\mu, \nu) = \int_{\mathcal{X} \times \mathcal{X}} c \, d\pi^*$ (see, e.g., Villani [52, theorem 4.1]).

Definition 2 (Wasserstein Distance). Let $p \in [1, \infty)$. The *Wasserstein distance* of order p between any two measures μ and ν on Polish metric space $(\mathcal{X}, \boldsymbol{d})$ is defined by

$$\boldsymbol{W}_p(\mu, \nu) = \left[\inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} \boldsymbol{d}^p \, d\pi \right]^{1/p}.$$

Throughout this paper, we make the following assumption on the cost function c.

Assumption 1. *Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ be a Borel space associated with $\mathcal{X}$. The cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+ \cup \{\infty\}$ is measurable and satisfies $c(x, y) = 0$ if and only if $x = y$.*

Assumption 1 implies that for $\mu, \nu \in \mathcal{P}(\mathcal{X})$, $\mu = \nu$ if and only if $\boldsymbol{K}_c(\mu, \nu) = 0$. When c is the metric $\boldsymbol{d}$ on $\mathcal{X}$, $\boldsymbol{K}_c(\mu, \nu)$ coincides with the Wasserstein distance of order 1 (Kantorovich-Rubinstein distance) between μ and ν defined in Definition 2.

For a given function $f : \mathcal{X} \to \mathbb{R}$, Blanchet and Murthy [5] define W-DMR as

$$\mathcal{I}_{\mathrm{DMR}}(\delta) := \sup_{\gamma \in \Sigma_{\mathrm{DMR}}(\delta)} \int_\mathcal{X} f \, d\gamma, \ \delta \geq 0,$$

where $\Sigma_{\mathrm{DMR}}(\delta)$ is the Wasserstein uncertainty set[4] centered at a reference measure $\mu \in \mathcal{P}(\mathcal{X})$ with radius $\delta \geq 0$, that is,

$$\Sigma_{\mathrm{DMR}}(\delta) := \{\gamma \in \mathcal{P}(\mathcal{X}) : \boldsymbol{K}_c(\mu, \gamma) \leq \delta\}.$$

Assumption 1 allows the cost function c to be asymmetric and take value ∞, where the latter corresponds to the case that there is no distributional shift in some marginal measure of μ.

Remark 1. Under Assumption 1, $\Sigma_{\mathrm{DMR}}(0) = \{\mu\}$ and $\mathcal{I}_{\mathrm{DMR}}(0) = \int_{\mathcal{X}} f\, d\mu$.

It is well-known that under mild conditions, strong duality holds for $\mathcal{I}_{\mathrm{DMR}}(\delta)$ when $\delta > 0$ (cf., Blanchet and Murthy [5], Gao and Kleywegt [25], and Zhang et al. [57]). To be self-contained, we restate the strong duality result in Zhang et al. [57] for Polish space below.[5]

Theorem 1 (Zhang et al. [57, Theorem 1]). *Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mu)$ be a probability space. Let $\delta \in (0, \infty)$ and $f : \mathcal{X} \to \mathbb{R}$ be a measurable function such that $\int_{\mathcal{X}} f\, d\mu > -\infty$. Suppose the cost function satisfies* Assumption 1. *Then, for any $\delta > 0$,*

$$\mathcal{I}_{\mathrm{DMR}}(\delta) = \inf_{\lambda \in \mathbb{R}_+} \left\{ \lambda\delta + \int_{\mathcal{X}} \sup_{x' \in \mathcal{X}} [f(x') - \lambda c(x, x')]\, d\mu(x) \right\}, \tag{1}$$

where $\lambda c(x, x')$ is defined to be ∞ when $\lambda = 0$ and $c(x, x') = \infty$.

In the rest of this paper, we keep the convention that for any cost function c, $\lambda c(x, y) = \infty$ when $\lambda = 0$ and $c(x, y) = \infty$.

2.2. W-DMR in Marginal Problems

2.2.1. Nonoverlapping Marginals.

Let $\mathcal{V} := \mathcal{S}_1 \times \mathcal{S}_2$ be the product space of two Polish spaces $\mathcal{S}_1$ and $\mathcal{S}_2$. Let μ_1 and μ_2 be Borel probability measures on $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Following Rüschendorf [48] (see also Embrechts and Puccetti [12]), we call the Fréchet class of all probability measures on $\mathcal{V}$ having marginals μ_1 and μ_2 the Fréchet class with nonoverlapping marginals, denoted as $\mathcal{F}(\mathcal{V}; \mu_1, \mu_2) := \mathcal{F}(\mu_1, \mu_2)$. Note that $\mathcal{F}(\mu_1, \mu_2) = \Pi(\mu_1, \mu_2)$.

Let $g : \mathcal{V} \to \mathbb{R}$ be a measurable function satisfying the following assumption.

Assumption 2. *The function $g : \mathcal{V} \to \mathbb{R}$ is measurable such that $\int_{\mathcal{V}} g d\gamma_0 > -\infty$ for some $\gamma_0 \in \Pi(\mu_1, \mu_2) \subset \mathcal{P}(\mathcal{V})$.*

The marginal problem associated with μ_1 and μ_2 is defined as

$$\mathcal{I}_{\mathrm{M}}(\mu_1, \mu_2) := \sup_{\gamma \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g\, d\gamma.$$

It is essentially an optimal transport problem, where the sup operation is replaced with the inf operation; see Kellerer [31], Rachev and Rüschendorf [45], Villani [52], Villani [53], or Appendix S.1.2 in the Online Supplement for a review of strong duality for $\mathcal{I}_{\mathrm{M}}(\mu_1, \mu_2)$.

The W-DMR with nonoverlapping marginals that we propose extends the marginal problem by allowing each marginal measure of γ to lie in a fixed Wasserstein distance away from a reference measure. Specifically, for any $\gamma \in \mathcal{P}(\mathcal{V})$, let γ_1 and γ_2 denote the projection of γ on $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. The W-DMR with nonoverlapping marginals is defined as

$$\mathcal{I}_{\mathrm{D}}(\delta) := \sup_{\gamma \in \Sigma_{\mathrm{D}}(\delta)} \int_{\mathcal{V}} g\, d\gamma, \quad \delta \in \mathbb{R}_+^2, \tag{2}$$

where $\Sigma_{\mathrm{D}}(\delta)$ is the uncertainty set given by

$$\Sigma_{\mathrm{D}}(\delta) := \Sigma_{\mathrm{D}}(\mu_1, \mu_2, \delta) = \{\gamma \in \mathcal{P}(\mathcal{V}) : \boldsymbol{K}_1(\mu_1, \gamma_1) \le \delta_1, \boldsymbol{K}_2(\mu_2, \gamma_2) \le \delta_2\},$$

in which $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are optimal transport costs associated with cost functions c_1 and c_2, respectively, and $\delta := (\delta_1, \delta_2) \in \mathbb{R}_+^2$ is the radius of the uncertainty set. For generality, we allow the cost functions c_1 and c_2 to be different and also allow δ_1 and δ_2 to be different. Obviously $\Sigma_{\mathrm{D}}(\delta)$ is nonempty for all $\delta \in \mathbb{R}_+^2$.

Remark 2. (i) Under Assumptions 1 and 2, it holds that $\mathcal{I}_{\mathrm{D}}(\delta) > -\infty$ for all $\delta \in \mathbb{R}_+^2$; see Lemma S.3(i) in the Online Supplement. (ii) Under Assumption 1, the uncertainty set $\Sigma_{\mathrm{D}}(0) = \Pi(\mu_1, \mu_2)$ and thus $\mathcal{I}_{\mathrm{D}}(0) = \mathcal{I}_{\mathrm{M}}(\mu_1, \mu_2)$.

2.2.2. Overlapping Marginals.

Let $\mathcal{S} := \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{X}$ be the product space of three Polish spaces $\mathcal{Y}_1$, $\mathcal{Y}_2$, and $\mathcal{X}$. Let $\mathcal{S}_1 := \mathcal{Y}_1 \times \mathcal{X}$ and $\mathcal{S}_2 := \mathcal{Y}_2 \times \mathcal{X}$. Let $\mu_{13} \in \mathcal{P}(\mathcal{S}_1)$ and $\mu_{23} \in \mathcal{P}(\mathcal{S}_2)$ be such that the projection of μ_{13} and the projection of μ_{23} on $\mathcal{X}$ are the same. Following Rüschendorf [48] (see also Embrechts and Puccetti [12]), we call the Fréchet class of all probability measures on $\mathcal{S}$ having marginals μ_{13} and μ_{23} the Fréchet class with overlapping marginals and denote it as $\mathcal{F}(\mathcal{S}; \mu_{13}, \mu_{23}) := \mathcal{F}(\mu_{13}, \mu_{23})$. Unlike the nonoverlapping case, $\mathcal{F}(\mu_{13}, \mu_{23})$ is different from the class of couplings $\Pi(\mu_{13}, \mu_{23})$. For example, for any given measures μ_{13} and μ_{23}, the product measure $\mu_{13} \otimes \mu_{23}$ belongs to $\Pi(\mu_{13}, \mu_{23})$, but does not belong to $\mathcal{F}(\mu_{13}, \mu_{23})$.

Let $f : \mathcal{S} \to \mathbb{R}$ be a measurable function satisfying the following assumption.

Assumption 3. *The function $f : \mathcal{S} \to \mathbb{R}$ is measurable such that $\int_{\mathcal{S}} f d\nu_0 > -\infty$ for some $\nu_0 \in \mathcal{F}(\mu_{13}, \mu_{23}) \subset \mathcal{P}(\mathcal{S})$.*

Rüschendorf [48] studies the following marginal problem with overlapping marginals:

$$\mathcal{I}_{\mathrm{M}}(\mu_{13}, \mu_{23}) := \sup_{\gamma \in \mathcal{F}(\mu_{13}, \mu_{23})} \int_{\mathcal{S}} f \, d\gamma.$$

As shown in Rüschendorf [48], the marginal problem with overlapping marginals can be computed via the marginal problem with nonoverlapping marginals through the following relation:

$$\mathcal{I}(0) = \int_{\mathcal{X}} \left[\sup_{\gamma(\cdot|x) \in \Pi(\mu_{1|3}, \mu_{2|3})} \int_{\mathcal{Y}_1 \times \mathcal{Y}_2} f(y_1, y_2, x) \, d\gamma(y_1, y_2|x) \right] d\gamma_X(x),$$

where γ_X denotes the projection of μ_{13} or μ_{23} onto $\mathcal{X}$, and $\mu_{\ell|3}(\mathrm{d}y_\ell|x)$ denote the conditional probability measures on $\mathcal{X}$ for $\ell \in \{1,2\}$. The inner optimization problem is a marginal problem with nonoverlapping marginals.

For any $\gamma \in \mathcal{P}(\mathcal{S})$, let γ_{13} and γ_{23} denote the projections of γ on $\mathcal{Y}_1 \times \mathcal{X}$ and $\mathcal{Y}_2 \times \mathcal{X}$, respectively. The W-DMR with overlapping marginals is defined as

$$\mathcal{I}(\delta) := \sup_{\gamma \in \Sigma(\delta)} \int_{\mathcal{S}} f \, d\gamma, \quad \delta \in \mathbb{R}_+^2, \tag{3}$$

where $\Sigma(\delta)$ is the uncertainty set given by

$$\Sigma(\delta) := \Sigma(\mu_{13}, \mu_{23}, \delta) = \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{K}_1(\mu_{13}, \gamma_{13}) \le \delta_1, \boldsymbol{K}_2(\mu_{23}, \gamma_{23}) \le \delta_2\},$$

in which $\delta := (\delta_1, \delta_2) \in \mathbb{R}_+^2$ is the radius of the uncertainty set, and $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ are optimal transport costs associated with c_1 and c_2. Similar to the nonoverlapping case, we allow the cost functions c_1 and c_2 to be different and also allow δ_1 and δ_2 to be different. In the examples in Section 2.3, when there is a shift in the distribution of X, different c_1 and c_2 allow us to incorporate potentially different covariances of X and Y_1 (X and Y_2) in the cost function; see Section 6.1.2. We note that $\Sigma(\delta)$ is nonempty for all $\delta \in \mathbb{R}_+^2$.

Remark 3. (i) Assumptions 1 and 3 imply that $\mathcal{I}(\delta) > -\infty$ for all $\delta \ge 0$; see Lemma S.3(ii) in the Online Supplement. (ii) When $\delta = 0$, the uncertainty set $\Sigma(0) = \mathcal{F}(\mu_{13}, \mu_{23})$ and $\mathcal{I}(0) = \mathcal{I}_{\mathrm{M}}(\mu_{13}, \mu_{23})$.

2.3. Motivating Examples

In this section, we present four distinct examples to demonstrate the wide applicability of the W-DMR in marginal problems. The first example is concerned with partial identification of treatment effect parameters when commonly used assumptions in the literature for point identification fail, the second example is concerned with distributionally robust optimal treatment choice, the third one is an application of W-DMR-MP in distributionally robust estimation under data combination, and the last one concerns measures of aggregate risk.

For the first two examples, we adopt the potential outcomes framework for a binary treatment. Let $D \in \{0,1\}$ represent an individual's treatment status, and $Y_1 \in \mathcal{Y}_1 \subset \mathbb{R}$ and $Y_2 \in \mathcal{Y}_2 \subset \mathbb{R}$ denote the potential outcomes under treatments $D = 0$ and $D = 1$, respectively. Let the observed outcome be $Y = DY_2 + (1 - D)Y_1$ and the observed pretreatment covariate be X. Suppose a random sample on (Y, X, D) is available.

To construct the uncertainty set $\Sigma(\delta)$, we choose the reference distributions μ_{13} and μ_{23} as the distributions of (Y_1, X) and of (Y_2, X) identified under Assumption 4 below.

Assumption 4 (Selection-on-Observables).

i. ***Conditional Independence:*** *The potential outcomes are independent of treatment assignment conditional on covariate $X \in \mathcal{X} \subset \mathbb{R}^q$ for $q \ge 1$, that is, $(Y_1, Y_2) \perp\!\!\!\perp D|X$.*

ii. ***Common Support:*** *For all $x \in \mathcal{X}$, $0 < p(x) < 1$, where $p(x) := \mathbb{P}(D = 1|X = x)$.*

Under Assumption 4, the conditional distribution functions of Y_1, Y_2 given $X = x$ are point identified from the sample information:

$$F_{Y_1|X}(y|x) = \mathbb{P}(Y_1 \le y|X = x) = \mathbb{P}(Y \le y|X = x, D = 0) \text{ and}$$
$$F_{Y_2|X}(y|x) = \mathbb{P}(Y_2 \le y|X = x) = \mathbb{P}(Y \le y|X = x, D = 1).$$

2.3.1. Partial Identification of Treatment Effects. Assumption 4 is commonly used to identify treatment effect parameters and optimal treatment choice. However, the validity of Assumption 4 may be questionable when there are unobserved confounders. W-DMR-MP presents a viable approach to studying robustness of causal inference to deviations from Assumption 4 by varying the joint distribution of (Y_1, Y_2, X) in the Wasserstein uncertainty set centered at the reference measures consistent with Assumption 4.

Formally, let f be a measurable function of Y_1, Y_2. Consider treatment effects of the form $\theta_o := \mathbb{E}_o[f(Y_1, Y_2)]$, where $\mathbb{E}_o$ denotes expectation with respect to the true distribution of (Y_1, Y_2). It includes the average treatment effect (ATE) for which $f(Y_1, Y_2) = Y_2 - Y_1$ and the distributional treatment effect such as $\mathbb{P}_o(Y_2 - Y_1 \geq 0)$, where $\mathbb{P}_o$ denotes the probability computed under the true distribution of (Y_1, Y_2).

Let

$$\Sigma(\delta) = \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{K}_1(\mu_{13}, \gamma_{13}) \leq \delta_1, \boldsymbol{K}_2(\mu_{23}, \gamma_{23}) \leq \delta_2\},$$

where μ_{13} and μ_{23} are the identified distributions under Assumption 4. Suppose the true distribution of (Y_1, Y_2, X) lies in the uncertainty set $\Sigma(\delta)$ for some δ. Then the identified set for θ_o is given by

$$\Theta(\delta) := \left\{ \int_{\mathcal{S}} f(y_1, y_2)\, d\gamma(y_1, y_2, x) : \gamma \in \Sigma(\delta) \right\}.$$

Under mild conditions, we show in Proposition 1 that the identified set $\Theta(\delta)$ is a closed interval given by

$$\Theta(\delta) = \left[\min_{\gamma \in \Sigma(\delta)} \int_{\mathcal{S}} f(y_1, y_2)\, d\gamma(s), \max_{\gamma \in \Sigma(\delta)} \int_{\mathcal{S}} f(y_1, y_2)\, d\gamma(s) \right],$$

where the lower and upper limits of the interval are characterized by the W-DMR-MP.[6] When $\delta = 0$, $\Theta(0)$ reduces to the characterization in Fan et al. [22].

Remark 4. The choice of the uncertainty set depends on the application of interest. Our objective is to assess stability/robustness of ATE to the violation of the selection-on-observables assumption. So, we construct our uncertainty set for the distribution of (Y_1, Y_2, X) such that the reference distributions for (Y_1, X) and (Y_2, X) are the ones identified under the selection-on-observables assumption.

Cheridito and Eckstein [9, section 4.1] imply that ATE is only continuous with respect to the causal optimal transport distance for the joint distribution of (Y, X, D), where Y is the observed outcome. Our formulation is different, because our uncertainty set is based on optimal transport distances for the distribution of (Y_1, X) and the distribution of (Y_2, X), where Y_1 and Y_2 are potential outcomes. In fact, we can show that when Y_1 and Y_2 are bounded, the following inequalities hold:

$$|\mathbb{E}_{\mu_{13}}[Y_1] - \mathbb{E}_{\nu_{13}}[Y_1]| \leq c\boldsymbol{W}_1(\mu_{13}, \nu_{13}) \text{ and } |\mathbb{E}_{\mu_{23}}[Y_2] - \mathbb{E}_{\nu_{23}}[Y_2]| \leq c\boldsymbol{W}_1(\mu_{23}, \nu_{23})$$

for some positive constant c. Consequently,

$$|\mathbb{E}_{\mu}[Y_1 - Y_2] - \mathbb{E}_{\nu}[Y_1 - Y_2]| \leq c\Big(\boldsymbol{W}_1(\mu_{13}, \nu_{13}) + \boldsymbol{W}_1(\mu_{23}, \nu_{23})\Big).$$

However, this does not contradict the result in Cheridito and Eckstein [9, section 4.1].

2.3.2. Robust Welfare Function. In empirical welfare maximization (EWM), an optimal choice/policy is chosen to maximize the expected welfare estimated from a training data set and then applied to a target population; see Kitagawa and Tetenov [34]. EWM assumes that the target population and the training data set come from the same underlying probability measure. This may not be valid in important applications. Motivated by designing externally valid treatment policy, Adjaho and Christensen [1] introduce a robust welfare function which allows the target population to differ from the training population. In this paper, we revisit Adjaho and Christensen's [1] robust welfare function and propose a new one based on W-DMR with overlapping marginals.

Adjaho and Christensen [1] adopt the following definition of a robust welfare function:

$$\mathrm{RW}_0(d) := \inf_{\gamma \in \Sigma_0(\delta_0)} \mathbb{E}_{\gamma}[Y_1(1 - d(X)) + Y_2 d(X)],$$

where $d : \mathcal{X} \to \{0, 1\}$ is a measurable policy function, that is, $d(X)$ is zero or one depending on X and $\Sigma_0(\delta_0)$ is the Wasserstein uncertainty set centered at a joint measure μ for (Y_1, Y_2, X) consistent with Assumption 4, that is,

$$\Sigma_0(\delta_0) := \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{K}_c(\mu, \gamma) \leq \delta_0\},$$

where $\boldsymbol{K}_c(\mu, \gamma)$ is the optimal transport cost with cost function $c : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+ \cup \{\infty\}$. Our robust welfare function

is defined as

$$\mathrm{RW}(d) := \inf_{\gamma\in\Sigma(\delta)} \mathbb{E}_\gamma[Y_1(1-d(X)) + Y_2 d(X)],$$

where $\Sigma(\delta) = \Sigma(\mu_{13}, \mu_{23}, \delta)$ is the uncertainty set for W-DMR with overlapping marginals.

The joint reference distribution μ of the uncertainty set $\Sigma_0(\delta_0)$ is unidentifiable under Assumption 4, because of the inherent missing-data nature of causal inference. Consequently, Adjaho and Christensen [1] suggest imposing either perfect negative or perfect positive dependence between $\mu_{1|3}$ and $\mu_{2|3}$ when constructing a joint reference distribution. In contrast, our new robust welfare function relies only on the marginal reference distributions μ_{13} and μ_{23}, both identified under Assumption 4.

2.3.3. W-DRO Under Data Combination. An important application of W-DMR is W-DRO. Let $f : \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{X} \times \Theta \to \mathbb{R}$ be a loss function with an unknown parameter $\theta \in \Theta \subset \mathbb{R}^q$. W-DRO under data combination is defined as

$$\min_{\theta\in\Theta} \sup_{\gamma\in\Sigma(\delta)} \int_{\mathcal{S}} f(y_1, y_2, x; \theta)\, d\gamma(y_1, y_2, x),$$

where $\Sigma(\delta)$ is the uncertainty set for the overlapping case. For each $\theta \in \Theta$, the inner optimization is a W-DMR with overlapping marginals. In practice, we need to choose the reference measures μ_{13} and μ_{23} based on the sample information. Focusing on the logit model, where $\mathcal{Y}_1 = \{+1, -1\}$ is the space for the dependent variable, and $\mathcal{Y}_2$ and $\mathcal{X}$ are feature spaces/covariate space, and

$$f(y_1, y_2, x; \theta) = \log(1 + \exp(-y_1\langle\theta, (y_2, x)\rangle)),$$

Awasthi et al. [2] propose a method dubbed "Robust Data Join" in which the empirical measures constructed from the two data sets are used as reference measures. Specifically, let $\hat{\mu}_{13}$ and $\hat{\mu}_{23}$ denote empirical measures based on two separate data sets. The uncertainty set in Awasthi et al. [2] takes the following form:

$$\Sigma_{\mathrm{RDJ}}(\delta) := \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{K}_1(\hat{\mu}_{13}, \gamma_{13}) \le \delta_1, \boldsymbol{K}_2(\hat{\mu}_{23}, \gamma_{23}) \le \delta_2\},$$

where

$$c_1((y_1, x), (y_1', x')) = \|x - x'\|_p + \kappa_1 |y_1 - y_1'| \quad \text{and}$$
$$c_2((y_2, x), (y_2, x')) = \|x - x'\|_p + \kappa_2 \|y_2 - y_2'\|_{p'}$$

with $\kappa_1 \ge 1$, $\kappa_2 \ge 1$, $p \ge 1$, and $p' \ge 1$.

Note that the "Robust Data Join" of Awasthi et al. [2] is different from our W-DMR with nonoverlapping marginals because the measure of interest $\gamma \in \mathcal{P}(\mathcal{S})$ has overlapping marginals. It is also different from our W-DMR with overlapping marginals because the reference measures $\hat{\mu}_{13}$ and $\hat{\mu}_{23}$ may not have overlapping marginals. Unlike the uncertainty set for W-DMR, $\Sigma_{RDJ}(\delta)$ may be empty when $\delta = 0$. This occurs when $\hat{\mu}_{13}$ and $\hat{\mu}_{23}$, estimated from separate data sets, do not have identical projections on the overlapping space $\mathcal{X}$. In this case, the constraints $\boldsymbol{K}_1(\hat{\mu}_{13}, \gamma_{13}) = 0$ and $\boldsymbol{K}_2(\hat{\mu}_{23}, \gamma_{23}) = 0$ cannot hold simultaneously, as γ_{13} and γ_{23} have the same marginal measure on $\mathcal{X}$.

2.3.4. Risk Aggregation. Let S_1, S_2 be random variables representing individual risks defined on Polish spaces $\mathcal{S}_1, \mathcal{S}_2$, respectively. Let μ_1, μ_2 be probability measures of S_1, S_2. Let $\mathcal{V} = \mathcal{S}_1 \times \mathcal{S}_2$ and $g : \mathcal{V} \to \mathbb{R}$ be a risk-aggregating function. Applying W-DMR with nonoverlapping marginals to the risk aggregation function g, we can compute the worst aggregate risk when the joint measure of the individual risks varies in the uncertainty set $\Sigma_{\mathrm{D}}(\delta)$. This is different from the set-up in Eckstein et al. [11]. Given a reference measure $\mu \in \Pi(\mu_1, \mu_2)$, they consider the following robust risk aggregation problem:

$$\mathcal{I}_\Pi(\delta_0) := \sup_{\gamma\in\Sigma_\Pi(\delta)} \int_{\mathcal{V}} g\, d\gamma,$$

where

$$\Sigma_\Pi(\delta_0) := \{\gamma \in \Pi(\mu_1, \mu_2) : \boldsymbol{K}_c(\gamma, \mu) \le \delta_0\},$$

in which $\boldsymbol{K}_c$ is the optimal transport cost associated with a cost function $c : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_+$. Because $\gamma \in \Sigma_\Pi(\delta_0)$ is a coupling of (μ_1, μ_2), we have that $\Sigma_\Pi(\delta_0) \subset \Sigma_{\mathrm{D}}(0)$ and thus $\mathcal{I}_\Pi(\delta_0) \le \mathcal{I}_{\mathrm{D}}(0)$.

3. Strong Duality and Distributionally Robust Makarov Bounds

In this section, we establish strong duality for our W-DMR-MP and apply it to develop Wasserstein distributionally robust Makarov bounds.

3.1. Nonoverlapping Marginals

For a measurable function $g : \mathcal{V} \to \mathbb{R}$ and $\lambda := (\lambda_1, \lambda_2) \in \mathbb{R}^2_+$, we define the function $g_\lambda : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ as

$$g_\lambda(v) := \sup_{v' \in \mathcal{V}} \varphi_\lambda(v, v'),$$

where $\varphi_\lambda : \mathcal{V} \times \mathcal{V} \to \mathbb{R} \cup \{-\infty\}$ is given by

$$\varphi_\lambda(v, v') = g(s_1', s_2') - \lambda_1 c_1(s_1, s_1') - \lambda_2 c_2(s_2, s_2'),$$

with $v := (s_1, s_2)$ and $v' := (s_1', s_2')$. Similarly, define $g_{\lambda_1,1} : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ and $g_{\lambda_2,2} : \mathcal{V} \to \mathbb{R} \cup \{+\infty\}$ as

$$g_{\lambda_1,1}(s_1, s_2) = \sup_{s_1' \in \mathcal{S}_1} \{g(s_1', s_2) - \lambda_1 c_1(s_1, s_1')\} \quad \text{and}$$

$$g_{\lambda_2,2}(s_1, s_2) = \sup_{s_2' \in \mathcal{S}_2} \{g(s_1, s_2') - \lambda_2 c_2(s_2, s_2')\}.$$

The dual problem $\mathcal{J}_{\mathrm{D}}(\delta)$ corresponding to the primal problem $\mathcal{I}_{\mathrm{D}}(\delta)$ is defined as follows:[7]

$$\mathcal{J}_{\mathrm{D}}(\delta) = \begin{cases} \inf_{\lambda \in \mathbb{R}^2_+} \left\{ \langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g_\lambda \, d\varpi \right\} & \text{if } \delta \in \mathbb{R}^2_{++}, \\ \inf_{\lambda_1 \in \mathbb{R}_+} \left\{ \lambda_1 \delta_1 + \sup_{\varpi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g_{\lambda_1,1} \, d\varpi \right\} & \text{if } \delta_1 > 0 \text{ and } \delta_2 = 0, \\ \inf_{\lambda_2 \in \mathbb{R}_+} \left\{ \lambda_2 \delta_2 + \sup_{\varpi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g_{\lambda_2,2} \, d\varpi \right\} & \text{if } \delta_1 = 0 \text{ and } \delta_2 > 0. \end{cases} \tag{4}$$

Theorem 2. *Suppose that* Assumptions 1 and 2 *hold. Then,* $\mathcal{I}_{\mathrm{D}}(\delta) = \mathcal{J}_{\mathrm{D}}(\delta)$ *for all* $\delta \in \mathbb{R}^2_+ \setminus \{0\}$.

Unlike the dual for W-DMR, the dual for W-DMR with nonoverlapping marginals in Theorem 2 involves a marginal problem with nonoverlapping marginals μ_1, μ_2 due to the lack of knowledge on the dependence of the joint measure μ. Computational algorithms developed for optimal transport can be used to solve the marginal problem; see Peyré and Cuturi [43]. For empirical measures μ_1, μ_2, the marginal problem is a discrete optimal transport problem and there are efficient algorithms to compute it; see Peyré and Cuturi [43]. For general measures μ_1, μ_2, strong duality may be employed in the numerical computation of the marginal problem. For instance, consider the case when $\delta > 0$. When $g_\lambda(v)$ is Borel measurable, several strong duality results are available; see, for example, Villani [52] and Villani [53]. For a general function g and cost functions c_1, c_2, $g_\lambda(v)$ is not guaranteed to be Borel measurable. However, for Polish spaces, the set $\{v \in \mathcal{V} : g_\lambda(v) \geq u\}$ is an analytic set for all $u \in \overline{\mathbb{R}}$ (and g_λ is universally measurable), because g, c_1, and c_2 are Borel measurable (see Blanchet and Murthy [5, p. 580] and Bertsekas and Shreve [4, lemma 7.22, lemma 7.30(i), and proposition 7.47]). This allows us to apply strong duality for the marginal problem in Kellerer [31] restated in Theorem S.1 in the Online Supplement to the marginal problem involving $g_\lambda(v)$; see Corollary S.1 in the Online Supplement.

Without additional assumptions on the function g and the cost functions, the dual $\mathcal{J}_{\mathrm{D}}(\delta)$ in Theorem 2 for interior points $\delta \in \mathbb{R}^2_{++}$ and the dual for boundary points may not be the same. To illustrate, plugging in $\delta_2 = 0$ in the dual form for interior points in Theorem 2, we obtain

$$\inf_{\lambda_1 \in \mathbb{R}_+} \left[\lambda_1 \delta_1 + \inf_{\lambda_2 \in \mathbb{R}_+} \sup_{\varpi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g_\lambda \, d\varpi \right].$$

It is different from the dual $\mathcal{J}_{\mathrm{D}}(\delta_1, 0)$ for $\delta_1 > 0$, because

$$\inf_{\lambda_2 \in \mathbb{R}_+} \sup_{\varpi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g_\lambda \, d\varpi \neq \sup_{\varpi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g_{\lambda_1,1} \, d\varpi.$$

When the function g and the cost functions satisfy assumptions in Theorem 8, the dual $\mathcal{J}_D(\delta)$ in Theorem 2 for interior points $\delta \in \mathbb{R}^2_{++}$ and the dual for boundary points are the same so that

$$\mathcal{I}_D(\delta) = \inf_{\lambda \in \mathbb{R}^2_+} \left[\langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g_\lambda \, d\varpi \right],$$

for all $\delta \in \mathbb{R}^2_+$.

Remark 5.

i. For Polish spaces, Theorem 2 generalizes the strong duality in Zhang et al. [57] restated in Theorem 1. Our proof is based on that in Zhang et al. [57]. However, because of the presence of two marginal measures in the uncertainty set $\Sigma_D(\delta)$, we need to verify the existence of a joint measure when some of its overlapping marginal measures are fixed, and we rely on existing results for a given consistent product marginal system studied in Vorob'ev [55], Kellerer [30], and Shortt [50]; see Appendix S.1.3 in the Online Supplement for a detailed review.

ii. Suppose that the assumptions of Theorem 2 hold, and c_2 is a real-valued function. Then for any $\delta_1 \geq 0$, one can show that

$$\lim_{\delta_2 \to \infty} \mathcal{I}_D(\delta) = \sup_{\gamma_1 : \boldsymbol{K}_1(\mu_1, \gamma_1) \leq \delta_1} \int \left[\sup_{s_2 \in \mathcal{S}_2} g(s_1, s_2) \right] d\gamma_1(s_1),$$

where the expression on the right-hand side of the above equation is the classical W-DMR with uncertainty set $\{\gamma_1 \in \mathcal{P}(\mathcal{S}_1) : \boldsymbol{K}_1(\mu_1, \gamma_1) \leq \delta_1\}$. In this case, Theorem 2 reduces to Theorem 1 where the loss function is given by $\sup_{s_2 \in \mathcal{S}_2} g(s_1, s_2)$. The proof is included in Appendix S.3 of the Online Supplement.

Remark 6.

i. Similar to Sinha et al. [51], for W-DMR in marginal problems, we can define an alternative W-DMR through linear penalty terms, that is,

$$\sup_{\gamma \in \mathcal{P}(\mathcal{V})} \left\{ \int_{\mathcal{V}} g d\gamma - \lambda_1 \boldsymbol{K}_1(\mu_1, \gamma_1) - \lambda_2 \boldsymbol{K}_2(\mu_2, \gamma_2) : \boldsymbol{K}_\ell(\mu_\ell, \gamma_\ell) < \infty \text{ for } \ell = 1, 2 \right\}, \tag{5}$$

with $\lambda_1, \lambda_2 \in \mathbb{R}_{++}$. The proof of Theorem 2 implies that the dual form of this problem is $\sup_{\varpi \in \Pi(\mu_1, \mu_2)} \int g_\lambda d\varpi$ under the condition in Theorem 2.

ii. As pointed out by an anonymous referee, one can consider W-DMR in marginal problems with a more general penalty given by

$$\sup_{\gamma \in \mathcal{P}(\mathcal{V})} \left\{ \int_{\mathcal{V}} g d\gamma - \varphi(\boldsymbol{K}_1(\mu_1, \gamma_1), \boldsymbol{K}_2(\mu_2, \gamma_2)) \, : \, \boldsymbol{K}_\ell(\mu_\ell, \gamma_\ell) < \infty \text{ for } \ell = 1, 2 \right\},$$

where φ is a convex function. This formulation for standard W-DMR has been studied by Bartl et al. [3], Jiang [28], and Eckstein et al. [11]. In contrast to the linear penalty in Equation (5), the proof of duality for the general penalty requires additional steps beyond that of Theorem 2; see Theorem S.4 and its proof in Appendix S.4 in the Online Supplement for details.

3.2. Overlapping Marginals

Let $\phi_\lambda : \mathcal{V} \times \mathcal{S} \to \mathbb{R} \cup \{-\infty\}$ be

$$\phi_\lambda(v, s') \quad := f(s') - \lambda_1 c_1(s_1, s'_1) - \lambda_2 c_2(s_2, s'_2),$$

where $v = (s_1, s_2)$, $s' = (y'_1, y'_2, x')$, $s'_\ell = (y'_\ell, x')$, and $s_\ell = (y_\ell, x_\ell)$. Define the function $f_\lambda : \mathcal{V} \to \overline{\mathbb{R}}$ associated with f as

$$f_\lambda(v) := \sup_{s' \in \mathcal{S}} \phi_\lambda(v, s').$$

Similarly, we define $f_{\lambda,1} : \mathcal{V} \to \overline{\mathbb{R}}$ and $f_{\lambda,2} : \mathcal{V} \to \overline{\mathbb{R}}$ as follows:

$$f_{\lambda_1,1}(s_1, s_2) = \sup_{y'_1 \in \mathcal{Y}_1} \{ f(y'_1, y_2, x_2) - \lambda_1 c_1((y_1, x_1), (y'_1, x_2)) \} \quad \text{and}$$

$$f_{\lambda_2,2}(s_1, s_2) = \sup_{y'_2 \in \mathcal{Y}_2} \{ f(y_1, y'_2, x_1) - \lambda_2 c_2((y_2, x_2), (y'_2, x_1)) \},$$

in which $s_1 = (y_1, x_1)$ and $s_2 = (y_2, x_2)$. The dual problem $\mathcal{J}(\delta)$ corresponding to the primal problem $\mathcal{I}(\delta)$ is defined

as follows:

$$\mathcal{J}(\delta) = \begin{cases} \inf_{\lambda \in \mathbb{R}^2_+} \left\{ \langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{V}} f_\lambda d\varpi \right\} & \text{if } \delta \in \mathbb{R}^2_{++}, \\ \inf_{\lambda_1 \in \mathbb{R}_+} \left\{ \lambda_1 \delta_1 + \sup_{\varpi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{V}} f_{\lambda_1,1} d\varpi \right\} & \text{if } \delta_1 > 0 \text{ and } \delta_2 = 0, \\ \inf_{\lambda_2 \in \mathbb{R}_+} \left\{ \lambda_2 \delta_2 + \sup_{\varpi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{V}} f_{\lambda_2,2} d\varpi \right\} & \text{if } \delta_1 = 0 \text{ and } \delta_2 > 0. \end{cases} \tag{6}$$

Theorem 3. *Suppose that* Assumptions 1 and 3 *hold. Then,* $\mathcal{I}(\delta) = \mathcal{J}(\delta)$ *for all* $\delta \in \mathbb{R}^2_+ \setminus \{0\}$.

An interesting feature of the dual for overlapping marginals is that it involves marginal problems with nonoverlapping marginals, that is, $\sup_{\varpi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{V}} f_\lambda(v) d\varpi(v)$, although the uncertainty set in the primal problem involves overlapping marginals. Compared with the nonoverlapping marginals case, overlapping marginals in the uncertainty set make the relevant consistent product marginal system in the verification of the existence of a joint measure more complicated; see the proof of Lemma A.5 in the appendix. Nonetheless, the nonoverlapping marginals in the dual allow us to apply Theorem S.1 in the Online Supplement to the marginal problem involving $f_\lambda, f_{\lambda,1}$, and $f_{\lambda,2}$; see Corollary S.2 in the Online Supplement.

Under the assumptions in Theorem 9, we have

$$\mathcal{I}(\delta) = \inf_{\lambda \in \mathbb{R}^2_+} \left[\langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{V}} f_\lambda \, d\varpi \right]$$

for all $\delta \in \mathbb{R}^2_+$.

Remark 7. Similar to the nonoverlapping case, we can also define an alternative W-DMR with overlapping marginals through linear penalty terms, that is,

$$\sup_{\gamma \in \mathcal{P}(\mathcal{S})} \left\{ \int_{\mathcal{S}} g \, d\gamma - \lambda_1 \boldsymbol{K}_1(\mu_{13}, \gamma_{13}) - \lambda_2 \boldsymbol{K}_2(\mu_{23}, \gamma_{23}) : \boldsymbol{K}_\ell(\mu_{\ell 3}, \gamma_{\ell 3}) < \infty \text{ for } \ell = 1, 2 \right\},$$

with $\lambda_1, \lambda_2 \in \mathbb{R}_{++}$. The proof of Theorem 3 implies that the dual form of this problem is $\sup_{\varpi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{V}} f_\lambda \, d\varpi$ under the conditions in Theorem 3. The general penalty case is discussed in Appendix S.4 in the Online Supplement; see that for details.

3.3. Wasserstein Distributionally Robust Makarov Bounds

Let $\mathcal{S}_1 = \mathbb{R}$, $\mathcal{S}_2 = \mathbb{R}$, $\mu_1 \in \mathcal{P}(\mathcal{S}_1)$, and $\mu_2 \in \mathcal{P}(\mathcal{S}_2)$. Further, let $Z = S_1 + S_2$, where S_1, S_2 are random variables whose distributions are μ_1 and μ_2, respectively. For a given $z \in \mathbb{R}$, let $F_Z(z) = \mathbb{E}_o[g(S_1, S_2)]$, where $g(s_1, s_2) = \mathbb{1}\{s_1 + s_2 \le z\}$.

Sharp bounds on the quantile function $F_Z^{-1}(\cdot)$ are established in Makarov [36] and referred to as the Makarov bounds. Inverting the Makarov bounds leads to sharp bounds on the distribution function $F_Z(z)$; see Rüschendorf [47] and Frank et al. [24]. They are given by

$$\inf_{\gamma \in \Pi(\mu_1, \mu_2)} \mathbb{E}_\gamma[g(S_1, S_2)] = \sup_{x \in \mathbb{R}} \max\{\mu_1(x) + \mu_2(z - x) - 1, 0\} \text{ and}$$

$$\sup_{\gamma \in \Pi(\mu_1, \mu_2)} \mathbb{E}_\gamma[g(S_1, S_2)] = 1 + \inf_{x \in \mathbb{R}} \min\{\mu_1(x) + \mu_2(z - x) - 1, 0\},$$

where with a slight abuse of notation, $\mu_j(t_o) := \mu_j(\{t' \in \mathbb{R} : t' \le t_o\})$ for all $t_o \in \mathbb{R}$ and $j = 1, 2$. Because the quantile bounds first established in Makarov [36] and the above distribution bounds are equivalent, we also refer to the latter as Makarov bounds. Makarov bounds have been successfully applied in distinct areas. For example, the upper bound on the quantile of Z is known as the worst VaR of Z; see Embrechts et al. [13] and Embrechts et al. [14]. Makarov bounds are also used to study partial identification of distributional treatment effects when the treatment assignment mechanism identifies the marginal measures of the potential outcomes such as in Assumption 4; see Fan and Park [18], Fan and Park [19], Fan and Park [20], Fan and Wu [21], Fan et al. [22], Ridder and Moffitt [46], and Firpo and Ridder [23].

Let $g(s_1, s_2) = \mathbb{1}(s_1 + s_2 \le z)$ and $c_\ell(s_\ell, s'_\ell) = |s_\ell - s'_\ell|^2$ for $\ell = 1, 2$. Theorem 3 implies the following corollary.

Corollary 1 (Wasserstein Distributionally Robust Makarov Bounds). *For all* $\delta \in \mathbb{R}^2_+$,

$$\sup_{\gamma\in\Sigma_D(\delta)} \mathbb{E}_\gamma[g(S_1,S_2)] = \inf_{\lambda\in\mathbb{R}^2_+}\left(\langle\lambda,\delta\rangle + \sup_{\varpi\in\Pi(\mu_1,\mu_2)}\left[\int_{\{s_1+s_2>z\}}\left[1-\frac{\lambda_1\lambda_2(s_1+s_2-z)^2}{\lambda_1+\lambda_2}\right]^+ d\varpi(s_1,s_2) + \mathbb{E}_\varpi[\mathbb{1}\{S_1+S_2\le z\}]\right)\right. \text{ and}$$

$$\inf_{\gamma\in\Sigma_D(\delta)} \mathbb{E}_\gamma[g(S_1,S_2)] = \sup_{\lambda\in\mathbb{R}^2_+}\left[-\langle\lambda,\delta\rangle + \inf_{\varpi\in\Pi(\mu_1,\mu_2)}\left\{-\int_{\{s_1+s_2\le z\}}\left[1-\frac{\lambda_1\lambda_2(s_1+s_2-z)^2}{\lambda_1+\lambda_2}\right]^+ d\varpi(s_1,s_2) + \mathbb{E}_\varpi[\mathbb{1}\{S_1+S_2\le z\}]\right.\right].$$

Suppose that the true joint distribution of S_1, S_2 belongs to the uncertainty set $\Sigma_D(\delta)$ for some $\delta \in \mathbb{R}^2_+$. Corollary 1 extends Makarov bounds to allow for possible misspecification of the marginal distributions of S_1, S_2 by μ_1, μ_2 respectively. We call the resulting bounds Wasserstein distributionally robust Makarov bounds.

We note that $g_\lambda(v)$ is bounded and continuous in v, and convex in λ, and $\Pi(\mu_1,\mu_2)$ is compact. Applying the minimax theorem of Fan [17, theorem 2], we can interchange the order of inf and sup in the dual in the above corollary and get

$$\sup_{\gamma\in\Sigma_D(\delta)} \mathbb{E}_\gamma[g(S_1,S_2)] = \sup_{\varpi\in\Pi(\mu_1,\mu_2)}\left[\inf_{\lambda\in\mathbb{R}^2_+}\left(\langle\lambda,\delta\rangle + \int_{\{s_1+s_2>z\}}\left[1-\frac{\lambda_1\lambda_2(s_1+s_2-z)^2}{\lambda_1+\lambda_2}\right]^+ d\varpi(s_1,s_2)\right) + \mathbb{E}_\varpi[\mathbb{1}\{S_1+S_2\le z\}]\right].$$

This expression is very insightful, where the inner infimum term characterizes possible deviations of the true marginal measures from the reference measures.

4. Finiteness of the W-DMR-MP and Existence of Optimizers

In this section, we assume that all the reference measures belong to appropriate Wasserstein spaces and prove finiteness of the W-DMR-MP and existence of an optimizer.

Definition 3 (Wasserstein Space). The Wasserstein space of order $p \ge 1$ on a Polish space $\mathcal{X}$ with metric $\boldsymbol{d}$ is defined as

$$\mathcal{P}_p(\mathcal{X}) = \left\{\mu \in \mathcal{P}(\mathcal{X}) : \int_{\mathcal{X}} \boldsymbol{d}(x_0,x)^p d\mu(x) < \infty\right\},$$

where $x_0 \in \mathcal{X}$ is arbitrary.

Assumption 5.

i. *In the nonoverlapping case, we assume that* $\mu_1 \in \mathcal{P}_{p_1}(\mathcal{S}_1)$ *and* $\mu_2 \in \mathcal{P}_{p_2}(\mathcal{S}_2)$ *for some* $p_1 \ge 1$ *and* $p_2 \ge 1$;
ii. *In the overlapping case, we assume that* $\mu_{13} \in \mathcal{P}_{p_1}(\mathcal{S}_1)$ *and* $\mu_{23} \in \mathcal{P}_{p_2}(\mathcal{S}_2)$ *for some* $p_1 \ge 1$ *and* $p_2 \ge 1$.

Assumption 6. *The cost function* $c_\ell : \mathcal{S}_\ell \times \mathcal{S}_\ell \to \mathbb{R} \cup \{\infty\}$ *is of the form* $c_\ell(s_\ell, s'_\ell) = \boldsymbol{d}_{\mathcal{S}_\ell}(s_\ell, s'_\ell)^{p_\ell}$, *where* $(\mathcal{S}_\ell, \boldsymbol{d}_{\mathcal{S}_\ell})$ *is a Polish space and* $p_\ell \ge 1$ *for* $\ell = 1, 2$.

4.1. Finiteness of the W-DMR-MP

For the nonoverlapping case, we establish the following result.

Theorem 4. *Suppose that* Assumptions 2, 5(i), and 6 *hold. Then for all* $\delta \in \mathbb{R}^2_{++}$, $\mathcal{I}_D(\delta) < \infty$ *if and only if there exist* $v^\star := (s_1^\star, s_2^\star) \in \mathcal{V}$ *and a constant* $M > 0$ *such that for all* $(s_1, s_2) \in \mathcal{V}$,

$$g(s_1,s_2) \le M[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1)^{p_1} + \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star, s_2)^{p_2}], \tag{7}$$

where p_1 *and* p_2 *are defined in Assumption* 5(i).

The inequality in Equation (7) is a growth condition on the function g. It extends the growth condition in Yue et al. [56] for W-DMR to our W-DMR with nonoverlapping marginals.

For the overlapping case, the following result holds.

Theorem 5. *Suppose that Assumptions* 3, 5(ii), *and* 6 *hold. Then for all* $\delta \in \mathbb{R}^2_{++}$, $\mathcal{I}(\delta) < \infty$ *if and only if there exist* $(s_1^\star, s_2^\star) \in \mathcal{S}_1 \times \mathcal{S}_2$ *and a constant* $M > 0$ *such that*

$$f(s) \le M[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1)^{p_1} + \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star, s_2)^{p_2}], \tag{8}$$

for all $s \in \mathcal{S}$, *where* $s := (y_1, y_2, x), s_\ell := (y_\ell, x)$ *and* $s_\ell^\star := (y_\ell^\star, x^\star)$ *for* $\ell = 1, 2$, *and* p_1 *and* p_2 *are defined in Assumption* 5(ii).

The growth condition (8) on the function f extends the growth condition in Yue et al. [56] for W-DMR. When

$$\boldsymbol{d}_{\mathcal{S}_\ell}((y_\ell, x), (y'_\ell, x')) = \boldsymbol{d}_{\mathcal{Y}_\ell}(y_\ell, y'_\ell) + \boldsymbol{d}_{\mathcal{X}}(x, x'),$$

Condition (8) is satisfied if and only if there exist $s^\star := (y_1^\star, y_2^\star, x^\star)$ and a constant $M > 0$ such that

$$f(s) \le M[1 + \boldsymbol{d}_{\mathcal{Y}_1}(y_1, y_1^\star)^{p_1} + \boldsymbol{d}_{\mathcal{Y}_2}(y_2, y_2^\star)^{p_2} + \boldsymbol{d}_{\mathcal{X}}(x, x^\star)^{p_1 \wedge p_2}],$$

for all $s = (y_1, y_2, x) \in \mathcal{S}$.

Remark 8. The conditions in Theorems 4 and 5 are sufficient conditions for $\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ to be finite for all $\delta \in \mathbb{R}_+^2$ including boundary points because $\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ are nondecreasing.

4.2. Existence of Optimizers

Definition 4. A metric space $(\mathcal{X}, \boldsymbol{d})$ is said to be proper if for any $r > 0$ and $x_0 \in \mathcal{X}$, the closed ball $\overline{B}(x_0, r) := \{x \in \mathcal{X} : \boldsymbol{d}(x, x_0) \le r\}$ is compact.

Examples of proper metric spaces include finite dimensional Banach spaces and complete Riemannian manifolds; see Yue et al. [56].

Assumption 7. *$(\mathcal{S}_1, \boldsymbol{d}_{\mathcal{S}_1})$ and $(\mathcal{S}_2, \boldsymbol{d}_{\mathcal{S}_2})$ are proper.*

Assumptions 5, 6, and 7 imply that $\Sigma_{\mathrm{D}}(\delta)$ and $\Sigma(\delta)$ are weakly compact for all $\delta \in \mathbb{R}_+^2$; see Propositions A.1 and A.2 in the appendix. Given weak compactness of the uncertainty sets $\Sigma_{\mathrm{D}}(\delta)$ and $\Sigma(\delta)$, it is sufficient to show that the mapping: $\gamma \to \int g d\gamma$ is upper semicontinuous over $\gamma \in \Sigma_{\mathrm{D}}(\delta)$ for the nonoverlapping case, and the mapping: $\gamma \to \int f d\gamma$ is upper semicontinuous over $\gamma \in \Sigma(\delta)$ for the overlapping case. In Theorems 6 and 7 below, we provide conditions for g and f ensuring upper semicontinuity of each map and thus the existence of optimal solutions for $\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ for all $\delta \in \mathbb{R}_+^2$.

Theorem 6. *Suppose that Assumptions 2, 5(i), 6, and 7 hold. Further, assume that g is upper semicontinuous, and there exist a constant $M > 0$, $v^\star := (s_1^\star, s_2^\star) \in \mathcal{V}$ and $p'_\ell \in (0, p_\ell)$ for $\ell = 1, 2$, such that*

$$g(v) \le M[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1)^{p'_1} + \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star, s_2)^{p'_2}], \tag{9}$$

for all $v := (s_1, s_2) \in \mathcal{V}$. Then an optimal solution of (2) exists for all $\delta \in \mathbb{R}_+^2$.

Theorem 7. *Suppose that Assumptions 3, 5(ii), 6, and 7 hold. Further, assume that f is upper semicontinuous, and there exist $(s_1^\star, s_2^\star) \in \mathcal{S}_1 \times \mathcal{S}_2$, a constant $M > 0$, $p'_\ell \in (0, p_\ell)$ for $\ell = 1, 2$, such that*

$$f(s) \le M[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1)^{p'_1} + \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star, s_2)^{p'_2}], \tag{10}$$

for all $s \in \mathcal{S}$ where $s := (y_1, y_2, x), s_\ell := (y_\ell, x)$, and $s_\ell^\star := (y_\ell^\star, x_\ell^\star)$ for $\ell = 1, 2$. Then an optimal solution of (3) exists for all $\delta \in \mathbb{R}_+^2$.

Remark 9. Theorems 6 and 7 state the existence of an optimizer for every $\delta \in \mathbb{R}_+^2$. Assumption 7 might not be required for $\Sigma_{\mathrm{D}}(\delta)$ and $\Sigma(\delta)$ to be weakly compact for some $\delta \in \mathbb{R}_+^2$. However, observation 1 in Yue et al. [56] implies that the properness is necessary for the Wasserstein ball to be weakly compact for every $\delta \in \mathbb{R}_+^2$. We provide counterexamples demonstrating the nonexistence of optimizers for both the nonoverlapping and overlapping cases when properness does not hold. For the nonoverlapping case, consider $\mathcal{S}_1 = \mathcal{S}_2 = \mathbb{R}$, where the metric d_j on $\mathcal{S}_j$ is defined as

$$d_j(s_j, s'_j) = \|s_j - s'_j\| \wedge 1,$$

where $\|\cdot\|$ denotes the Euclidean norm. Under this metric, when $\delta_1, \delta_2 \ge 1$, $\Sigma_{\mathrm{D}}(\mu, \delta) = \mathcal{P}(\mathcal{V})$ is not weakly compact for any reference measure $\mu \in \mathcal{P}(\mathcal{V})$; see remark 6.19 in Villani [52]. Let $g(s_1, s_2) = \exp(|s_1| + |s_2|) / [1 + \exp(|s_1| + |s_2|)]$. It is straightforward to verify that the growth condition holds and g is upper semicontinuous, ensuring that the optimal value $\mathcal{I}_{\mathrm{D}}(\delta)$ is finite, but optimizers fail to exist when $\delta_1, \delta_2 \ge 1$.

For the overlapping case, let $\mathcal{Y}_1 = \mathcal{Y}_1 = \mathcal{X} = \mathbb{R}$, and define the metric on $\mathcal{S}_\ell = \mathcal{Y}_\ell \times \mathcal{X}$ as $d(s_j, s'_j) = \|s_j - s'_j\| \wedge 1$. Under this metric, when $\delta_1, \delta_2 \ge 1$, $\Sigma_{\mathrm{D}}(\delta) = \mathcal{P}(\mathcal{S})$ is not weakly compact, where $\mathcal{S} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \mathcal{X}$. Let $f(s) = \exp(\|s\|) / [1 + \exp(\|s\|)]$, where $s = (y_1, y_2, x) \in \mathbb{R}^3$. Similar to the nonoverlapping case, when $\delta_1, \delta_2 \ge 1$, the optimal value $\mathcal{I}(\delta)$ is finite but the optimizer does not exist.

4.3. Characterization of Identified Sets

In some applications, such as the partial identification of treatment effects introduced in Section 2.3.1, the identified sets of $\theta_{Do} := \mathbb{E}_o[g(S_1, S_2)]$ and $\theta_o := \mathbb{E}_o[f(S)]$ are of interest, where S is a random variable whose distribution belongs to $\Sigma(\delta)$, and S_1 and S_2 are random variables whose joint probability distribution belongs to $\Sigma_D(\delta)$. They are

$$\Theta_{\mathrm{D}}(\delta) := \left\{ \int_{\mathcal{S}_1 \times \mathcal{S}_2} g \, d\gamma : \gamma \in \Sigma_{\mathrm{D}}(\delta) \right\} \text{ and } \Theta(\delta) := \left\{ \int_{\mathcal{S}} f \, d\gamma : \gamma \in \Sigma(\delta) \right\}.$$

By applying finiteness and existence results, we show below that under mild conditions, the identified sets $\Theta_{\mathrm{D}}(\delta)$ and $\Theta(\delta)$ are both closed intervals.

Proposition 1.

i. *Suppose Assumptions 5(i), 6, and 7 hold. In addition, g is continuous, and $|g|$ satisfies Condition (9). Then, for $\delta \in \mathbb{R}^2_+$, we have*

$$\Theta_{\mathrm{D}}(\delta) = \left[\min_{\gamma \in \Sigma_{\mathrm{D}}(\delta)} \int_{\mathcal{S}_1 \times \mathcal{S}_2} g \, d\gamma, \max_{\gamma \in \Sigma_{\mathrm{D}}(\delta)} \int_{\mathcal{S}_1 \times \mathcal{S}_2} g \, d\gamma \right],$$

where both the lower and upper bounds are finite.

ii. *Suppose Assumptions 5(ii), 6, and 7 hold. In addition, f is continuous and $|f|$ satisfies Condition (10). Then for $\delta \in \mathbb{R}^2_+$, we have*

$$\Theta(\delta) = \left[\min_{\gamma \in \Sigma(\delta)} \int_{\mathcal{S}} f \, d\gamma, \max_{\gamma \in \Sigma(\delta)} \int_{\mathcal{S}} f \, d\gamma \right],$$

where both the lower and upper bounds are finite.

The strong duality in Section 3 can be used to evaluate the lower and upper bounds.

5. Continuity of the DMR-MP Functions

In this section, we establish continuity of the W-DMR-MP functions $\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ for all $\delta \in \mathbb{R}^2_+$ under conditions similar to those in Zhang et al. [57]. Compared with Zhang et al. [57], our analysis is more involved, because the boundary in our case includes not only the origin $(0,0)$ but also $(\delta_1, 0)$ and $(0, \delta_2)$ for all $\delta_1 > 0$ and $\delta_2 > 0$.

5.1. Nonoverlapping Marginals

Theorem S.3(i) in the Online Supplement implies that under Assumptions 1 and 2, $\mathcal{I}_{\mathrm{D}}(\delta)$ is a concave function for $\delta \in \mathbb{R}^2_+$ and hence is continuous on $\mathbb{R}^2_{++}$. We provide the main assumption for the continuity of $\mathcal{I}_{\mathrm{D}}(\delta)$ on $\mathbb{R}^2_+$ in this subsection.

Assumption 8. *Let $\Psi : \mathbb{R}^2_+ \to \mathbb{R}_+$ be a continuous, nondecreasing, and concave function with $\Psi(0,0) = 0$. Suppose the function $g : \mathcal{V} \to \mathbb{R}$ satisfies*

$$g(v) - g(v') \le \Psi(c_1(s_1, s_1'), c_2(s_2, s_2')), \tag{11}$$

for all $v = (s_1, s_2) \in \mathcal{V}$ and $v' = (s_1', s_2') \in \mathcal{V}$.

The function Ψ in Assumption 8 plays the role of the modulus of continuity of g. To illustrate, consider the following example.

Example 1. Suppose Assumption 6 holds, that is, $c_\ell(s_\ell, s_\ell') = \boldsymbol{d}_{\mathcal{S}_\ell}(s_\ell, s_\ell')^{p_\ell}$ for some $p_\ell \ge 1$, $\ell = 1, 2$.

i. Define a product metric $\boldsymbol{d}_{\mathcal{V}}$ on $\mathcal{V} = \mathcal{S}_1 \times \mathcal{S}_2$ as

$$\boldsymbol{d}_{\mathcal{V}}((s_1, s_2), (s_1', s_2')) = \boldsymbol{d}_{\mathcal{S}_1}(s_1, s_1') + \boldsymbol{d}_{\mathcal{S}_2}(s_2, s_2').$$

Let $\Psi(x, y) = x^{1/p_1} + y^{1/p_2}$. Then, $\boldsymbol{d}_{\mathcal{V}}((s_1, s_2), (s_1', s_2')) = \Psi(c_1(s_1, s_1'), c_2(s_2, s_2'))$. On the metric space $(\mathcal{V}, \boldsymbol{d}_{\mathcal{V}})$, the function g is continuous and has $\omega : x \longmapsto x$ as the modulus of continuity. Moreover, Assumption 8 implies the growth condition in Theorem 6.

ii. Suppose $p_1 = p_2$. Define a product metric $\boldsymbol{d}_{\mathcal{V}}$ on $\mathcal{V} = \mathcal{S}_1 \times \mathcal{S}_2$ as

$$\boldsymbol{d}_{\mathcal{V}}((s_1, s_2), (s_1', s_2')) = [\boldsymbol{d}_{\mathcal{S}_1}(s_1, s_1')^p + \boldsymbol{d}_{\mathcal{S}_2}(s_2, s_2')^p]^{1/p}.$$

Let $\Psi(x,y) = (x+y)^{1/p}$. Then, $\boldsymbol{d}_{\mathcal{V}}((s_1,s_2),(s_1',s_2')) = \Psi(c_1(s_1,s_1'),c_2(s_2,s_2'))$. On the metric space $(\mathcal{V},\boldsymbol{d}_{\mathcal{V}})$, the function g is continuous and has $\omega : x \longmapsto x$ as the modulus of continuity. Assumption 8 also implies the growth condition in Theorem 6.

iii. Suppose $p_1 \neq p_2$. Define a product metric $\boldsymbol{d}_{\mathcal{V}}$ on $\mathcal{V} = \mathcal{S}_1 \times \mathcal{S}_2$ as

$$\boldsymbol{d}_{\mathcal{V}}((s_1,s_2),(s_1',s_2')) = \boldsymbol{d}_{\mathcal{S}_1}(s_1,s_1') \vee \boldsymbol{d}_{\mathcal{S}_2}(s_2,s_2').$$

Then, Assumption 8 implies

$$g(v) - g(v') \leq \Psi(\boldsymbol{d}_{\mathcal{V}}(v,v'),\boldsymbol{d}_{\mathcal{V}}(v,v')) = \omega(\boldsymbol{d}_{\mathcal{V}}(v,v')),$$

where $\omega : x \longmapsto \Psi(x,x)$ is a concave function. On the metric space $(\mathcal{V},\boldsymbol{d}_{\mathcal{V}})$, the function g is continuous and has $\omega : x \longmapsto \Psi(x,x)$ as the modulus of continuity.

Theorem 8. *Suppose Assumptions 1, 2, and 8 hold and $\mathcal{I}_{\mathrm{D}}(\delta) < \infty$ for some $\delta > 0$. Then, the function $\mathcal{I}_{\mathrm{D}}(\delta)$ is continuous on $\mathbb{R}^2_+$.*

Two implications follow. First, under Assumptions 1 and 2,

$$\mathcal{I}_{\mathrm{D}}(0) = \sup_{\gamma \in \Pi(\mu_1,\mu_2)} \int_{\mathcal{V}} g \, d\gamma.$$

Continuity facilitates stability/robustness analysis as δ approaches zero. Second, under the assumptions in Theorem 8, we have

$$\mathcal{I}_{\mathrm{D}}(\delta) = \inf_{\lambda \in \mathbb{R}^2_+} \left[\langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\mu_1,\mu_2)} \int_{\mathcal{V}} g_\lambda \, d\varpi \right]$$

for all $\delta \in \mathbb{R}^2_+$. As a result, the dual $\mathcal{J}_{\mathrm{D}}(\delta)$ in (4) is continuous for all $\delta \in \mathbb{R}^2_+$.

5.2. Overlapping Marginals

Lemma S.3(ii) in the Online Supplement implies that under Assumptions 1 and 3, $\mathcal{I}(\delta)$ is a concave function for $\delta \in \mathbb{R}^2_+$ and hence is continuous on $\mathbb{R}^2_{++}$. We provide the main assumption for the continuity of $\mathcal{I}(\delta)$ on $\mathbb{R}^2_+$ below.

To simplify the technical analysis, we maintain Assumption 6 in this section. Because the metrics in $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are not specified, we introduce an auxiliary function ρ_ℓ from $\mathcal{Y}_\ell \times \mathcal{Y}_\ell$ to $\mathbb{R}_+$ induced by the cost function c_ℓ, $\ell = 1,2$.

Assumption 9. *For $\ell = 1,2$, there exists a function ρ_ℓ from $\mathcal{Y}_\ell \times \mathcal{Y}_\ell$ to $\mathbb{R}_+$ such that*

i. *ρ_ℓ is symmetric, that is, $\rho_\ell(y_\ell,y_\ell') = \rho_\ell(y_\ell',y_\ell)$ for all $y_\ell,y_\ell' \in \mathcal{Y}_\ell$;*

ii. *there is $q_\ell \in [1,p_\ell]$ such that $\rho_\ell(y_\ell,y_\ell') \leq \boldsymbol{d}_{\mathcal{S}_\ell}(s_\ell,s_\ell')^{q_\ell}$ for all $s_\ell \equiv (y_\ell,x) \in \mathcal{S}_\ell$ and $s_\ell' \equiv (y_\ell',x') \in \mathcal{S}_\ell$;*

iii. *there is a constant $N > 0$ such that $\rho_\ell(y_\ell,y_\ell') \leq N[\rho_\ell(y_\ell,y_\ell^\star) + \rho_\ell(y_\ell^\star,y_\ell')]$ for all $y_\ell,y_\ell',y_\ell^\star \in \mathcal{Y}_\ell$.*

We now introduce the main assumption on f.

Assumption 10. *For $\ell = 1,2$, let $\Psi_\ell : \mathbb{R}^2_+ \to \mathbb{R}_+$ be continuous, nondecreasing, and concave satisfying $\Psi_\ell(0,0) = 0$. Suppose for all $s = (y_1,y_2,x)$ and $s' = (y_1',y_2',x')$, it holds that*

$$f(y_1,y_2,x) - f(y_1',y_2',x') \leq \Psi_1(c_1(s_1,s_1'),\rho_2(y_2,y_2')) \text{ and}$$
$$f(y_1,y_2,x) - f(y_1',y_2',x') \leq \Psi_2(\rho_1(y_1,y_1'),c_2(s_2,s_2')).$$

Like Assumption 8, Assumption 10 depends on the cost functions c_1,c_2. It also depends on the auxiliary functions ρ_1,ρ_2. The functions Ψ_1,Ψ_2 play the role of the modulus of continuity.

Example 2 ($\boldsymbol{p_j}$-Product Metric). Let $(\mathcal{Y}_1,\boldsymbol{d}_{\mathcal{Y}_1})$, $(\mathcal{Y}_2,\boldsymbol{d}_{\mathcal{Y}_2})$, and $(\mathcal{X},\boldsymbol{d}_{\mathcal{X}})$ be Polish (metric) spaces. For $p_\ell \geq 1$, define the p_ℓ-product metric on $\mathcal{S}_\ell$ as

$$\boldsymbol{d}_{\mathcal{S}_\ell}(s_\ell,s_\ell') = [\boldsymbol{d}_{\mathcal{Y}_\ell}(y_\ell,y_\ell')^{p_\ell} + \boldsymbol{d}_{\mathcal{X}}(x,x')^{p_\ell}]^{1/p_\ell}.$$

Let

$$\rho_\ell(y_\ell,y_\ell') := \inf_{x_\ell,x_\ell' \in \mathcal{X}} \boldsymbol{d}_{\mathcal{S}_\ell}((y_\ell,x_\ell),(y_\ell',x_\ell'))^{p_\ell}.$$

It is easy to show that $\rho_\ell(y_\ell, y'_\ell) = \boldsymbol{d}_{\mathcal{Y}_\ell}(y_\ell, y'_\ell)^{p_\ell}$ and Assumption 9 is satisfied with $N = 2^{p_\ell}$. Moreover, Assumption 10 reduces to

$$f(y_1, y_2, x) - f(y'_1, y'_2, x') \le \Psi_1(\boldsymbol{d}_{\mathcal{S}_1}(s_1, s'_1)^{p_1}, \boldsymbol{d}_{\mathcal{Y}_2}(y_2, y'_2)^{p_2}) \quad \text{and}$$
$$f(y_1, y_2, x) - f(y'_1, y'_2, x') \le \Psi_2(\boldsymbol{d}_{\mathcal{Y}_1}(y_1, y'_1)^{p_1}, \boldsymbol{d}_{\mathcal{S}_2}(s_2, s'_2)^{p_2}).$$

When $p_1 = p_2 = p$, Assumption 10 may be reduced to a simpler form. To see this, define two functions ψ_1 and ψ_2 from $\mathbb{R}^3$ to $\mathbb{R}^2$ as $\psi_1 : (z_1, z_2, z) \longmapsto (z_1 + z, z_2)$ and $\psi_2 : (z_1, z_2, z) \longmapsto (z_1, z_2 + z)$. We can see that

$$\Psi_1(\boldsymbol{d}_{\mathcal{S}_1}(s_1, s'_1)^p, \rho_2(y_1, y'_1)^p) = \Psi_1 \circ \psi_1(\boldsymbol{d}_{\mathcal{Y}_1}(y_1, y'_1)^p, \boldsymbol{d}_{\mathcal{Y}_2}(y_2, y'_2)^p, \boldsymbol{d}_{\mathcal{X}}(x, x')^p),$$
$$\Psi_2(\rho_1(y_1, y'_1)^p, \boldsymbol{d}_{\mathcal{S}_2}(s_2, s'_2)^p) = \Psi_2 \circ \psi_2(\boldsymbol{d}_{\mathcal{Y}_1}(y_1, y'_1)^p, \boldsymbol{d}_{\mathcal{Y}_2}(y_2, y'_2)^p, \boldsymbol{d}_{\mathcal{X}}(x, x')^p).$$

Because ψ_j is linear, $\Phi_j = \Psi_j \circ \psi_j$ is still continuous, nondecreasing, and concave. Assumption 10 is reduced to the following condition:

$$f(y_1, y_2, x) - f(y'_1, y'_2, x') \le \Phi_j(\boldsymbol{d}_{\mathcal{Y}_1}(y_1, y'_1)^p, \boldsymbol{d}_{\mathcal{Y}_2}(y_2, y'_2)^p, \boldsymbol{d}_{\mathcal{X}}(x, x')^p)$$

for all $(y_1, y_2, x) \in \mathcal{S}$ and $(y'_1, y'_2, x') \in \mathcal{S}$.

Theorem 9. *Suppose Assumptions 3, 5(ii), 6, 9, and 10 hold, and $\mathcal{I}(\delta) < \infty$ for some $\delta > 0$. Then the function $\mathcal{I}(\delta)$ is continuous on $\mathbb{R}^2_+$.*

Like the nonoverlapping case, two implications follow. First, under Assumptions 1 and 2,

$$\mathcal{I}(0) = \sup_{\gamma \in \mathcal{F}(\mu_{13}, \mu_{23})} \int_{\mathcal{S}} f \, d\gamma.$$

Continuity facilitates stability/robustness analysis as δ approaches zero. Second, under the assumptions in Theorem 9, we have

$$\mathcal{I}(\delta) = \inf_{\lambda \in \mathbb{R}^2_+} \left[\langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{V}} f_\lambda \, d\varpi \right]$$

for all $\delta \in \mathbb{R}^2_+$. As a result, the dual $\mathcal{J}(\delta)$ in (6) is continuous for all $\delta \in \mathbb{R}^2_+$.

6. Motivating Examples Revisited

In this section, we apply the results in Sections 3–5 to the examples introduced in Section 2.

6.1. Partial Identification of Treatment Effects

In addition to characterizing $\Theta(\delta)$ introduced in Section 2, we also study the identified set for $\theta_{Do} = \mathbb{E}_o[f(Y_1, Y_2)]$ without using the covariate information:

$$\Theta_{\mathrm{D}}(\delta) := \left\{ \int_{\mathcal{Y}_1 \times \mathcal{Y}_2} f(y_1, y_2) \, d\gamma(y_1, y_2) : \gamma \in \Sigma_{\mathrm{D}}(\delta) \right\},$$

where

$$\Sigma_{\mathrm{D}}(\delta) = \{\gamma \in \mathcal{P}(\mathcal{Y}_1 \times \mathcal{Y}_2) : \boldsymbol{K}_{Y_1}(\mu_{Y_1}, \gamma_1) \le \delta_1, \boldsymbol{K}_{Y_1}(\mu_{Y_2}, \gamma_2) \le \delta_2\},$$

in which $\boldsymbol{K}_{Y_1}$ and $\boldsymbol{K}_{Y_2}$ are the optimal transport costs associated with cost functions c_{Y_1} and c_{Y_2}, respectively.

6.1.1. Characterization of the Identified Sets. When f is continuous and conditions in Proposition 1 are satisfied, the identified sets $\Theta_{\mathrm{D}}(\delta)$ and $\Theta(\delta)$ are both closed intervals with upper limits given by W-DMR for nonoverlapping and overlapping marginals, respectively. This allows us to apply our duality results in Section 3 to evaluate and compare $\Theta_{\mathrm{D}}(\delta)$ and $\Theta(\delta)$.

Let $\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ denote the upper bounds of $\Theta_{\mathrm{D}}(\delta)$ and $\Theta(\delta)$, respectively, where

$$\mathcal{I}_{\mathrm{D}}(\delta) = \sup_{\gamma\in\Sigma_{\mathrm{D}}(\delta)} \int_{\mathcal{Y}_1\times\mathcal{Y}_2} f(y_1,y_2)\, d\gamma(y_1,y_2) \text{ and}$$

$$\mathcal{I}(\delta) = \sup_{\gamma\in\Sigma(\delta)} \int_{\mathcal{S}} f(y_1,y_2)\, d\gamma(y_1,y_2,x).$$

Proposition 1 establishes robust versions of existing results on the identified sets of treatment effects under Assumption 4; see Fan et al. [22]. Robustness to deviations from Assumption 4 can be examined via $\Theta_{\mathrm{D}}(\delta)$ and $\Theta(\delta)$ by varying δ. For example, when f satisfies assumptions in Theorems 8 and 9, $\mathcal{I}(\delta)$ and $\mathcal{I}_{\mathrm{D}}(\delta)$ are continuous on $\mathbb{R}^2_+$. As a result,

$$\lim_{\delta\to 0}\mathcal{I}(\delta) = \mathcal{I}(0) \quad \text{and} \quad \lim_{\delta\to 0}\mathcal{I}_{\mathrm{D}}(\delta) = \mathcal{I}_{\mathrm{D}}(0).$$

For a general function f, the lower and upper limits of the identified sets $\Theta_{\mathrm{D}}(\delta)$ and $\Theta(\delta)$ need to be computed numerically. When f is additively separable, we show that duality results in Section 3 simplify the evaluation of $\Theta_D(\delta)$ and $\Theta(\delta)$. Because the lower bounds of $\Theta_D(\delta)$ and $\Theta(\delta)$ can be computed in a similar way by applying duality to $-f(y_1,y_2)$, we omit details for the lower bounds.

Assumption 11. *Let $f : (y_1,y_2,x) \longmapsto f_1(y_1)+f_2(y_2)$ from $\mathcal{S}$ to $\mathbb{R}$, where $f_\ell \in L^1(\mu_{\ell 3})$ for $\ell = 1,2$.*

To avoid tedious notation, we also treat f as a function from $\mathcal{Y}_1\times\mathcal{Y}_2$ to $\mathbb{R}$. Under Assumptions 1 and 11, it is easy to show that

$$\begin{aligned}\mathcal{I}_{\mathrm{D}}(\delta) &= \sup_{\gamma_1:\boldsymbol{K}_{Y_1}(\mu_{Y_1},\gamma_1)\le\delta_1}\int_{\mathcal{Y}_1} f_1\, d\gamma_1 + \sup_{\gamma_2:\boldsymbol{K}_{Y_2}(\mu_{Y_2},\gamma_2)\le\delta_2}\int_{\mathcal{Y}_2} f_2\, d\gamma_2\\ &= \inf_{\lambda_1\ge 0}\left[\lambda_1\delta_1 + \int_{\mathcal{Y}_1}(f_1)_{\lambda_1} d\mu_1\right] + \inf_{\lambda_2\ge 0}\left[\lambda_2\delta_2 + \int_{\mathcal{Y}_2}(f_2)_{\lambda_2} d\mu_2\right],\end{aligned}$$

where $(f_\ell)_{\lambda_\ell} : \mathcal{Y}_\ell \to \mathbb{R}$ is given by

$$(f_\ell)_{\lambda_\ell}(y_\ell) = \sup_{y'_\ell\in\mathcal{Y}_\ell}\{f_\ell(y'_\ell) - \lambda_\ell c_{Y_\ell}(y_\ell,y'_\ell)\}.$$

That is, when f is an additively separable function, the W-DMR for nonoverlapping marginals is the sum of two W-DMRs associated with the marginals regardless of the cost functions.

Depending on the cost functions, the W-DMR for overlapping marginals may be different from the sum of two W-DMRs associated with the marginals.

Definition 5 (Refer to Chen et al. [8]). We say that a function $\boldsymbol{f} : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is separable if each $\boldsymbol{x}$ and $\boldsymbol{y}$ can be optimized regardless of the other variable. In other words,

$$\arg\min_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f(x,y) = \left[\arg\min_{x\in\mathcal{X}} f(x,y')\right] \times \left[\arg\min_{y\in\mathcal{Y}} f(x',y)\right],$$

for any $x'\in\mathcal{X}$ and $y'\in\mathcal{Y}$.

Assumption 12. *For $\ell = 1,2$, the cost function $c_\ell((y_\ell,x_\ell),(y'_\ell,x'_\ell))$ is separable with respect to (y_ℓ,y'_ℓ) and (x_ℓ,x'_ℓ).*

Example 3. Let $a_\ell : \mathcal{Y}_\ell\times\mathcal{Y}_\ell\to\mathbb{R}_+\cup\{\infty\}$ and $b_\ell : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}_+\cup\{\infty\}$ satisfy Assumption 1. Let $s=(y,x)$ and $s'=(y',x')$. Then $c(s,s') = a(y,y')+b(x,x')$ is separable with respect to (x,x') and (y,y'). Also, both $c(s,s') = (a(y,y')+1)(b(x,x')+1)-1$ and $c(s,s') = [a(y,y')^p + b(x,x')^p]^{1/p}$ for $p\ge 1$ are separable with respect to (x,x') and (y,y') even though they are not additively separable.

Proposition 2. *For $\ell = 1,2$, let $c_\ell : (\mathcal{Y}_\ell\times\mathcal{X})\times(\mathcal{Y}_\ell\times\mathcal{X})\to\mathbb{R}_+$ denote the cost function for $\Theta(\delta)$. Suppose that c_ℓ satisfies Assumption 1 and the marginal measure of $\mu_{\ell 3}$ on $\mathcal{Y}_\ell$ coincides with μ_ℓ, that is, $\mu_{\ell,3} = \mathrm{Law}(Y_\ell,X)$ with $\mu_\ell = \mathrm{Law}(Y_\ell)$. Under Assumptions 11 and 12, one has $\mathcal{I}(\delta) = \mathcal{I}_{\mathrm{D}}(\delta)$, where $\mathcal{I}_{\mathrm{D}}(\delta)$ is based on the cost function c_{Y_ℓ} on $\mathcal{Y}_\ell\times\mathcal{Y}_\ell$ given by*

$$c_{Y_\ell}(y_\ell,y'_\ell) = \inf_{x_\ell,x'_\ell\in\mathcal{X}} c_\ell((y_\ell,x_\ell),(y'_\ell,x'_\ell)).$$

It is easy to verify that $c_{Y_\ell}(y_\ell,y'_\ell) = 0$ if and only if $y_\ell = y'_\ell$.

This proposition implies that for separable cost functions, the W-DMR for overlapping marginals equals the W-DMR for nonoverlapping marginals with cost function $c_{Y_\ell}(y_\ell, y'_\ell)$. As a result, the covariate information does not help shrink the identified set.

6.1.2. Identified Sets for Average Treatment Effect. Suppose $f(y_1, y_2) = y_2 - y_1$ and $c_\ell((y,x),(y_\ell,x_\ell)) = |y - y'|^2 + \|x_\ell - x'_\ell\|^2$ for $\ell = 1,2$. Let $\tau_{ATE} = \mathbb{E}[Y_2 - Y_1]$. Then Proposition 2 implies that the upper bound on τ_{ATE} is given by

$$\mathcal{I}(\delta) = \mathcal{I}_{\mathrm{D}}(\delta) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + \sqrt{\delta_1} + \sqrt{\delta_2}.$$

In this section, we demonstrate that when Assumption 12 is violated, the W-DMR for overlapping marginals may be smaller than the W-DMR for nonoverlapping marginals and, as a result, $\Theta(\delta)$ is a proper subset of $\Theta_D(\delta)$.

Consider the squared Mahalanobis distance with respect to a positive definite matrix. That is,

$$c_\ell(s_\ell, s'_\ell) = (s_\ell - s'_\ell)^\top V_\ell^{-1}(s_\ell - s'_\ell),$$

where $V_\ell = \begin{pmatrix} V_{\ell,YY} & V_{\ell,YX} \\ V_{\ell,XY} & V_{\ell,XX} \end{pmatrix}$ is a positive definite matrix. It is easy to show that

$$\begin{aligned} c_{Y_\ell}(y_\ell, y'_\ell) &= \min_{x_\ell, x'_\ell \in \mathcal{X}'_\ell} c_\ell(s_\ell, s'_\ell) \\ &= (y_\ell - y'_\ell)^\top V_{\ell,YY}^{-1}(y_\ell - y'_\ell), \end{aligned}$$

where $s_\ell = (y_\ell, x_\ell)$ and $s'_\ell = (y'_\ell, x'_\ell)$.

Proposition 3. *Let $\mathcal{I}$ be the primal of the overlapping W-DMR problem under*

$$c_\ell(s_\ell, s'_\ell) = (s_\ell - s'_\ell)^\top V_\ell^{-1}(s_\ell - s'_\ell).$$

Let $\mathcal{I}_{\mathrm{D}}$ be the primal of the nonoverlapping W-DMR problem under $c_{Y_\ell}(y_\ell, y'_\ell)$. Assume that $\mathbb{E}\|X|\|_2^2 < \infty$, $\mathbb{E}|Y_1|^2 < \infty$, and $\mathbb{E}|Y_2|^2 < \infty$. Then, $\mathcal{I}(\delta) \le \mathcal{I}_{\mathrm{D}}(\delta)$ for all $\delta > 0$.

Proposition 4. *Suppose that all the conditions in Proposition 3 hold. Then,*

i. *for all $\delta \in \mathbb{R}^2_+$,*

$$\mathcal{I}_{\mathrm{D}}(\delta) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + V_{1,YY}^{1/2}\delta_1^{1/2} + V_{2,YY}^{1/2}\delta_2^{1/2},$$

$$\mathcal{I}(\delta) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + \inf_{\lambda \in \mathbb{R}^2_{++}} \left\{ \lambda_1\delta_1 + \lambda_2\delta_2 + \frac{1}{4\lambda_1}(V_1/V_{1,XX}) + \frac{1}{4\lambda_2}(V_2/V_{2,XX}) + \frac{1}{4}V_o^\top(\lambda_1 V_{1,XX}^{-1} + \lambda_2 V_{2,XX}^{-1})^{-1}V_o \right\},$$

where $V_\ell/V_{\ell,XX} := V_{\ell,YY} - V_{\ell,YX}V_{\ell,XX}^{-1}V_{\ell,XY}$ is the Schur complement of $V_{\ell,XX}$ in V_ℓ for $\ell = 1,2$, and $V_o = V_{2,XX}^{-1}V_{2,XY} - V_{1,XX}^{-1}V_{0,XY}$;

i. *$\mathcal{I}_{\mathrm{D}}(\delta) = \mathcal{I}(\delta)$ for all $\delta \in \mathbb{R}^2_+$ if and only if $V_{1,XY} = V_{2,XY} = 0$;*

ii. *$\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ are continuous on $\mathbb{R}^2_+$.*

Propositions 3 and 4 imply that for nonseparable Mahalanobis cost functions, the information in covariates may help shrink the identified set because $\mathcal{I}_{\mathrm{D}}(\delta) < \mathcal{I}(\delta)$ for some δ under mild conditions. Proposition 4 also implies that (i) $\mathcal{I}(0) = \mathcal{I}_{\mathrm{D}}(0) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1]$ and (ii) $\mathcal{I}(\delta_1, 0) = \mathcal{I}_{\mathrm{D}}(\delta_1, 0)$ and $\mathcal{I}(0, \delta_2) = \mathcal{I}_{\mathrm{D}}(0, \delta_2)$ for all $\delta_1 \ge 0$ and $\delta_2 \ge 0$.

6.2. Comparison of Robust Welfare Functions

Recall that

$$\begin{aligned} \mathrm{RW}_0(d) &:= \inf_{\gamma \in \Sigma_0(\delta)} \mathbb{E}[Y_1(1 - d(X)) + Y_2 d(X)] \quad \text{and} \\ \mathrm{RW}(d) &:= \inf_{\gamma \in \Sigma(\delta)} \mathbb{E}[Y_1(1 - d(X)) + Y_2 d(X)], \end{aligned}$$

where

$$\begin{aligned} \Sigma_0(\delta_0) &= \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{K}(\mu, \gamma) \le \delta_0\} \quad \text{and} \\ \Sigma(\delta) &= \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{K}_\ell(\mu_{\ell,3}, \gamma_{\ell,3}) \le \delta_\ell, \ \forall \ell = 1,2\}. \end{aligned}$$

Consider the following cost function c_ℓ for $\ell = 1,2$:

$$c_\ell(s_\ell, s'_\ell) = c_{Y_\ell}(y_\ell, y'_\ell) + b(x, x'),$$

where $s_\ell = (y_\ell, x_\ell)$, $s'_\ell = (y'_\ell, x'_\ell)$, and $c_{Y_1}(y_1, y'_1)$ and $c_{Y_2}(y_2, y'_2)$ are cost functions for Y_1 and Y_2, respectively, and $b(x, x')$ is some function on the space $\mathcal{X}$ satisfying Assumption 1. When $b(x, x') = \infty \mathbb{1}\{x \neq x'\}$, $\mathbb{P}(X = X') = 1$ for any probability measure in the uncertainty set.

Adjaho and Christensen [1] establish strong duality for $\mathrm{RW}_0(d)$ under several cost functions. For comparison purposes, we restate the following proposition in Adjaho and Christensen [1] which allows distributional shifts in covariate X.

Proposition 5 (Adjaho and Christensen [1, Proposition 4.1])**.** *Suppose Y_1 and Y_2 are unbounded and $\mathbb{E}\|X\|_2^2$ is finite. Let the cost function $c : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ be given by*

$$c(s, s') = |y_1 - y'_1| + |y_2 - y'_2| + \|x' - x\|_2,$$

for $s = (y_1, y_2, x)$ and $s' = (y'_1, y'_2, x')$. Then

$$\mathrm{RW}_0(d) = \sup_{\eta \geq 1} \{\mathbb{E}_\mu[\max\{Y_2 + \eta h_1(X), Y_1 + \eta h_0(X)\}] - \eta \delta_0\},$$

where $h_0(x) = \inf_{u \in \mathcal{X}: d(u) = 0} \|x - u\|_2$ and $h_1(x) = \inf_{u \in \mathcal{X}: d(u) = 1} \|x - u\|_2$.

This proposition implies that $\mathrm{RW}_0(d)$ depends on the choice of the reference measure μ. Because only the marginals μ_{13} and μ_{23} are identified under Assumption 4, Adjaho and Christensen [1] suggest three possible choices for μ by imposing specific dependence structures on μ:

- Y_1 and Y_2 are perfectly positively dependent conditional on $X = x$;
- Y_1 and Y_2 are conditionally independent given $X = x$;
- Y_1 and Y_2 are perfectly negatively dependent conditional on $X = x$.

Section 4.3.1 in Adjaho and Christensen [1] shows that their robust welfare function $\mathrm{RW}_0(d)$ is minimized when Y_1 and Y_2 are perfectly negatively dependent conditional on $X = x$.

The following proposition evaluates $\mathrm{RW}(d)$ via the duality result in Section 3 and compares it with $\mathrm{RW}_0(d)$.

Proposition 6. *Consider $c_\ell(s_\ell, s'_\ell) = |y_\ell - y'_\ell| + \|x_\ell - x'_\ell\|_2$. Assume that Y is unbounded and $\mathbb{E}|Y_1|$, $\mathbb{E}|Y_2|$, and $\mathbb{E}\|X\|_2^2$ are finite. Then,*

i. *the robust welfare function $\mathrm{RW}(d)$ based on $\Sigma(\delta)$ has the following dual reformulation:*

$$\mathrm{RW}(d) = \sup_{\lambda \geq 1} \left[\inf_{\pi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{V}} \min\{y_2 + \varphi_{\lambda,1}(x_1, x_2), y_1 + \varphi_{\lambda,0}(x_1, x_2)\} d\pi(v) - \langle \lambda, \delta \rangle \right],$$

where $v = (y_1, x_1, y_2, x_2)$, and

$$\varphi_{\lambda,0}(x_1, x_2) = \min_{x' : d(x') = 0} (\lambda_1 \|x_1 - x'\|_2 + \lambda_2 \|x_2 - x'\|_2),$$
$$\varphi_{\lambda,1}(x_1, x_2) = \min_{x' : d(x') = 1} (\lambda_1 \|x_1 - x'\|_2 + \lambda_2 \|x_2 - x'\|_2);$$

ii. *When $\delta_0 = \delta_1 = \delta_2$, $\mathrm{RW}(d) \leq \mathrm{RW}_0^*(d)$, where $\mathrm{RW}_0^*(d)$ is the robust welfare function $\mathrm{RW}_0(d)$ based on the reference measure $\pi^* = \int \max\{\mu_{1|3} + \mu_{2|3} - 1, 0\} d\mu_3$.*

Part (ii) of the above proposition implies that $\mathrm{RW}(d) \leq \mathrm{RW}_0(d)$ for any reference measure $\mu \in \mathcal{F}(\mu_{13}, \mu_{23})$.

6.3. W-DRO for Logit Model Under Data Combination

We revisit the logit model in Section 2.3.3 and make the following assumption.

Assumption 13. (i) *Let (Y_1, Y_2, X) follow some unknown measure μ. Let D denote a binary random variable independent of (Y_1, Y_2, X) such that we observe (Y_1, X) when $D = 0$ and (Y_2, X) when $D = 1$.* (ii) *Let $\{Y_{1i}, X_{1i}\}_{i=1}^{n_1}$ be the data set from (Y_1, X), and $\{Y_{2i}, X_{2i}\}_{i=1}^{n_2}$ be the data set from (Y_2, X).*

Under this assumption, $X|D = 1$ has the same distribution as $X|D = 0$ and the empirical distributions of the two data sets are consistent estimators of the population reference measures for (Y_1, X) and (Y_2, X).

Suppose Assumptions 1 and 3 hold. Then Theorem 3 implies that for all $\delta > 0$,

$$\mathcal{I}(\delta) = \inf_{\lambda \in \mathbb{R}_+^2} \left[\langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{V}} f_{\theta,\lambda} \, d\varpi \right],$$

where

$$f_{\theta,\lambda}(v) = \sup_{y_1', y_2', x'} [f(y_1', y_2', y; \theta) - \lambda_1 c_1((y_1, x_1), (y_1', x')) - \lambda_2 c_2((y_2, x_2, y_2', x'))]$$

with $v = (y_1, x_1, y_2, x_2)$.

Let $\hat{\mu}_{13}$ and $\hat{\mu}_{23}$ denote the empirical measures based on the two data sets. The dual form of $\mathcal{I}(\delta)$ can be estimated by

$$\hat{\mathcal{I}}(\delta) := \inf_{\lambda \in \mathbb{R}_+^2} \left[\langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\hat{\mu}_{13}, \hat{\mu}_{23})} \int_{\mathcal{V}} f_{\theta,\lambda} \, d\varpi \right].$$

A direct consequence of Kellerer [31, proposition 2.1] is that

$$\hat{\mathcal{I}}(\delta) = \inf_{\lambda \in \mathbb{R}_{++}^2, \{\varphi_i\}_{i=1}^{n_1}, \{\varphi_j\}_{j=1}^{n_2}} \left[\langle \lambda, \delta \rangle + \frac{1}{n_1} \sum_{i=1}^{n_1} \varphi_i + \frac{1}{n_2} \sum_{j=1}^{n_2} \varphi_j \right]$$
$$\text{such that } f_{\theta,\lambda}(s_{1i}, s_{2j}) \le \varphi_i + \varphi_j' \text{ for any } i \in [n_1] \text{ and } j \in [n_2],$$

where the last expression reduces to the dual in Awasthi et al. [2] for the cost functions

$$c_1((y_1, x), (y_1', x')) = \|x - x'\|_p + \kappa_1 |y_1 - y_1'| \quad \text{and}$$
$$c_2((y_2, x), (y_2, x')) = \|x - x'\|_p + \kappa_2 \|y_2 - y_2'\|_{p'}.$$

7. W-DMR with Multimarginals

Sections 2–6 present a detailed study of W-DMR with two marginals. In this section, we briefly introduce W-DMR with more than two marginals or multimarginals and discuss strong duality for nonoverlapping and overlapping marginals.[8] Applications include extension of risk aggregation in Section 2.3.4 to any finite number of individual risks and robust treatment choice in Section 2.3.4 to multivalued treatment.

7.1. Nonoverlapping Marginals

Let $\mathcal{V} := \prod_{\ell \in [L]} \mathcal{S}_\ell$ for Polish spaces $\mathcal{S}_\ell$ for $\ell \in [L]$, and μ_ℓ be a probability measure on $(\mathcal{S}_\ell, \mathcal{B}_{\mathcal{S}_\ell})$. Let $\Pi(\mu_1, \dots, \mu_L)$ be the set of all possible couplings of $\mu_1, \dots, \mu_L$. Further, let $g : \mathcal{V} \to \mathbb{R}$ be a measurable function satisfying the following assumption.

Assumption 14. *The function $g : \mathcal{V} \to \mathbb{R}$ is a measurable function such that $\int_{\mathcal{V}} g d\gamma_0 > -\infty$ for some $\gamma_0 \in \Pi(\mu_1, \dots, \mu_L) \subset \mathcal{P}(\mathcal{V})$.*

For any $\gamma \in \mathcal{P}(\mathcal{V})$, let γ_ℓ denote the projection of γ on $\mathcal{S}_\ell$ for $\ell \in [L]$. The W-DMR with nonoverlapping multimarginals is formulated as

$$\mathcal{I}_{\mathrm{D}}(\delta) = \sup_{\gamma \in \Sigma_{\mathrm{D}}(\delta)} \int_{\mathcal{V}} g d\gamma,$$

where $\Sigma_{\mathrm{D}}(\delta)$ is the uncertainty set defined as

$$\Sigma_{\mathrm{D}}(\delta) = \{\gamma \in \mathcal{P}(\mathcal{V}) : \boldsymbol{K}_\ell(\mu_\ell, \gamma_\ell) \le \delta_\ell, \ \forall \ell \in [L]\},$$

in which $\delta = (\delta_1, \dots, \delta_L) \in \mathbb{R}_+^L$ is the radius of the uncertainty set.

For a generic vector $v \in \mathbb{R}^L$ and $A \subset [L]$, we write $v_A = (v_{A,1}, \dots, v_{A,L}) \in \mathbb{R}^L$ as follows:

$$v_{A,\ell} = \begin{cases} v_\ell & \text{if } \ell \in A, \\ 0 & \text{if } \ell \notin A. \end{cases}$$

We also define $\tilde{c}_\ell : \mathcal{S}_\ell \times \mathcal{S}_\ell \to \mathbb{R}_+ \cup \{\infty\}$ as

$$\tilde{c}_\ell(s_\ell, s_\ell') = \begin{cases} c_\ell(s_\ell, s_\ell') & \text{if } \ell \in A, \\ \infty \mathbb{1}\{s_\ell \neq s_\ell'\} & \text{if } \ell \notin A. \end{cases}$$

For a function $g : \mathcal{V} \to \mathbb{R}$ and $\lambda := (\lambda_1, \dots, \lambda_L) \in \mathbb{R}_+^L$, we define the function $g_{\lambda,A} : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$ as

$$g_{\lambda,A}(v) = \sup_{v' \in \mathcal{V}} \left\{ g(v') - \sum_{\ell=1}^{L} \lambda_\ell \tilde{c}_\ell\{s_\ell, s_\ell'\} \right\}$$

with $v := (s_1, \dots, s_L)$ and $v' := (s_1', \dots, s_L')$.

Theorem 10 (Nonoverlapping Case). *Suppose that Assumptions 1 and 14 hold. Then, for any $\delta \in \mathbb{R}^L_{++}$ and $A \subset [L]$, we have*

$$\mathcal{I}_D(\delta_A) = \inf_{\lambda \in \mathbb{R}^L_+} \left[\langle \lambda, \delta_A \rangle + \sup_{\pi \in \Pi(\mu_1, \dots, \mu_L)} \int_{\mathcal{V}} g_{\lambda, A} \, d\pi \right].$$

By choosing the set A as a proper subset of $[L]$, Theorem 10 includes the boundary case when some entries of $\delta \in \mathbb{R}^L_+$ are zero. In practice, the dual in Theorem 10 involves the computation of the multimarginal problem, $\sup_{\pi \in \Pi(\mu_1, \dots, \mu_L)} \int_{\mathcal{V}} g_\lambda \, d\pi$; see Pass [40], Pass [41], Pass [42], von Lindheim [54], Nenna and Pass [39], and Mehta et al. [37] for detailed studies of properties and computation of multimarginal problems for specific functions g_λ. For general possibly non-Borel-measurable g_λ, the strong duality in Kellerer [31] could be applied. The established result is stated in Corollary S.3 in the Online Supplement.

7.2. Overlapping Marginals

Let $\mathcal{S} := (\prod_{\ell \in [L]} \mathcal{Y}_\ell) \times \mathcal{X}$, where $\mathcal{Y}_\ell$ for $\ell \in [L]$ and $\mathcal{X}$ are Polish spaces. Let $\mathcal{S}_\ell := \mathcal{Y}_\ell \times \mathcal{X}$ for $\ell \in [L]$. Let $\mu_{\ell, L+1} \in \mathcal{P}(\mathcal{S}_\ell)$ for $\ell \in [L]$ be such that the projections of $\mu_{\ell, L+1}$ on $\mathcal{X}$ are the same for $\ell \in [L]$. We call the Fréchet class of all probability measures on $\mathcal{S}$ having marginals $(\mu_{\ell, L+1})_{\ell \in [L]}$ the Fréchet class with overlapping marginals and denote it as $\mathcal{F}(\mathcal{S}; (\mu_{\ell, L+1})_{\ell \in [L]}) := \mathcal{F}((\mu_{\ell, L+1})_{\ell \in [L]})$. This class is the star-like system of marginals in Rüschendorf [48] and Embrechts and Puccetti [12]; see also Doan et al. [10].

Moreover, let $f : \mathcal{S} \to \mathbb{R}$ be a measurable function satisfying the following assumption.

Assumption 15. *The function $f : \mathcal{S} \to \mathbb{R}$ is a measurable function such that $\int_{\mathcal{S}} f \, d\nu_0 > -\infty$ for some $\nu_0 \in \Pi(\mu_{1, L+1}, \dots, \mu_{L, L+1}) \subset \mathcal{P}(\mathcal{S})$.*

For any $\gamma \in \mathcal{P}(\mathcal{S})$, let $\gamma_{\ell, L+1}$ denote the projection of γ on $\mathcal{Y}_\ell \times \mathcal{X}$ for $\ell \in [L]$. Similar to the two-marginals case, the W-DMR with overlapping multimarginals is defined as

$$\mathcal{I}(\delta) = \sup_{\gamma \in \Sigma(\delta)} \int_{\mathcal{S}} f \, d\gamma,$$

where $\Sigma(\delta)$ is the uncertainty set defined as

$$\Sigma(\delta) = \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{K}_\ell(\mu_{\ell, L+1}, \gamma_{\ell, L+1}) \le \delta_\ell \text{ for } \ell \in [L]\},$$

in which $\delta = (\delta_1, \dots, \delta_L) \in \mathbb{R}^L_+$ is the radius of the uncertainty set.

For a function $f : \mathcal{V} \to \mathbb{R}$, $\lambda := (\lambda_1, \dots, \lambda_L) \in \mathbb{R}^L_+$, and $A \subset [L]$, we define the function $f_{\lambda, A} : \mathcal{V} \to \overline{\mathbb{R}}$ as follows:

$$f_{\lambda, A}(v) = \sup_{s' \in \mathcal{S}} \left\{ f(s') - \sum_{\ell=1}^{L} \lambda_\ell \tilde{c}_\ell(s_\ell, s'_\ell) \right\},$$

where $v = (s_1, \dots, s_L)$, $s' = (y'_1, \dots, y'_L, x')$, $s'_\ell = (y'_\ell, x')$, and $s_\ell = (y_\ell, x_\ell)$, and

$$\tilde{c}_\ell(s_\ell, s'_\ell) = \begin{cases} c_\ell(s_\ell, s'_\ell) & \text{if } \ell \in A, \\ \infty \mathbb{1}\{s_\ell \neq s'_\ell\} & \text{if } \ell \notin A. \end{cases}$$

Theorem 11 (Overlapping Case). *Suppose that Assumptions 1 and 15 hold. Then, for any $\delta \in \mathbb{R}^L_{++}$ and $A \subset [L]$, we have*

$$\mathcal{I}(\delta_A) = \inf_{\lambda \in \mathbb{R}^L_+} \left[\langle \lambda, \delta_A \rangle + \sup_{\pi \in \Pi(\mu_{1, L+1}, \dots, \mu_{L, L+1})} \int_{\mathcal{V}} f_{\lambda, A} \, d\pi \right].$$

When A is a proper subset of $[L]$, Theorem 11 is a duality result for the boundary case. Similar to the nonoverlapping case, strong duality holds for the inner multimarginal problem under additional conditions. The result is stated in Corollary S.4 in the Online Supplement.

7.3. Treatment Choice for Multivalued Treatment

We apply strong duality to multivalued treatment in Kido [33]. Let $d : \mathcal{X} \to [L]$ be a policy function or treatment rule on $\mathcal{X}$ and $Y_\ell \in \mathbb{R}$ denote the potential outcome under the treatment ℓ for $\ell \in [L]$. Consider the policy function

defined as

$$Y(d) := \sum_{\ell=1}^{L} Y_\ell \times 1\{d(X) = \ell\}.$$

Kido [33] introduces the following robust welfare function:

$$\mathrm{RW}_{\mathrm{C}}(d) = \sup_{\gamma \in \Sigma_{\mathrm{M}}(\delta_0)} \mathbb{E}_\gamma \left[\sum_{\ell=1}^{L} Y_\ell \mathbb{1}\{d(X) = \ell\} \right],$$

where the uncertainty set $\Sigma_{\mathrm{M}}(\delta_0)$ is based on the conditional distribution of $(Y_\ell)_{\ell \in [L]}$ given X:

$$\Sigma_{\mathrm{M}}(\delta_0) := \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{K}(\mu_{(Y_1,\ldots,Y_L)|X=x}, \gamma_{(Y_1,\ldots,Y_L)|X=x}) \le \delta_0 \text{ for all } x,\ \mu_X = \gamma_X\},$$

in which the cost function c associated with $\boldsymbol{K}$ is

$$c((y_1,\ldots,y_L),(y'_1,\ldots,y'_L)) = \sum_{\ell=1}^{L} |y_\ell - y'_\ell|.$$

Note that the uncertainty set $\Sigma_{\mathrm{M}}(\delta_0)$ does not allow any potential shift[9] in X. When $Y_1,\ldots,Y_L$ are unbounded, Kido [33] shows that

$$\begin{aligned}\mathrm{RW}_{\mathrm{C}}(d) &= \sum_{\ell=1}^{L} \mathbb{E}_{(Y_\ell,X)\sim\mu_{\ell,L+1}}[(Y_\ell - \delta_0) I(D(X) = \ell)] \\ &= \mathbb{E}_X \left[\sum_{\ell=1}^{L} (\mathbb{E}[Y_\ell|X] - \delta_0) I(D(X) = \ell) \right].\end{aligned}$$

We apply W-DMR for overlapping marginals with the following cost function,

$$c_\ell(s_\ell, s'_\ell) = |y_\ell - y'_\ell| + \|x_\ell - x'_\ell\|_2,$$

and define a robust welfare function as

$$\mathrm{RW}(d) = \sup_{\gamma \in \Sigma(\delta)} \mathbb{E}_\gamma \left[\sum_{\ell=1}^{L} Y_\ell I(d(X) = \ell) \right].$$

Proposition 7. *For $\ell \in [L]$, let $c_\ell(s_\ell, s'_\ell) = |y_\ell - y'_\ell| + \|x_\ell - x'_\ell\|_2$. Assume that Y_ℓ is unbounded, $\mathbb{E}[\|X\|_2^2] < \infty$, and $\mathbb{E}[|Y_\ell|] < \infty$. Then*

$$\mathrm{RW}(d) = \sup_{\lambda \ge 1} \left\{ \inf_{\pi \in \Pi(\mu_{1,L+1},\cdots,\mu_{L,L+1})} \int_{\mathcal{V}} \min_{\ell \in [L]} \{y_\ell + \phi_{\lambda,\ell}(x_1,\ldots,x_L)\} d\pi(s) - \langle \lambda, \delta \rangle \right\},$$

where $\varphi_{\lambda,\ell}(x_1,\ldots,x_L) = \min_{x', d(x')=\ell} \sum_{\ell=1}^{L} \lambda_\ell \|x_\ell - x'\|_2$.

Proposition 7 is an extension of Proposition 6.

8. Concluding Remarks

In this paper, we have introduced W-DMR in marginal problems for both nonoverlapping and overlapping marginals and established fundamental results including strong duality, finiteness of the proposed W-DMR, and existence of an optimizer at each radius. We have also shown continuity of the W-DMR-MP as a function of the radius. Applicability of the proposed W-DMR in marginal problems and established properties is demonstrated via distinct applications when the sample information comes from multiple data sources and only some marginal reference measures are identified. To the best of the authors' knowledge, this paper is the first systematic study of W-DMR in marginal problems. Many open questions remain including the structure of optimizers of W-DMR for both nonoverlapping and overlapping marginals, efficient numerical algorithms, and estimation and inference in each motivating example. Another useful extension is to consider objective functions that are nonlinear in the joint probability measure such as the Value-at-Risk of a linear portfolio of risks in Puccetti and Rüschendorf [44] and robust spectral measures of risk in Ghossoub et al. [26] and Ennaji et al. [16].

Acknowledgments
The authors acknowledge valuable feedback from seminar participants at UIUC and participants of the Optimization-Conscious Econometrics Conference II at the University of Chicago, the KI+Scale MoDL Retreat at the University of Washington, and the Econometrics and Optimal Transport Workshop at the University of Washington.

Appendix. Proofs of Main Results
The technical lemmas can be found in the Online Supplement to this paper.

A.1. Proofs in Section 3
A.1.1. Proof of Theorem 2. The expressions of $\mathcal{I}_D(\delta_1,0)$ and $\mathcal{I}_D(0,\delta_2)$ can be derived from $\mathcal{I}_D(\delta_1,\delta_2)$ for $\delta_1,\delta_2>0$ with appropriate modifications of the cost function. In particular, consider another cost function $\hat{c}_2(s_2,s_2')=\infty\mathbb{1}\{s_2\neq s_2'\}$ and the optimal transport distance $\hat{\boldsymbol{K}}_2$ associated with $\hat{c}_2$. Define an uncertainty set $\hat{\Sigma}_D(\delta_1,\delta_2)$ depending on $\boldsymbol{K}_1$ and $\hat{\boldsymbol{K}}_2$ as

$$\hat{\Sigma}_D(\delta_1,\delta_2)=\{\gamma\in\mathcal{P}(\mathcal{S}_1\times\mathcal{S}_2):\boldsymbol{K}_1(\gamma_1,\mu_1)\le\delta_1,\hat{\boldsymbol{K}}_2(\gamma_2,\mu_2)\le\delta_2\}.$$

Moreover, we define $\hat{\mathcal{I}}_D:\mathbb{R}_+^2\to\mathbb{R}$ as

$$\hat{\mathcal{I}}_D(\delta_1,\delta_2)=\sup_{\gamma\in\hat{\Sigma}_D(\delta_1,\delta_2)}\int_{\mathcal{V}}g(s_1,s_2)\,d\gamma(s_1,s_2).$$

We note $\hat{\boldsymbol{K}}_2(\mu,\nu)=0$ if and only if $\mu=\nu$. For all $\delta_2>0$, $\hat{\Sigma}_D(\delta_1,\delta_2)=\Sigma_D(\delta_1,0)$ and $\hat{\mathcal{I}}_D(\delta_1,\delta_2)=\mathcal{I}_D(\delta_1,0)$. Using the dual reformulation of $\hat{\mathcal{I}}_D$ on $\mathbb{R}_{++}^2$, we have

$$\mathcal{I}_D(\delta_1,0)=\hat{\mathcal{I}}_D(\delta_1,\delta_2)=\inf_{\lambda\in\mathbb{R}_+^2}\left[\langle\lambda,\delta\rangle+\sup_{\varpi\in\Pi(\mu_1,\mu_2)}\int_{\mathcal{V}}g_\lambda(s_1,s_2)d\varpi(s_1,s_2)\right],$$

where

$$\begin{aligned}g_\lambda(s_1,s_2)&=\sup_{s_1'\in\mathcal{S}_1,s_2'\in\mathcal{S}_2}\{g(s_1',s_2')-\lambda_1c_1(s_1,s_1')-\lambda_2\hat{c}_2(s_2,s_2')\}\\&=\sup_{s_1'\in\mathcal{S}_1}\{g(s_1',s_2)-\lambda_1c_1(s_1,s_1')\}=g_{\lambda,1}(s_1,s_2).\end{aligned}$$

Because $g_{\lambda,1}(s_1,s_2)$ is independent of λ_2, letting $\lambda_2=0$ yields

$$\mathcal{I}_D(\delta_1,0)=\inf_{\lambda_1\in\mathbb{R}_+}\left[\lambda_1\delta_1+\sup_{\varpi\in\Pi(\mu_1,\mu_2)}\int_{\mathcal{V}}g_{\lambda,1}(v)\,d\varpi(v)\right].$$

Using the same reasoning, we can get the expression of $\mathcal{I}_D(0,\delta_2)$.

In the rest of the proof, we show the dual reformulation of $\mathcal{I}_D$ on $\mathbb{R}_{++}^2$. Let $\mathcal{P}_D$ denote the set of $\gamma\in\mathcal{P}(\mathcal{V})$ that satisfies $\boldsymbol{K}_1(\mu_1,\gamma_1)<\infty$, $\boldsymbol{K}_2(\mu_2,\gamma_2)<\infty$, and $\int_{\mathcal{V}}g\,d\gamma>-\infty$. Taking the Legendre transform on $\mathcal{I}$ yields that any $\lambda\in\mathbb{R}_{++}^2$,

$$\begin{aligned}\mathcal{I}_D^\star(\lambda)&:=\sup_{\delta\in\mathbb{R}_+^2}\{\mathcal{I}_D(\delta)-\langle\lambda,\delta\rangle\}=\sup_{\delta\in\mathbb{R}_+^2}\sup_{\gamma\in\Sigma(\delta)}\left\{\int_{\mathcal{V}}g\,d\gamma-\langle\lambda,\delta\rangle\right\}\\&=\sup_{\delta\in\mathbb{R}_+^2}\sup_{\gamma\in\mathcal{P}(\mathcal{V})}\left\{\int_{\mathcal{V}}g\,d\gamma-\langle\lambda,\delta\rangle:\boldsymbol{K}_\ell(\mu_\ell,\gamma_\ell)\le\delta_\ell,\ \forall\ell\in[2]\right\}\\&=\sup_{\gamma\in\mathcal{P}(\mathcal{V})}\sup_{\delta\in\mathbb{R}_+^2}\left\{\int_{\mathcal{V}}g\,d\gamma-\langle\lambda,\delta\rangle:\boldsymbol{K}_\ell(\mu_\ell,\gamma_\ell)\le\delta_\ell,\ \forall\ell\in[2]\right\}\\&=\sup_{\gamma\in\mathcal{P}_D}\underbrace{\left\{\int_{\mathcal{V}}g\,d\gamma-\lambda_1\boldsymbol{K}_1(\mu_1,\gamma_1)-\lambda_2\boldsymbol{K}_2(\mu_2,\gamma_2)\right\}}_{:=I_{D,\lambda}[\gamma]}=\sup_{\gamma\in\mathcal{P}_D}I_{D,\lambda}[\gamma].\end{aligned}$$

We note that the expression above also holds for $\lambda\in\mathbb{R}_+^2\setminus\mathbb{R}_{++}^2$. Recall that

$$\varphi_\lambda(v,v')=g(s_1',s_2')-\lambda_1c_1(s_1,s_1')-\lambda_2c_2(s_2,s_2').$$

Let $\mathcal{G}_{D,\lambda}$ be the set of all probability measures π on $\mathcal{V}\times\mathcal{V}$ such that $\int_{\mathcal{V}\times\mathcal{V}}\varphi_\lambda d\pi$ is well-defined and the first and second marginals are μ_1 and μ_2.[10] Lemma A.3 implies $\mathcal{I}_D^\star(\lambda)=\sup_{\pi\in\mathcal{G}_{D,\lambda}}\int_{\mathcal{V}\times\mathcal{V}}\varphi_\lambda d\pi$. By Lemma A.4, we have for all $\lambda\in\mathbb{R}_+^2$

$$\mathcal{I}_D^\star(\lambda)=\sup_{\pi\in\mathcal{G}_{D,\lambda}}\int_{\mathcal{V}\times\mathcal{V}}\varphi_\lambda d\pi=\sup_{\pi\in\overline{\Gamma}}\int_{\mathcal{V}\times\mathcal{V}}\varphi_\lambda d\pi,$$

where we write $\overline{\Gamma}=\Gamma(\Pi(\mu_1,\mu_2),\varphi_\lambda)$ for simplicity. From Lemma S.3(i) in the Online Supplement, $\mathcal{I}_D$ is bounded from

below, nondecreasing, and concave. As a result, $\mathcal{I}_{\mathrm{D}} < \infty$ or $\mathcal{I}_{\mathrm{D}} = \infty$ on $\delta \in \mathbb{R}^2_{++}$. In the first case, by Lemma S.4 in the Online Supplement, for all $\delta \in \mathbb{R}^2_{+}$,

$$\mathcal{I}_{\mathrm{D}}(\delta) = \inf_{\lambda \in \mathbb{R}^2_+} \left\{ \langle \lambda, \delta \rangle + \sup_{\pi \in \overline{\Gamma}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi \right\}.$$

In the second case, by definition $\mathcal{I}^\star_{\mathrm{D}}(\lambda) = \infty$ for all $\lambda \in \mathbb{R}^2_+$ and the above is also true. Moreover, example 2 of Zhang et al. [57] implies that φ_λ satisfies the interchangeability principle with respect to $\Pi(\mu_1, \mu_2)$. So, Lemma S.1 in the Online Supplement implies that for all $\lambda \in \mathbb{R}^2_{++}$

$$\sup_{\pi \in \overline{\Gamma}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi = \sup_{\gamma \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g_\lambda(v) \, d\gamma(v),$$

where $g_\lambda(v) = \sup_{v' \in \mathcal{V}} \varphi_\lambda(v, v')$. This shows for all $\delta \in \mathbb{R}^2_{++}$

$$\mathcal{I}_{\mathrm{D}}(\delta) = \inf_{\lambda \in \mathbb{R}^2_+} \left\{ \langle \lambda, \delta \rangle + \sup_{\gamma \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{V}} g_\lambda \, d\gamma \right\}. \quad \square$$

Lemma A.1. *If $\lambda_1 > 0$ and $\lambda_2 > 0$, then*

$$\sup_{\gamma \in \mathcal{P}_{\mathrm{D}}} I_{\mathrm{D},\lambda}[\gamma] = \sup_{\gamma \in \mathcal{P}_{\mathrm{D}}} \sup_{\pi \in \Pi(\mu_1, \mu_2, \gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi.$$

Proof of Lemma A.1. Fix any $\epsilon > 0$ and $\gamma \in \mathcal{P}_{\mathrm{D}}$. By the definition of $\mathcal{P}_{\mathrm{D}}$, we have $\boldsymbol{K}_\ell(\mu_\ell, \gamma_\ell) < \infty$ and hence there is $\nu_\ell \in \Pi(\mu_\ell, \gamma_\ell)$ such that $\boldsymbol{K}_\ell(\mu_\ell, \gamma_\ell) \geq \int_{\mathcal{S}_\ell \times \mathcal{S}_\ell} c_\ell \, d\nu_\ell - \epsilon/(\lambda_1 + \lambda_2)$. Let $K = \{K_1, K_2, K_3\}$ with $K_1 = \{1,3\}$, $K_2 = \{2,4\}$ and $K_3 = \{3,4\}$. Because K is decomposable, then by Proposition S.1 in the Online Supplement, there is a measure $\tilde{\pi}$ on $\mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{S}_1 \times \mathcal{S}_2$ with marginals given by $\pi_{1,3} = \nu_1$, $\pi_{2,4} = \nu_2$, and $\pi_{3,4} = \gamma$. Moreover, we note $\int_{\mathcal{V}\times\mathcal{V}} c_\ell(s_\ell, s'_\ell) \, d\tilde{\pi} = \int_{\mathcal{S}_\ell \times \mathcal{S}_\ell} c_\ell \, d\nu_\ell \leq \boldsymbol{K}_\ell(\mu_\ell, \gamma_\ell) + \epsilon/(\lambda_1 + \lambda_2) < \infty$. Now, we show the left-hand side (LHS) is not greater than the right-hand side (RHS). When $I_{\mathrm{D},\lambda}[\gamma] = \infty$, provided $\boldsymbol{K}_\ell(\mu_\ell, \gamma_\ell) \in (0, \infty)$ for $\ell = 1, 2$, we must have $\int_{\mathcal{V}} g \, d\gamma = \infty$. Then, it is apparent that $\int \varphi_\lambda \, d\tilde{\pi} = \infty$ and hence $I_{\mathrm{D},\lambda}[\gamma] \leq \int \varphi_\lambda d\tilde{\pi} + \epsilon$. When $I_{\mathrm{D},\lambda}[\gamma] < \infty$, then $\int_{\mathcal{V}} g \, d\gamma < \infty$. Therefore, the integral given by

$$\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\tilde{\pi} = \int_{\mathcal{V}} g \, d\gamma - \int_{\mathcal{S}_1 \times \mathcal{S}_1} \lambda_1 c_1 \, d\nu_1 - \int_{\mathcal{S}_2 \times \mathcal{S}_2} \lambda_2 c_2 \, d\nu_2 < \infty$$

is well-defined. The desired result follows from the estimate below:

$$\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\tilde{\pi} \geq \int_{\mathcal{V}} g \, d\gamma - \lambda_1 \boldsymbol{K}_1(\mu_1, \gamma_1) - \lambda_2 \boldsymbol{K}_2(\mu_2, \gamma_2) - \epsilon = I_{\mathrm{D},\lambda}[\gamma] - \epsilon.$$

Therefore, we have $I_{\mathrm{D},\lambda}[\gamma] \leq \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\tilde{\pi} + \epsilon$. Because $\epsilon > 0$ and $\gamma \in \mathcal{P}_{\mathrm{D}}$ are arbitrary, we have

$$\sup_{\gamma \in \mathcal{P}_{\mathrm{D}}} I_{\mathrm{D},\lambda}[\gamma] \leq \sup_{\gamma \in \mathcal{P}_{\mathrm{D}}} \sup_{\pi \in \Pi(\mu_1, \mu_2, \gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi.$$

Next, we prove that the reversed direction holds by showing that if $\gamma \in \mathcal{P}_{\mathrm{D}}$, then $I_{\mathrm{D},\lambda}[\gamma] \geq \sup_{\pi \in \Pi(\mu_1, \mu_2, \gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi$. Fix $\gamma \in \mathcal{P}_{\mathrm{D}}$. When $\int_{\mathcal{V}} g \, d\gamma = \infty$, $I_{\mathrm{D},\lambda}[\gamma] = \infty$ and then the proof is done. Next, when $\int_{\mathcal{V}} g \, d\gamma < \infty$, for any $\pi \in \Pi(\mu_1, \mu_2, \gamma)$ such that $\int \varphi_\lambda \, d\pi$ is well-defined,

$$\begin{aligned} I_{\mathrm{D},\lambda}[\gamma] &= \int_{\mathcal{V}} g \, d\gamma - \lambda_1 \boldsymbol{K}_1(\mu_1, \gamma_1) - \lambda_2 \boldsymbol{K}_2(\mu_2, \gamma_2) \\ &\geq \int_{\mathcal{V}} g(s'_1, s'_2) \, d\pi_{3,4} - \lambda_1 \int_{\mathcal{S}_1 \times \mathcal{S}_1} c_1(s_1, s'_1) \, d\pi_{1,3} - \lambda_2 \int_{\mathcal{S}_2 \times \mathcal{S}_2} c_2(s_2, s'_2) \, d\pi_{2,4} \\ &= \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi. \end{aligned}$$

With the convention that $\sup = -\infty$, if the integral $\int \varphi_\lambda \, d\pi$ is not well-defined for all $\pi \in \Pi(\mu_1, \mu_2, \gamma)$, then $I_{\mathrm{D},\lambda}[\gamma] \geq \sup_{\pi \in \Pi(\mu_1, \mu_2, \gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi$ holds trivially. Otherwise, taking the supremum over $\pi \in \Pi(\mu_1, \mu_2, \gamma)$ on the RHS of the inequality above yields $I_{\mathrm{D},\lambda}[\gamma] \geq \sup_{\pi \in \Pi(\mu_1, \mu_2, \gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi$. The desired result follows. $\square$

Lemma A.2. *If $\lambda_1 > 0$ and $\lambda_2 > 0$, then*

$$\sup_{\gamma \in \mathcal{P}_{\mathrm{D}}} \sup_{\pi \in \Pi(\mu_1, \mu_2, \gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi = \sup_{\pi \in \mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi.$$

Proof of Lemma A.2. We divide the proof into the following two steps. The first step is to show that the LHS is less than or equal to the RHS. Fix any $\gamma \in \mathcal{P}_{\mathrm{D}}$. If $\int_{\mathcal{V}} g\, d\gamma = \infty$, from the proof of Lemma A.1, we can see that $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\tilde{\pi} = \infty$ for some $\tilde{\pi} \in \Pi(\mu_1, \mu_2, \gamma)$ and the LHS is ∞. So, the integral $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\tilde{\pi}$ is well-defined and $\tilde{\pi} \in \mathcal{G}_{\mathrm{D},\lambda}$. We must have $\sup_{\pi\in\mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\pi = \infty$ and the statement of the lemma is true. Now suppose $\int_{\mathcal{V}} g\, d\gamma < \infty$ holds. For any $\pi \in \Pi(\mu_1, \mu_2, \gamma)$, because $\int_{\mathcal{V}\times\mathcal{V}} (\lambda_1 c_1 + \lambda_2 c_2)\, d\pi \geq 0$, the integral

$$\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\pi = \int_{\mathcal{V}} g\, d\gamma - \int_{\mathcal{V}\times\mathcal{V}} (\lambda_1 c_1 + \lambda_2 c_2)\, d\pi < \infty$$

is well-defined. This shows $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$, and we have $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\pi \leq \sup_{\pi\in\mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\pi$. Taking the supremum over $\pi \in \Pi(\mu_1, \mu_2, \gamma)$ yields

$$\sup_{\pi\in\Pi(\mu_1,\mu_2,\gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\pi \leq \sup_{\pi\in\mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\pi.$$

Thus, we showed that the inequality above holds for all $\gamma \in \mathcal{P}_{\mathrm{D}}$, and this ends the first step.

The second step is to show that the LHS is greater than or equal to the RHS. Fix any $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$. It suffices to show

$$\sup_{\gamma\in\mathcal{P}_{\mathrm{D}}} \sup_{\pi\in\Pi(\mu_1,\mu_2,\gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\pi \geq \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\pi. \tag{A.1}$$

When $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\pi > -\infty$, we have $\int (\lambda_1 c_1 + \lambda_2 c_2)\, d\pi < \infty$ and hence $\int_{\mathcal{V}} g\, d\pi_{3,4} > -\infty$. It follows that $\pi \in \Pi(\mu_1, \mu_2, \pi_{3,4})$ and

$$\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\pi \leq \sup_{\tilde{\pi}\in\Pi(\mu_1,\mu_2,\gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\tilde{\pi} \leq \sup_{\gamma\in\mathcal{P}_{\mathrm{D}}} \sup_{\tilde{\pi}\in\Pi(\mu_1,\mu_2,\gamma)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\tilde{\pi}.$$

When $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\pi = -\infty$, the inequality (A.1) holds trivially. □

Lemma A.3. *For all $\lambda \in \mathbb{R}^2_+$, one has*

$$\mathcal{I}^\star_{\mathrm{D}}(\lambda) = \sup_{\pi\in\mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\pi. \tag{A.2}$$

Proof of Lemma A.3. We divide the proof into the following four cases. When $\lambda_1, \lambda_2 > 0$, the Equality (A.2) follows from Lemmas A.1 and A.2. When $\lambda_1 = \lambda_2 = 0$, we show that Equality (A.2) holds. Let $A_\ell = \{(v, v') \in \mathcal{V}\times\mathcal{V} : c_\ell(s_\ell, s'_\ell) < \infty\})$, and for simplicity, we write $g : (v, v') \longmapsto g(v')$ and $c_\ell : (v, v') \longmapsto c_\ell(s_\ell, s'_\ell)$ for $\ell = 1, 2$. By convention, $0c_\ell = 0, \pi$-a.s. if and only if $c_\ell < \infty, \pi$-a.s., so it follows that

$$\begin{aligned}
\sup_{\pi\in\mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\pi &= \sup\left\{ \int_{\mathcal{V}\times\mathcal{V}} g(v')\, d\pi(v, v') : \pi \in \mathcal{G}_{\mathrm{D},\lambda}, \pi(A_1 \cap A_2) = 1, \right\} \\
&\geq \sup\left\{ \int_{\mathcal{V}\times\mathcal{V}} g\, d\pi : \pi \in \mathcal{G}_{\mathrm{D},\lambda}, \int c_\ell\, d\pi < \infty \text{ for } \ell = 1, 2 \right\} \\
&\geq \sup\left\{ \int_{\mathcal{V}} g\, d\gamma : \gamma \in \mathcal{P}_{\mathrm{D}} \right\},
\end{aligned}$$

where the last inequality holds because for all $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$ with $\int c_\ell d\pi < \infty$ for $\ell = 1, 2$, the marginal $\pi_{3,4} \in \mathcal{P}_{\mathrm{D}}$, that is, $\pi(\mathcal{V}\times\cdot) \in \mathcal{P}_{\mathrm{D}}$ and vise versa. On the other hand, for any $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$ with $\pi(A_1 \cap A_2) = 1$, define a measure π_n on $\mathcal{V}\times\mathcal{V}$ as

$$\pi_n(\cdot) = \frac{\pi(\cdot \cap (A_{1n} \cap A_{2n}))}{\pi(A_{1n} \cap A_{2n})},$$

where $A_{\ell n} = \{(v, v') \in \mathcal{V}\times\mathcal{V} : c_\ell(s_\ell, s'_\ell) < n\}$ for $\ell = 1, 2$. Because $c_\ell < n$, π_n-a.s. for $\ell = 1, 2$, then the second marginal of π_n is in $\mathcal{P}_{\mathrm{D}}$.[11] By the monotone convergence theorem,

$$\lim_{n\to\infty} \int_{\mathcal{V}\times\mathcal{V}} g^+ \mathbb{1}_{A_{1n}\cap A_{2n}} d\pi = \int_{\mathcal{V}\times\mathcal{V}} g^+\, d\pi, \text{ and } \lim_{n\to\infty} \int_{\mathcal{V}\times\mathcal{V}} g^- \mathbb{1}_{A_{1n}\cap A_{2n}}\, d\pi = \int_{\mathcal{V}\times\mathcal{V}} g^-\, d\pi.$$

Moreover, because $\pi(A_{1n} \cap A_{2n}) \to 1$,

$$\lim_{n\to\infty} \int_{\mathcal{V}\times\mathcal{V}} g^+ d\pi_n = \lim_{n\to\infty} \frac{\int_{\mathcal{V}\times\mathcal{V}} g^+ \mathbb{1}_{A_{1n}\cap A_{2n}} d\pi}{\pi(A_{1n} \cap A_{2n})} = \int_{\mathcal{V}\times\mathcal{V}} g^+ d\pi.$$

Similarly, $\lim_{n\to\infty} \int_{\mathcal{V}\times\mathcal{V}} g^- d\pi_n = \int_{\mathcal{V}\times\mathcal{V}} g^- d\pi$. Because $\int g d\pi$ is well-defined, we can exclude the case $\int g^+ d\pi = \int g^- d\pi = \infty$. Therefore,

$$\int_{\mathcal{V}\times\mathcal{V}} g\, d\pi = \lim_{n\to\infty} \int_{\mathcal{V}\times\mathcal{V}} g\, d\pi_n \leq \sup_{\gamma\in\mathcal{P}_{\mathrm{D}}} \int_{\mathcal{V}} g\, d\gamma.$$

This shows $\sup_{\pi\in\mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda\, d\pi = \sup_{\pi\in\mathcal{P}_{\mathrm{D}}} \int_{\mathcal{V}\times\mathcal{V}} g\, d\pi$, and hence Equality (A.2) holds for $\lambda_1 = \lambda_2 = 0$.

Next, we show that Equality (A.2) holds when $\lambda_1 > 0, \lambda_2 = 0$. By definition, the integral $\int \varphi_\lambda \, d\pi$ is well-defined for all $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$. If $\int \varphi_\lambda \, d\pi = \infty$ for some $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$, then $\sup_{\pi \in \mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi \geq \sup_{\gamma \in \mathcal{P}_{\mathrm{D}}} \int g \, d\gamma$. Without loss of generality, assume $\int \varphi_\lambda \, d\pi < \infty$ for all $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$. It follows that for some $\pi \in \mathcal{G}_{D,\lambda}$,

$$\lambda_1 \int c_1 d\pi \leq \int (g^- + \lambda_1 c_1 + \lambda_2 c_2) d\pi < \infty,$$

and $\int c_1 d\pi < \infty$ and $\pi(A_1) = 1$. By convention, $0 \times c_2 = 0, \pi$-a.s. if and only if $0 \times c_2 < \infty, \pi$-a.s. We find that

$$\begin{aligned}
\sup_{\pi \in \mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi &= \sup \left\{ \int_{\mathcal{V}\times\mathcal{V}} \left[g(v') - \lambda_1 c_1(s_1, s_1') \right] d\pi(v, v') : \pi \in \mathcal{G}_{\mathrm{D},\lambda}, \pi(A_2) = 1 \right\} \\
&= \sup \left\{ \int_{\mathcal{V}\times\mathcal{V}} \left[g(v') - \lambda_1 c_1(s_1, s_1') \right] d\pi(v, v') : \pi \in \mathcal{G}_{\mathrm{D},\lambda}, \pi(A_1 \cap A_2) = 1 \right\} \\
&\geq \sup_{\gamma \in \mathcal{P}_{\mathrm{D}}} I_{D,\lambda}[\gamma].
\end{aligned}$$

On the other hand, for any $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$ with $\pi(A_2) = 1$, define a measure π_n' on $\mathcal{V} \times \mathcal{V}$ as

$$\pi_n(\cdot) = \frac{\pi(\cdot \cap (A_{1n}))}{\pi(A_{1n})}.$$

Using a similar argument as shown above, we can show $\int_{\mathcal{V}\times\mathcal{V}} [g - \lambda_1 c_1] \, d\pi \leq \sup_{\gamma \in \mathcal{P}_{\mathrm{D}}} I_{D,\lambda}[\gamma]$, and hence Equality (A.2) holds when $\lambda_1 > 0$ and $\lambda_2 = 0$. In the same way, we can show that Equality (A.2) holds when $\lambda_1 = 0, \lambda_2 > 0$. □

Lemma A.4. *Let $\lambda \in \mathbb{R}_+^2$. If φ_λ is interchangeable with respect to $\Pi(\mu_1, \mu_2)$, then*

$$\sup_{\pi \in \mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\pi = \sup_{\pi \in \Gamma(\Pi(\mu_1,\mu_2), \varphi_\lambda)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\pi.$$

Proof of Lemma A.4. For any $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$, it is obvious that $\pi_{1,2} \in \Pi(\mu_1, \mu_2)$ and hence $\pi \in \Gamma(\Pi(\mu_1, \mu_2), \varphi_\lambda)$. This shows $\mathcal{G}_{\mathrm{D},\lambda} \subset \Gamma(\Pi(\mu_1, \mu_2), \varphi_\lambda)$ and the LHS is less than or equal to the RHS.

Next, we show the LHS is not less than the RHS. We adopt the convention that the supremum of an empty set is $-\infty$. If $\int \varphi_\lambda d\pi$ is not well-defined for all $\pi \in \Gamma(\Pi(\mu_1, \mu_2), \varphi_\lambda)$, then the proof is done trivially. Now let π be any measure in $\Gamma(\Pi(\mu_1, \mu_2), \varphi_\lambda)$ for which integral $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\pi$ is well-defined. To finish the proof, it suffices to show

$$\sup_{\pi \in \mathcal{G}_{\mathrm{D},\lambda}} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda(v, v') d\pi(v, v') \geq \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda(v, v') d\pi(v, v'). \tag{A.3}$$

When $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda, d\pi = -\infty$, Inequality (A.3) holds trivially. Now suppose $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi = \infty$. Because $c_1, c_2 \geq 0$, we have $\int_{\mathcal{V}\times\mathcal{V}} g(v')$ $d\pi(v, v') = \infty$ and is well-defined. We note $\varphi_\lambda = g^+ - g^- - (\lambda_1 c_1 + \lambda_2 c_2)$ and hence $\varphi_\lambda^+ = g^+$ and $\varphi_\lambda^- = g^- + (\lambda_1 c_1 + \lambda_2 c_2)$. Because $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi$ is well-defined, then $\int_{\mathcal{V}\times\mathcal{V}} (\lambda_1 c_1 + \lambda_2 c_2) \, d\pi \leq \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda^- \, d\pi < \infty$. This shows that $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$ and Inequality (A.3) holds. Next, suppose $\int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi < \infty$. Given that the integral is well-defined, using the same reasoning as demonstrated above, we have $\int_{\mathcal{V}\times\mathcal{V}} g(v') \, d\pi(v, v') < \infty$ and $\int_{\mathcal{V}\times\mathcal{V}} (\lambda_1 c_1 + \lambda_2 c_2) \, d\pi < \infty$. So, $\pi \in \mathcal{G}_{\mathrm{D},\lambda}$ and the proof is done. □

A.1.2. Proof of Corollary 1. We provide only the derivation of the upper bound $\mathcal{I}_{\mathrm{D}}(\delta) = \sup_{\gamma \in \Sigma_{\mathrm{D}}(\delta)} \int \mathbb{1}(s_1 + s_2 \leq z) \, d\gamma(s_1, s_2)$. We can derive the expression of the lower bound $\inf_{\gamma \in \Sigma_{\mathrm{D}}} \int \mathbb{1}(s_1 + s_2 \leq z) \, d\gamma(s_1, s_2)$ by similar reasoning and the following identity:

$$\inf_{\gamma \in \Sigma_{\mathrm{D}}(\delta)} \int \mathbb{1}(s_1 + s_2 \leq z) \, d\gamma(s_1, s_2) = 1 - \sup_{\gamma \in \Sigma_{\mathrm{D}}(\delta)} \int \mathbb{1}(\{s_1 + s_2 > z\}) \, d\gamma(s_1, s_2).$$

When $\lambda_1 = 0$ or $\lambda_2 = 0$, $g_\lambda(s_1, s_2) = 0$ for all $(s_1, s_2) \in \mathcal{S}_1 \times \mathcal{S}_2$. When $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$, we have

$$\begin{aligned}
g_\lambda(s_1, s_2) &= \sup_{s_1', s_2'} [\mathbb{1}(s_1' + s_2' \leq z) - \lambda_1 |s_1 - s_1'|^2 - \lambda_2 |s_2 - s_2'|^2] \\
&= \left(1 - \inf_{s_1' + s_2' \leq z} [\lambda_1 |s_1 - s_1'|^2 + \lambda_2 |s_2 - s_2'|^2] \right)^+ \\
&= \begin{cases} 1 & \text{if } s_1 + s_2 \leq z \\ \left[1 - \frac{\lambda_1 \lambda_2 (s_1 + s_2 - z)^2}{\lambda_1 + \lambda_2} \right]^+ & \text{if } \{s_1 + s_2 > z\}. \end{cases}
\end{aligned}$$

By some simple algebra, we have

$$g_{\lambda,1}(s_1,s_2) = \sup_{s_1'}[\mathbb{1}(s_1' + s_2 \le z) - \lambda_1|s_1 - s_1'|^2]$$
$$= \begin{cases} 1 & \text{if } s_1 + s_2 \le z, \\ (1 - \lambda_1|s_1 + s_2 - z|^2)^+ & \text{if } \{s_1 + s_2 > z\}, \end{cases}$$

and

$$g_{\lambda,2}(s_1,s_2) = \sup_{s_2'}[\mathbb{1}(s_1 + s_2' \le z) - \lambda_2|s_2 - s_2'|^2]$$
$$= \begin{cases} 1 & \text{if } s_1 + s_2 \le z, \\ (1 - \lambda_2|s_1 + s_2 - z|^2)^+ & \text{if } \{s_1 + s_2 > z\}. \end{cases}$$

By applying Theorem 2, we have that for each $\delta = (\delta_1,\delta_2) \in \mathbb{R}^2_{++}$,

$$\mathcal{I}_D(\delta) = \inf_{\lambda\in\mathbb{R}^2_+}\left[\langle\lambda,\delta\rangle + \sup_{\pi\in\Pi(\mu_1,\mu_2)}\int g_\lambda(s_1,s_2)\,d\pi(s_1,s_2)\right].$$

However, in the rest of the proof, we show, for all $\delta = (\delta_1,\delta_2) \in \mathbb{R}^2_+$,

$$\mathcal{I}_D(\delta) = \inf_{\lambda\in\mathbb{R}^2_+}\sup_{\pi\in\Pi(\mu_1,\mu_2)}\left[\langle\lambda,\delta\rangle + \int_{\mathcal{V}} g_\lambda\,d\pi\right] = \sup_{\pi\in\Pi(\mu_1,\mu_2)}\inf_{\lambda\in\mathbb{R}^2_+}\left[\langle\lambda,\delta\rangle + \int_{\mathcal{V}} g_\lambda\,d\pi\right].$$

Define a function $F : \Pi(\mu_1,\mu_2)\times\mathbb{R}^2_+ \to \mathbb{R}$ as

$$F : (\pi,\lambda) \longmapsto -\langle\lambda,\delta\rangle - \int_{\mathcal{S}_1\times\mathcal{S}_2} g_\lambda\,d\pi.$$

We note that for any (s_1,s_2), the function $\lambda \longmapsto g_\lambda(s_1,s_2)$ is convex because it is the supremum of a set of affine functions in λ. As a result, $\lambda \longmapsto -\int g_\lambda d\pi$ is concave for each fixed π. For any $\lambda \in \mathbb{R}^2_+$, the function $\pi \longmapsto F(\pi,\lambda)$ is continuous because of continuous and bounded g_λ and the Portmanteau theorem. Moreover, it is easy to verify that $\pi \longmapsto F(\pi,\lambda)$ is convex. By Fan's [17, theorem 2] minimax theorem, we have

$$\inf_{\pi\in\Pi(\mu_1,\mu_2)}\sup_{\lambda\in\mathbb{R}^2_+} F(\pi,\lambda) = \sup_{\lambda\in\mathbb{R}^2_+}\inf_{\pi\in\Pi(\mu_1,\mu_2)} F(\pi,\lambda).$$

As a result, we have for all $\delta = (\delta_1,\delta_2) \in \mathbb{R}^2_{++}$

$$\mathcal{I}_D(\delta) = \inf_{\lambda\in\mathbb{R}^2_+}\sup_{\pi\in\Pi(\mu_1,\mu_2)} -F(\pi,\lambda) = -\sup_{\lambda\in\mathbb{R}^2_+}\inf_{\pi\in\Pi(\mu_1,\mu_2)} F(\pi,\lambda)$$
$$= -\inf_{\pi\in\Pi(\mu_1,\mu_2)}\sup_{\lambda\in\mathbb{R}^2_+} F(\pi,\lambda) = \sup_{\pi\in\Pi(\mu_1,\mu_2)}\inf_{\lambda\in\mathbb{R}^2_+} -F(\pi,\lambda)$$
$$= \sup_{\pi\in\Pi(\mu_1,\mu_2)}\inf_{\lambda\in\mathbb{R}^2_+}\left[\langle\lambda,\delta\rangle + \int_{\mathcal{V}} g_\lambda\,d\pi\right].$$

Using the same reasoning as above, the application of Fan [17, theorem 2] to $\mathcal{I}_D(\delta_1,0)$ yields

$$\mathcal{I}_D(\delta_1,0) = \sup_{\pi\in\Pi(\mu_1,\mu_2)}\inf_{\lambda_1\in\mathbb{R}_+}\left[\lambda_1\delta_1 + \int_{\mathcal{V}} g_{\lambda,1}\,d\pi\right].$$

Because $g_\lambda \downarrow g_{\lambda,1}$ as $\lambda_2 \uparrow \infty$, the monotone convergence theorem implies

$$\inf_{\lambda\in\mathbb{R}^2_+}\left[\langle\lambda,(\delta_1,0)\rangle + \int_{\mathcal{V}} g_\lambda d\pi\right] = \inf_{\lambda_1\in\mathbb{R}_+}\left[\lambda_1\delta_1 + \inf_{\lambda_2\in\mathbb{R}_+}\int_{\mathcal{V}} g_\lambda\,d\pi\right]$$
$$= \inf_{\lambda_1\in\mathbb{R}_+}\left[\lambda_1\delta_1 + \lim_{\lambda_2\to\infty}\int_{\mathcal{V}} g_\lambda\,d\pi\right]$$
$$= \inf_{\lambda_1\in\mathbb{R}_+}\left[\lambda_1\delta_1 + \int g_{\lambda,1}\,d\pi\right].$$

Taking the supremum over $\pi \in \Pi(\mu_1, \mu_2)$ on both sides yields that for $\delta_1 > 0$

$$\mathcal{I}_{\mathrm{D}}(\delta_1, 0) = \sup_{\pi \in \Pi(\mu_1,\mu_2)} \inf_{\lambda_1 \in \mathbb{R}_+} \left[\lambda_1 \delta_1 + \int g_{\lambda,1}\, d\pi \right] = \sup_{\pi \in \Pi(\mu_1,\mu_2)} \inf_{\lambda \in \mathbb{R}^2_+} \left[\langle \lambda, (\delta_1, 0) \rangle + \int_{\mathcal{V}} g_\lambda\, d\pi \right].$$

Similarly, we can show that for $\delta_2 > 0$

$$\mathcal{I}_{\mathrm{D}}(0, \delta_2) = \sup_{\pi \in \Pi(\mu_1,\mu_2)} \inf_{\lambda_2 \in \mathbb{R}_+} \left[\lambda_2 \delta_2 + \int_{\mathcal{V}} g_{\lambda,2}\, d\pi \right] = \sup_{\pi \in \Pi(\mu_1,\mu_2)} \inf_{\lambda \in \mathbb{R}^2_+} \left[\langle \lambda, (0, \delta_2) \rangle + \int_{\mathcal{V}} g_\lambda\, d\pi \right].$$

In addition, when $\delta_1 = \delta_2 = 0$, we note $g_\lambda \downarrow g$ as $\lambda_1, \lambda_2 \uparrow \infty$ and the monotone convergence theorem implies $\inf_{\lambda \in \mathbb{R}^2_+} \int g_\lambda d\pi = \int g d\pi$ and

$$\mathcal{I}_{\mathrm{D}}(0) = \sup_{\pi \in \Pi(\mu_1,\mu_2)} \inf_{\lambda \in \mathbb{R}^2_+} \int g_\lambda\, d\pi = \inf_{\lambda \in \mathbb{R}^2_+} \sup_{\pi \in \Pi(\mu_1,\mu_2)} \int g_\lambda\, d\pi = \sup_{\pi \in \Pi(\mu_1,\mu_2)} \int g\, d\pi.$$

This completes the proof that for all $\delta = (\delta_1, \delta_2) \in \mathbb{R}^2_+$

$$\mathcal{I}_{\mathrm{D}}(\delta) = \sup_{\pi \in \Pi(\mu_1,\mu_2)} \inf_{\lambda \in \mathbb{R}^2_+} \left[\langle \lambda, \delta \rangle + \int g_\lambda\, d\pi \right] = \inf_{\lambda \in \mathbb{R}^2_+} \sup_{\pi \in \Pi(\mu_1,\mu_2)} \left[\langle \lambda, \delta \rangle + \int g_\lambda\, d\pi \right]. \quad \square$$

A.1.3. Proof of Theorem 3. The expressions of $\mathcal{I}(\delta_1, 0)$ and $\mathcal{I}(0, \delta_2)$ can be derived from $\mathcal{I}(\delta_1, \delta_2)$ for $\delta_1, \delta_2 > 0$ with appropriate modifications of the cost function. In particular, consider another cost function $\hat{c}_2(s_2, s_2') = \infty \mathbb{1}\{s_2 \neq s_2'\}$ and the optimal transport distance $\hat{\boldsymbol{K}}_2$ associated with $\hat{c}_2$. Define an uncertainty set $\hat{\Sigma}(\delta_1, \delta_2)$ depending on $\boldsymbol{K}_1$ and $\hat{\boldsymbol{K}}_2$ as

$$\hat{\Sigma}(\delta_1, \delta_2) = \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{K}_1(\gamma_{13}, \mu_{13}) \le \delta_1, \hat{\boldsymbol{K}}_2(\gamma_{23}, \mu_{23}) \le \delta_2\}.$$

Moreover, we define $\hat{\mathcal{I}} : \mathbb{R}^2_+ \to \mathbb{R}$ as

$$\hat{\mathcal{I}}(\delta_1, \delta_2) = \sup_{\gamma \in \hat{\Sigma}(\delta_1,\delta_2)} \int_{\mathcal{V}} f(v)\, d\gamma(v).$$

We note $\hat{\boldsymbol{K}}_2(\mu, \nu) = 0$ if and only if $\mu = \nu$. So, for all $\delta_2 > 0$, $\hat{\Sigma}(\delta_1, \delta_2) = \Sigma(\delta_1, 0)$ and $\hat{\mathcal{I}}(\delta_1, \delta_2) = \mathcal{I}(\delta_1, 0)$. Using the dual reformulation of $\hat{\mathcal{I}}$ on $\mathbb{R}^2_{++}$, we have

$$\mathcal{I}(\delta_1, 0) = \hat{\mathcal{I}}(\delta_1, \delta_2) = \inf_{\lambda_1 \in \mathbb{R}_+} \left[\langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\mu_{13},\mu_{23})} \int_{\mathcal{V}} f_\lambda(s_1, s_2)\, d\varpi(s_1, s_2) \right],$$

where

$$\begin{aligned} f_\lambda(s_1, s_2) &= \sup_{(y_1', y_2', x') \in \mathcal{S}} \{ f(y_1', y_2', x') - \lambda_1 c_1(s_1, (y_1', x')) - \lambda_2 \hat{c}_2(s_2, (y_2', x')) \} \\ &= \sup_{s_1' \in \mathcal{S}_1} \{ f(y_1', y_2, x_2) - \lambda_1 c_1(s_1, (y_1', x_2)) \} = f_{\lambda,1}(s_1, s_2). \end{aligned}$$

Because $f_{\lambda,1}(s_1, s_2)$ is independent of λ_2, letting $\lambda_2 = 0$ yields

$$\mathcal{I}(\delta_1, 0) = \inf_{\lambda_1 \in \mathbb{R}_+} \left[\lambda_1 \delta_1 + \sup_{\varpi \in \Pi(\mu_{13},\mu_{23})} \int_{\mathcal{V}} f_{\lambda,1}(v)\, d\varpi(v) \right].$$

Using the same reasoning, we can get the expression of $\mathcal{I}(0, \delta_2)$.

In the rest of the proof, we show that the dual reformulation of $\mathcal{I}$ on $\mathbb{R}^2_{++}$ holds. Let $\overline{\mathcal{P}}$ denote the set of $\gamma \in \mathcal{P}(\mathcal{S})$ such that $\boldsymbol{K}_\ell(\mu_{\ell 3}, \gamma_{\ell 3}) < \infty$ for $\ell = 1, 2$ and $\int_{\mathcal{S}} f d\gamma > -\infty$. Taking the Legendre transform on $\mathcal{I}$ gives

$$\begin{aligned} \mathcal{I}^\star(\lambda) &:= \sup_{\delta \in \mathbb{R}^2_+} \{ \mathcal{I}(\delta) - \langle \lambda, \delta \rangle \} = \sup_{\delta \in \mathbb{R}^2_+} \sup_{\gamma \in \Sigma(\delta)} \left\{ \int_{\mathcal{S}} f\, d\gamma - \langle \lambda, \delta \rangle \right\} \\ &= \sup_{\delta \in \mathbb{R}^2_+} \sup_{\gamma \in \overline{\mathcal{P}}} \left\{ \int_{\mathcal{S}} f\, d\gamma - \langle \lambda, \delta \rangle : \boldsymbol{K}_\ell(\mu_{\ell 3}, \gamma_{\ell 3}) \le \delta_\ell,\ \forall \ell \in [2] \right\} \\ &= \sup_{\gamma \in \overline{\mathcal{P}}} \sup_{\delta \in \mathbb{R}^2_+} \left\{ \int_{\mathcal{S}} f\, d\gamma - \langle \lambda, \delta \rangle : \boldsymbol{K}_\ell(\mu_{\ell 3}, \gamma_{\ell 3}) \le \delta_\ell,\ \forall \ell \in [2] \right\} \\ &= \sup_{\gamma \in \overline{\mathcal{P}}} \underbrace{\left\{ \int_{\mathcal{S}} f\, d\gamma - \lambda_1 \boldsymbol{K}_1(\mu_{13}, \gamma_{13}) - \lambda_2 \boldsymbol{K}_2(\mu_{23}, \gamma_{23}) \right\}}_{:= I_\lambda[\gamma]} = \sup_{\gamma \in \overline{\mathcal{P}}} I_\lambda[\gamma]. \end{aligned}$$

We note that the expression above still holds when $\lambda \in \mathbb{R}^2_+ \setminus \mathbb{R}^2_{++}$. Recall the definition of the function $\phi_\lambda : \mathcal{V} \times \mathcal{S} \to \mathbb{R}$. Let $\mathcal{G}_\lambda$ denote the set of $\pi \in \mathcal{P}(\mathcal{V} \times \mathcal{S})$ such that $\int_{\mathcal{V}\times\mathcal{S}} \phi_\lambda \, d\pi$ is well-defined and the first and second marginals coincide with μ_{13} and μ_{23} respectively.[12] Lemma A.7 implies $\mathcal{I}^\star(\lambda) = \sup_{\pi\in\mathcal{G}_\lambda} \int_{\mathcal{V}\times\mathcal{S}} \phi_\lambda \, d\pi$. By Lemma A.8, we have for all $\lambda \in \mathbb{R}^2_+$

$$\mathcal{I}^\star(\lambda) = \sup_{\pi \in \Gamma(\Pi(\mu_{13},\mu_{23}),\phi_\lambda)} \int_{\mathcal{V}\times\mathcal{V}} \phi_\lambda \, d\pi.$$

Example 2 of Zhang et al. [57] implies that $\phi_\lambda : \mathcal{V} \times \mathcal{S} \to \mathbb{R}$ satisfies the interchangeability principle with respect to $\Pi(\mu_{13}, \mu_{23})$. As a result, Lemma S.1 in the Online Supplement implies that for all $\lambda \in \mathbb{R}^2_+$

$$\mathcal{I}^\star(\lambda) = \sup_{\gamma \in \Pi(\mu_{13},\mu_{23})} \int_{\mathcal{V}} f_\lambda(v) \, d\gamma(v),$$

where $f_\lambda(v) = \sup_{s\in\mathcal{S}} \phi_\lambda(v,s)$.

From Lemma S.3(i) in the Online Supplement, $\mathcal{I}$ is bounded from below, nondecreasing, and concave. As a result, $\mathcal{I}(\delta) = \infty$ for all $\delta \in \mathbb{R}^2_+$ or $\mathcal{I}(\delta) < \infty$ for all $\delta \in \mathbb{R}^2_+$. In the first case, $\mathcal{I}^\star = \infty$ on $\mathbb{R}^2_+$ by definition and hence we have $\mathcal{I}(\delta) = \inf_{\lambda\in\mathbb{R}^2_+} \{\langle \lambda, \delta\rangle + \mathcal{I}^\star(\lambda)\} = \infty$. For the second case, by Lemma S.4 in the Online Supplement, for all $\delta \in \mathbb{R}^2_{++}$,

$$\mathcal{I}(\delta) = \inf_{\lambda\in\mathbb{R}^2_+} \{\langle \lambda, \delta\rangle + \mathcal{I}^\star(\lambda)\} = \inf_{\lambda\in\mathbb{R}^2_+} \left\{ \langle \lambda, \delta\rangle + \sup_{\gamma\in\Pi(\mu_{13},\mu_{23})} \int_{\mathcal{V}} f_\lambda(v) \, d\gamma(v) \right\},$$

and the proof is complete. □

Lemma A.5. *If $\lambda_1 > 0$ and $\lambda_2 > 0$, then*

$$\sup_{\gamma\in\overline{\mathcal{P}}} I_\lambda[\gamma] = \sup_{\gamma\in\overline{\mathcal{P}}} \sup_{\pi\in\Pi(\mu_{13},\mu_{23},\gamma)} \int_{\mathcal{V}\times\mathcal{S}} \phi_\lambda(v,s') \, d\pi(v,s').$$

Proof of Lemma A.5. The proof is almost identical to that of Lemma A.1, so we only give the sketch. For notational convenience, we write $c_\ell : (s_1, s_2, y_1, y_2, x) \longmapsto c_\ell(s_\ell, (y_\ell, x))$ for $\ell = 1, 2$ and $f : (s_1, s_2, s') \longmapsto f(s')$.

Fix any $\epsilon > 0$ and $\gamma \in \overline{\mathcal{P}}$. Let $K = \{K_1, K_2, K_3\}$ with $K_1 = \{3,4,5\}$, $K_2 = \{1,3,5\}$, and $K_3 = \{2,4,5\}$, and we note that K is decomposable. By Proposition S.1 in the Online Supplement, there is a $\tilde{\pi} \in \Pi(\mu_{13}, \mu_{23}, \gamma)$ satisfying $I_\lambda[\gamma] \le \int_{\mathcal{V}\times\mathcal{S}} \phi_\lambda \, d\tilde{\pi} + \epsilon$. Because $\epsilon > 0$ and $\gamma \in \overline{\mathcal{P}}$ are arbitrary, this shows LHS $\le$ RHS. The proof of LHS $\ge$ RHS is identical to the proof of Lemma A.1. □

Lemma A.6. *If $\lambda_1 > 0$ and $\lambda_2 > 0$, then*

$$\sup_{\gamma\in\overline{\mathcal{P}}} \sup_{\pi\in\Pi(\mu_{13},\mu_{23},\gamma)} \int_{\mathcal{V}\times\mathcal{S}} \phi_\lambda \, d\pi = \sup_{\pi\in\mathcal{G}_\lambda} \int_{\mathcal{V}\times\mathcal{S}} \phi_\lambda \, d\pi.$$

Proof of Lemma A.6. The proof is the same as that of Lemma A.2. □

Lemma A.7. *For all $\lambda \in \mathbb{R}^2_+$, one has $\mathcal{I}^\star(\lambda) = \sup_{\pi\in\mathcal{G}_\lambda} \int_{\mathcal{V}\times\mathcal{S}} \phi_\lambda \, d\pi$.*

Proof of Lemma A.7. The proof is almost the same as Lemma A.3 as long as we replace g with f, φ_λ with ϕ_λ, A_ℓ with B_ℓ, $A_{\ell n}$ with $B_{\ell n}$, and $\mathcal{P}_{\mathrm{D}}$ with $\overline{\mathcal{P}}$, where

$$B_\ell = \{((s_1,s_2),(y_1,y_2,x)) \in \mathcal{V}\times\mathcal{S} : c_\ell(s_\ell,(y_\ell,x)) < \infty\},$$

and

$$B_{\ell n} = \{((s_1,s_2),(y_1,y_2,x)) \in \mathcal{V}\times\mathcal{S} : c_\ell(s_\ell,(y_\ell,x)) < n\},$$

for $\ell = 1, 2$. □

Lemma A.8. *Let $\lambda \in \mathbb{R}^2_+$. If $\phi_\lambda : \mathcal{V} \times \mathcal{S} \to \mathbb{R}$ is interchangeable with respect to $\Pi(\mu_{13}, \mu_{23})$, then*

$$\sup_{\pi\in\mathcal{G}_\lambda} \int_{\mathcal{V}\times\mathcal{S}} \phi_\lambda \, d\pi = \sup_{\pi\in\Gamma(\Pi(\mu_{13},\mu_{23}),\phi_\lambda)} \int_{\mathcal{V}\times\mathcal{S}} \phi_\lambda \, d\pi.$$

Proof of Lemma A.8. The proof is the same as Lemma A.4. □

A.2. Proofs in Section 4

A.2.1. Proof of Theorem 4.

First, assuming that Condition (7) does not hold, we show $\mathcal{I}_{\mathrm{D}}(\delta) = \infty$. Fix any $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2_+$ and $v = (s_1, s_2) \in \mathcal{V}$. For any $B \ge \lambda_1 \vee \lambda_2$, there is $v' = (s_1', s_2') \in \mathcal{V}$ such that

$$g(s_1', s_2') > B[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1, s_1')^{p_1} + \boldsymbol{d}_{\mathcal{S}_2}(s_2, s_2')^{p_2}],$$

and hence

$$\begin{aligned}\varphi_\lambda(v,v') &= g(s_1',s_2') - \lambda_1 \boldsymbol{d}_{\mathcal{S}_1}(s_1,s_1')^{p_1} - \lambda_2 \boldsymbol{d}_{\mathcal{S}_2}(s_2,s_2')^{p_2} \\ &> B[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1,s_1')^{p_1} + \boldsymbol{d}_{\mathcal{S}_2}(s_2,s_2')^{p_1}] - \lambda_1 \boldsymbol{d}_{\mathcal{S}_1}(s_1,s_1')^{p_1} - \lambda_2 \boldsymbol{d}_{\mathcal{S}_2}(s_2,s_2')^{p_2} \\ &\geq B + (B-\lambda_1)\boldsymbol{d}_{\mathcal{S}_1}(s_1,s_1')^{p_1} + (B-\lambda_2)\boldsymbol{d}_{\mathcal{S}_2}(s_2,s_2')^{p_2} \geq B.\end{aligned}$$

This shows that for all $\lambda \in \mathbb{R}_+^2$ and B large enough, we have $g_\lambda(v) = \sup_{v' \in \mathcal{V}} \varphi_\lambda(v,v') \geq B$ for all $v \in \mathcal{V}$. Therefore, by Theorem 2, we have

$$\mathcal{I}_{\mathrm{D}}(\delta) \geq \sup_{\pi \in \Pi(\mu_1,\mu_2)} \int_{\mathcal{V}} g_\lambda(v)\, d\pi(v) \geq B,$$

for all B large enough. As a result, $\mathcal{I}_{\mathrm{D}}(\delta) = \infty$.

Conversely, assuming that the growth condition (7) holds, we show $\mathcal{I}_{\mathrm{D}}(\delta) < \infty$. For all $\pi \in \Sigma_{\mathrm{D}}(\delta)$,

$$\begin{aligned}\int_{\mathcal{V}} f(v)\, d\pi(v) &\leq \int_{\mathcal{S}_1 \times \mathcal{S}_2} M[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star,s_1)^{p_1} + \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star,s_2)^{p_2}]\, d\pi(s_1,s_2) \\ &= M + M\boldsymbol{W}_{p_1}(\pi_1,\delta_{s_1^\star})^{p_1} + M\boldsymbol{W}_{p_2}(\pi_2,\delta_{s_2^\star})^{p_2} \\ &\leq M + \sum_{j=1}^{2} M[\boldsymbol{W}_{p_j}(\pi_j,\mu_j) + \boldsymbol{W}_{p_j}(\mu_j,\delta_{s_j^\star})]^{p_j} < \infty,\end{aligned}$$

where π_j denotes the marginal measure of π on $\mathcal{S}_j$ and $\delta_{s_j^\star}$ denotes the Dirac measure at $s_j^\star \in \mathcal{S}_j$. The last step follows from $\mu_j \in \mathcal{P}_{p_j}(\mathcal{S}_j)$ for $j = 1,2$ and $\pi \in \Sigma_{\mathrm{D}}(\delta)$, that is, $\boldsymbol{W}_{p_j}(\pi_j,\mu_j)^{p_j} \leq \delta_j$ for $j = 1,2$. □

A.2.2. Proof of Theorem 5. First, we assume Condition (8) does not hold and aim to show $\mathcal{I}(\delta) = \infty$. Fix any $\lambda = (\lambda_1,\lambda_2) \in \mathbb{R}_+^2$. For any $v = (s_1,s_2) \in \mathcal{V}$ and $B \geq \lambda_1 \vee \lambda_2$, there exists $s' = (y_1',y_2',x')$ such that

$$f(s') \geq B[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1,s_1')^{p_1} + \boldsymbol{d}_{\mathcal{S}_2}(s_2,s_2')^{p_2}].$$

Therefore,

$$\begin{aligned}\phi_\lambda(v,s') &= f(s') - \lambda_1 \boldsymbol{d}_{\mathcal{S}_1}(s_1,s_1')^{p_1} - \lambda_2 \boldsymbol{d}_{\mathcal{S}_2}(s_2,s_2')^{p_2} \\ &\geq B + (B-\lambda_1)\boldsymbol{d}_{\mathcal{S}_1}(s_1,s_1')^{p_1} + (B-\lambda_2)\boldsymbol{d}_{\mathcal{S}_2}(s_2,s_2')^{p_2} \geq B.\end{aligned}$$

As a result, $f_\lambda(v) = \sup_{s' \in \mathcal{S}} \phi_\lambda(v,s') \geq B$ for all $v \in \mathcal{V}$ and all B large enough. Because $B > 0$ is arbitrary, we must have $\sup_{\varpi \in \Pi(\mu_{13},\mu_{23})} \int_{\mathcal{V}} f_\lambda(v) d\varpi(v) = \infty$. By Theorem 3, we have $\mathcal{I}(\delta) = \infty$.

Conversely, we show that the condition (8) implies $\mathcal{I}(\delta) < \infty$. For any $\gamma \in \Sigma(\delta)$,

$$\begin{aligned}\int_{\mathcal{S}} f(s)\, d\gamma(s) &\leq \int_{\mathcal{S}} M[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star,s_1)^{p_1} + \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star,s_2)^{p_2}]\, d\gamma(s) \\ &\leq M + M\boldsymbol{W}_{p_1}(\delta_{s_1^\star},\gamma_{13})^{p_1} + M\boldsymbol{W}_{p_2}(\delta_{s_2^\star},\gamma_{23})^{p_2} \\ &\leq M + \sum_{j=1}^{2} M[\boldsymbol{W}_{p_j}(\delta_{s_j^\star},\mu_{j3}) + \boldsymbol{W}_{p_j}(\mu_{j3},\gamma_{j3})]^{p_j} < \infty,\end{aligned}$$

where γ_{j3} is the marginal measure of γ on $\mathcal{S}_j = \mathcal{Y}_j \times \mathcal{X}$ and $\delta_{s_j^\star}$ is the Dirac measure concentrated at $\{s_j^\star\}$. The last step follows from $\gamma \in \Sigma(\delta)$ and $\mu_{j3} \in \mathcal{P}_{p_j}(\mathcal{S}_j)$ for $j = 1,2$. □

A.2.3. Proof of Theorem 6. In this section, we first prove the weak compactness of $\Sigma_{\mathrm{D}}(\delta)$ for all $\delta \in \mathbb{R}_+^2$ when $\mathcal{S}_1$ and $\mathcal{S}_2$ are both proper and $c_j = \boldsymbol{d}_{\mathcal{S}_j}^{p_j}$ for some $p_j \geq 1$. As a result, $\boldsymbol{K}_j = \boldsymbol{W}_{p_j}^{p_j}$ and the set $\Sigma_{\mathrm{D}}(\delta)$ can be written as

$$\Sigma_{\mathrm{D}}(\delta) := \{\gamma \in \mathcal{P}(\mathcal{S}_1 \times \mathcal{S}_2) : \boldsymbol{W}_{p_1}(\gamma_1,\mu_1) \leq \delta_1^{1/p_1},\, \boldsymbol{W}_{p_2}(\gamma_2,\mu_2) \leq \delta_2^{1/p_2}\}.$$

For any Polish metric space $\mathcal{X}$, let $B_{\mathcal{P}_p(\mathcal{X})}(\mu,\delta) := \{\gamma \in \mathcal{P}(\mathcal{X}) : \boldsymbol{W}_p(\mu,\gamma) \leq \delta\}$ denote the ball centered at μ in Wasserstein space $\mathcal{P}_p(\mathcal{X})$. When there is no ambiguity, we will abbreviate this notation by referring to $B_p(\mu,\delta)$.

Proposition A.1. *Suppose Assumptions 5(i), 6, and 7 hold. Then, $\Sigma_{\mathrm{D}}(\delta)$ is weakly compact.*

Proof of Proposition A.1. Theorem 1 of Yue et al. [56] implies that $B_p(\mu,\delta)$ is weakly compact whenever μ has a finite p-th moment. As a result, the set $\Sigma_{\mathrm{D}}(\delta)$ can be written as

$$\Sigma_{\mathrm{D}}(\delta) = \Pi(\mathcal{B}_1,\mathcal{B}_2), \quad \text{where } \mathcal{B}_1 = B_{p_1}(\mu_1,\delta_1^{1/p_1}) \text{ and } \mathcal{B}_2 = B_{p_2}(\mu_2,\delta_2^{1/p_2}).$$

By Assumption 7, $\mathcal{B}_1$ and $\mathcal{B}_2$ are weakly compact in $\mathcal{P}(\mathcal{S}_1)$ and $\mathcal{P}(\mathcal{S}_2)$, respectively. Hence, they are both uniformly tight by Prokhorov's theorem. By lemma 4.4 of Villani [52], $\Sigma_{\mathrm{D}}(\delta)$ is tight in $\mathcal{P}(\mathcal{S}_1 \times \mathcal{S}_2)$. By Prokhorov's theorem again, $\Sigma_{\mathrm{D}}(\delta)$ has a compact closure under the topology of weak convergence. To show the weak compactness of $\Sigma_{\mathrm{D}}(\delta)$, it suffices to show it is closed.

Let $\pi^n \in \Sigma_D(\delta) \equiv \Pi(\mathcal{B}_1, \mathcal{B}_2)$ be a sequence converging weakly to $\pi^\infty \in \mathcal{P}(\mathcal{S}_1 \times \mathcal{S}_2)$. We have

$$\boldsymbol{W}_{p_1}(\pi_1^n, \mu_1) \le \delta_1^{1/p_1} \text{ and } \boldsymbol{W}_{p_2}(\pi_2^n, \mu_2) \le \delta_1^{1/p_2}.$$

Let π_j^n denote the marginal distribution of π^n on $\mathcal{S}_j$. For any open U_1 in $\mathcal{S}_1$, the Portmanteau theorem implies

$$\liminf_{n\to\infty} \pi_1^n(U_1) = \liminf_{n\to\infty} \pi^n(U_1 \times \mathcal{S}_2) \ge \pi^\infty(U_1 \times \mathcal{S}_2) = \pi_1^\infty(U_1).$$

This shows π_1^n weakly converges to π_1^∞. Moreover, $\boldsymbol{W}_{p_1}(\pi_1^\infty, \mu_1) \le \delta_1^{1/p_1}$ can be seen from the weakly closedness of $\mathcal{B}_1$. Using the identical argument, we can show π_2^n weakly converges to π_2^∞ and $\boldsymbol{W}_{p_2}(\pi_2^\infty, \mu_2) \le \delta_1^{1/p_2}$. This shows $\pi^\infty \in \boldsymbol{W}_{p_2}(\pi_2^n, \mu_2) \le \delta_1^{1/p_2}$ and hence $\Sigma_D(\delta)$ is weakly closed. □

The weak compactness of $\Sigma_D(\delta)$ does not depend on the functional forms of metrics $\boldsymbol{d}_{\mathcal{S}_1}$ and $\boldsymbol{d}_{\mathcal{S}_2}$. Essentially, the topological properties of $\mathcal{S}_1$ and $\mathcal{S}_2$, mainly properness, determine the weak compactness of $\Sigma_D(\delta)$.

Proof of Theorem 6. Because Proposition A.1 implies that $\Sigma_D(\delta)$ is weakly compact, by the Weierstrass theorem, it suffices to show $\pi \mapsto \int_{\mathcal{V}} g\, d\pi$ is weakly upper semicontinuous. Let $\{\pi^k\}_{k=1}^\infty$ be any sequence in $\Sigma_D(\delta)$ that weakly converges to $\pi^\infty \in \Sigma_D(\delta)$; we show $\limsup_{n\to\infty} \int_{\mathcal{V}} g\, d\pi^k \le \int_{\mathcal{V}} g\, d\pi^\infty$. For any $\rho > 0$, define an auxiliary function $f_\rho : \mathcal{V} \to \mathbb{R}$ as $g_\rho(v) = f(v) \wedge [M(1 + \rho^{p_0'} + \rho^{p_1'})]$. Let $A_1 = \{(s_1, s_2) \in \mathcal{V} : \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1) \ge \rho\}$ and $A_2 = \{(s_1, s_2) \in \mathcal{V} : \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star, s_2) \ge \rho\}$. It is easy to verify that for all $v \in \mathcal{V}$,

$$|g(v) - g_\rho(v)| \le \begin{cases} M[\boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1)^{p_1'} + \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star, s_2)^{p_2'}] & \text{if } v \in A_1 \cap A_2, \\ M\, \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1)^{p_1'} & \text{if } v \in A_1 \cap A_2^c, \\ M\, \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star, s_2)^{p_2'} & \text{if } v \in A_1^c \cap A_2, \\ 0 & \text{otherwise.} \end{cases}$$

For any $\pi \in \Sigma_D(\delta)$, we have

$$\begin{aligned} \left| \int_{\mathcal{V}} g\, d\pi - \int_{\mathcal{V}} g_\rho\, d\pi \right| &\le \int_{\mathcal{V}} |g - g_\rho|\, d\pi \\ &\le \int_{A_1 \cap A_2} |g - g_\rho|\, d\pi + \int_{A_1 \cap A_2^c} |g - g_\rho|\, d\pi + \int_{A_1^c \cap A_2} |g - g_\rho|\, d\pi. \end{aligned}$$

By lemma 1 in Yue et al. [56], there exists $B > 0$ such that $W_{p_j}(\pi_j, \delta_{s_j^\star})^{p_j} \le B$ for $j = 1, 2$ and all $\pi \in \Sigma_D(\delta)$, where π_j is the marginal of π on $\mathcal{S}_j$ and $\delta_{s_j^\star}$ is a Dirac measure at $\{s_j^\star\}$. Therefore, we have

$$\begin{aligned} \int_{A_1 \cap A_2^c} |g - g_\rho|\, d\pi &\le M \int_{A_1 \cap A_2^c} \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1)^{p_1'} d\pi \le M\rho^{p_1 - p_1'} \int_{A_1 \cap A_2^c} \boldsymbol{d}_{\mathcal{S}_1}(s_1, s_1^\star)^{p_1}\, d\pi \\ &\le M\rho^{p_1 - p_1'} W_{p_1}(\pi_1, \delta_{s_1^\star})^{p_1} \le B\rho^{p_1' - p_1}. \end{aligned}$$

Similarly, we can show $\int_{A_1^c \cap A_2} |g - g_\rho| d\pi \le B\rho^{p_2' - p_2}$ and

$$\begin{aligned} \int_{A_1 \cap A_2} |g - g_\rho|\, d\pi &\le \int_{A_1 \cap A_2} M[\boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1)^{p_1'} + \boldsymbol{d}_{\mathcal{S}_2}(s_2, s_1^\star)^{p_2'}] d\pi(s_1, s_2) \\ &\le B(\rho^{p_1' - p_1} + \rho^{p_2' - p_2}). \end{aligned}$$

Therefore, we have for all $\pi \in \Sigma_D(\delta)$

$$\left| \int_{\mathcal{V}} g\, d\pi - \int_{\mathcal{V}} g_\rho\, d\pi \right| \le \int_{\mathcal{V}} |g - g_\rho|\, d\pi \le 2B(\rho^{p_1' - p_1} + \rho^{p_2' - p_2}).$$

For any $\epsilon > 0$, there is a $\rho > 0$ large enough such that $4B(\rho^{p_1' - p_1} + \rho^{p_2' - p_2}) < \epsilon/2$. By lemma 3 in Yue et al. [56], we have $\limsup_{k\to\infty} \int_{\mathcal{V}} g_\rho\, d\pi^k \le \int_{\mathcal{V}} g_\rho\, d\pi^\infty$, and hence there is a $k(\epsilon)$ large enough such that

$$\int_{\mathcal{V}} g_\rho\, d\pi^k - \int_{\mathcal{V}} g_\rho\, d\pi^\infty < \frac{\epsilon}{2}, \quad \text{for all } k > k(\epsilon).$$

Consequently, for all $k > k(\epsilon)$, the following holds:

$$\begin{aligned} \int_{\mathcal{V}} g\, d\pi^k - \int_{\mathcal{V}} g\, d\pi^\infty &\le \int_{\mathcal{V}} |g - g_\rho|\, d\pi^k + \int_{\mathcal{V}} g_\rho\, d\pi^k - \int_{\mathcal{V}} g_\rho\, d\pi^\infty + \int_{\mathcal{V}} |g_\rho - g|\, d\pi^\infty \\ &\le 4B(\rho^{p_1' - p_1} + \rho^{p_2' - p_2}) + \int_{\mathcal{V}} g_\rho\, d\pi^k - \int_{\mathcal{V}} g_\rho\, d\pi^\infty < \epsilon. \end{aligned}$$

Because ϵ is arbitrary, we must have $\limsup_{k\to\infty} \int_{\mathcal{V}} g\, d\pi^k \le \int_{\mathcal{V}} g\, d\pi^\infty$. This completes the proof. □

A.2.4. Proof of Theorem 7. Here, we will only show that $\Sigma(\delta)$ is weakly compact. This is because the upper semicontinuity of $\gamma \to \int f d\gamma$ over $\gamma \in \Sigma(\delta)$ can be shown using the same argument for the proof of Theorem 6. We write

$$\Sigma(\delta) = \{\gamma \in \mathcal{P}(\mathcal{S}) : \boldsymbol{W}_{p_1}(\gamma_1, \mu_1) \le \delta_1^{1/p_1},\ \boldsymbol{W}_{p_2}(\gamma_2, \mu_2) \le \delta_2^{1/p_2}\}.$$

Lemma A.9. *For $j = 1, 2$, let $\mathcal{G}_j$ be a uniformly tight subset of $\mathcal{P}(\mathcal{S}_j)$. Then the following set*

$$\Gamma(\mathcal{G}_1, \mathcal{G}_2) := \{\gamma \in \mathcal{P}(\mathcal{S}) : \gamma_{13} \in \mathcal{G}_1, \gamma_{23} \in \mathcal{G}_2\}$$

is tight in $\mathcal{P}(\mathcal{S})$.

Proof of Lemma A.9. First, we assume there exist $\mu \in \mathcal{G}_1$ and $\nu \in \mathcal{G}_2$ such that $\mu(\mathcal{Y}_1 \times A) = \nu(\mathcal{Y}_2 \times A)$ for all $A \in \mathcal{B}_{\mathcal{X}}$; that is, μ and ν have the same marginal distribution on $\mathcal{X}$. Otherwise, $\Gamma(\mathcal{G}_1, \mathcal{G}_2)$ will be empty and hence the statement holds trivially.

Because $\mathcal{G}_1$ is uniformly tight, then for any $\epsilon > 0$, there is a compact set $K_\epsilon \subset \mathcal{S}_1 \equiv \mathcal{Y}_1 \times \mathcal{X}$ such that $\mu(K_\epsilon^c) \le \epsilon$ for all $\mu \in \mathcal{G}_1$. Similarly, there is a compact set $L_\epsilon \subset \mathcal{S}_2 \equiv \mathcal{Y}_2 \times \mathcal{X}$ such that $\nu(L_\epsilon^c) \le \epsilon$ for all $\nu \in \mathcal{G}_2$. Moreover, define a mapping $\sigma : \mathcal{S} \to \mathcal{S}$ as $\sigma : (y_1, y_2, x) \mapsto (y_1, x, y_2)$. Trivially, σ is a homeomorphism (a continuous mapping whose inverse is also continuous) from $\mathcal{S}$ to $\mathcal{S}$. Let $E_\epsilon = \sigma^{-1}(K_\epsilon \times \mathcal{Y}_2)$ and $G_\epsilon = \mathcal{Y}_1 \times L_\epsilon$. Explicitly, $(y_1, y_2, x) \in E_\epsilon \Leftrightarrow (y_1, x) \in K_\epsilon$. Fix any $\gamma \in \Gamma(\mathcal{G}_1, \mathcal{G}_2)$; let $S = (Y_1, Y_2, X)$ be a random variable with γ as its law, that is, $\mathrm{Law}(S) = \gamma$. We must have $\gamma_{j3} \in \mathcal{G}_j$ for $j = 1, 2$. Then,

$$\begin{aligned}
\mathbb{P}[S \notin E_\epsilon \cap G_\epsilon] &\le \mathbb{P}[S \notin E_\epsilon] + \mathbb{P}[S \notin G_\epsilon] \\
&= \mathbb{P}[(Y_1, Y_2, X) \notin E_\epsilon] + \mathbb{P}[(Y_1, Y_2, X) \notin G_\epsilon] \\
&= \mathbb{P}[(Y_1, X) \notin K_\epsilon] + \mathbb{P}[(Y_2, X) \notin L_\epsilon] \\
&\le \gamma_{13}(K_\epsilon^c) + \gamma_{23}(L_\epsilon^c) \\
&\le 2\epsilon.
\end{aligned}$$

The desired result follows from the compactness of $E_\epsilon \cap G_\epsilon$ in $\mathcal{S}$. To see this, we note $\mathrm{proj}_{\mathcal{Y}_1} : (y_1, x) \mapsto y_1$ is continuous from $\mathcal{S}_1$ to $\mathcal{Y}_1$ and hence $\mathrm{proj}_{\mathcal{Y}_1}(K_\epsilon)$ is compact. As a result, $\mathrm{proj}_{\mathcal{Y}_1}(K_\epsilon) \times L_\epsilon$ is compact, because $E_\epsilon \cap G_\epsilon$ is a subset of a compact set and its compactness follows from the closedness of E_ϵ and G_ϵ. □

Proposition A.2. *Suppose Assumptions 5(ii), 6, and 7 hold. Then, $\Sigma(\delta)$ is weakly compact.*

Proof of Proposition A.2. By abuse of notations, let $\mathcal{B}_1 = B_{p_1}(\mu_{13}, \delta_1^{1/p_1})$ and $\mathcal{B}_2 = B_{p_2}(\mu_{23}, \delta_2^{1/p_2})$. We can rewrite $\Sigma(\delta) = \Gamma(\mathcal{B}_1, \mathcal{B}_2)$. By Lemma A.9, $\Sigma(\delta)$ is tight and hence has a compact closure under weak topology. Using a similar argument in the proof of Proposition A.1, we can show $\Sigma(\delta)$ is weakly closed. Therefore, $\Sigma(\delta)$ is weakly compact in $\mathcal{P}(\mathcal{S})$. □

A.2.5. Proof of Proposition 1. We focus on $\Theta(\delta)$ because the proof of $\Theta_{\mathrm{D}}(\delta)$ is identical to that of $\Theta(\delta)$. The proof of Proposition 1 for $\Theta(\delta)$ follows from the following two lemmas.

Lemma A.10. *Suppose that the assumptions in Proposition 1 hold. Then, the linear functional $T : \Sigma(\delta) \to \mathbb{R}$ given by $\pi \mapsto \int_{\mathcal{S}} f d\pi$ is continuous.*

Proof of Lemma A.10. Because $\mu_{\ell 3}$ has finite p_ℓ-th moment, then for all $\pi \in \Sigma(\delta)$, $\pi_{\ell 3}$, that is, the projection onto $\mathcal{Y}_\ell \times \mathcal{X}$ also has finite p_ℓ-th moment. Define a function $h : \mathcal{S} \to \mathbb{R}$ as

$$h(s) = M[1 + \boldsymbol{d}_{\mathcal{S}_1}(s_1^\star, s_1)^{p_1'} + \boldsymbol{d}_{\mathcal{S}_2}(s_2^\star, s_2)^{p_2'}],$$

where $s = (y_1, y_2, x)$, $s_1 = (y_1, x)$, and $s_2 = (y_2, x)$. We note $h \in L^1(\pi)$ for all $\pi \in \Sigma(\delta)$. Using the identical argument in the proof of Theorem 6, we can show that $\pi \mapsto \int f d\pi$ is upper semicontinuous on $\Sigma(\delta)$. By replacing f by $-f$, we can see that $\pi \mapsto \int(-f)d\pi$ is upper semicontinuous and hence $\pi \mapsto \int f d\pi$ is lower semicontinuous on $\Sigma(\delta)$. As a result, $\pi \mapsto \int f d\pi$ is continuous on $\Sigma(\delta)$. □

Lemma A.11. *Suppose that Assumptions 5(ii) and 6 hold. Then $\Sigma(\delta)$ is connected under weak topology.*

Proof of Lemma A.11. Fix any π and π' in $\Sigma(\delta)$. It suffices to show $\nu : t \mapsto t\pi + (1-t)\pi'$ is continuous from [0, 1] into $\Sigma(\delta)$. We note $\Sigma(\delta) \subset \mathcal{P}_p(\mathcal{S})$ is metrizable under $\boldsymbol{W}_p$ for $p = p_1 \wedge p_2$. Fix any $t_0 \in [0, 1]$. Let $t_1 \neq t_0$ be any point in [0, 1] such that $\Delta = |t_1 - t_0| > 0$ is sufficiently small. Without loss of generality, we assume $t_0 < t_1$. For simplicity, we write $\gamma = t_0\pi + (1 - t_1)\pi' \ge 0$. By the triangle inequality,

$$\begin{aligned}
\boldsymbol{W}_p(\nu(t_0), \nu(t_1)) &= \boldsymbol{W}_p(\nu(t_0), \gamma + \Delta\pi') \\
&\le (1-\Delta)\boldsymbol{W}_p(\nu(t_0), (1-\Delta)^{-1}\gamma) + \underbrace{\Delta \boldsymbol{W}_p(\nu(t_0), \pi')}_{=O(\Delta)}.
\end{aligned}$$

Consider the following derivation:

$$\boldsymbol{W}_p(\nu(t_0),(1-\Delta)^{-1}\gamma) = \boldsymbol{W}_p\left(\nu(t_0), \underbrace{\frac{\nu(t_0)-\Delta\pi'}{1-\Delta}}_{=\rho_\Delta}\right) = \boldsymbol{W}_p((1-\Delta)\rho_\Delta + \Delta\pi', \rho_\Delta)$$
$$\le \Delta \boldsymbol{W}_p(\pi', \rho_\Delta) = \Delta \boldsymbol{W}_p\left(\pi', \frac{\nu(t_0)-\Delta\pi'}{1-\Delta}\right).$$

Because $\lim_{\Delta\to 0}\frac{\nu(t_0)-\Delta\pi'}{1-\Delta} = \nu(t_0)$ in weak topology induced by $\boldsymbol{W}_p$, then

$$\lim_{\Delta\to 0} \boldsymbol{W}_p\left(\pi', \frac{\nu(t_0)-\Delta\pi'}{1-\Delta}\right) = \boldsymbol{W}_p(\pi', \nu(t_0)) < \infty.$$

As a result,

$$\boldsymbol{W}_p(\nu(t_0),(1-\Delta)^{-1}\gamma) \le \Delta \boldsymbol{W}_p\left(\pi', \frac{\nu(t_0)-\Delta\pi'}{1-\Delta}\right) \to 0, \quad \text{as } \Delta \to 0,$$

and hence

$$\boldsymbol{W}_p(\nu(t_0), \nu(t_1)) \to 0, \quad \text{as } \Delta \to 0.$$

Interchanging the role of t_0 and t_1, we can show the case when $\boldsymbol{W}_p(\nu(t_0),\nu(t_1)) \to 0$ as $\Delta = |t_1 - t_0| \to 0$. This shows $\nu : t \longmapsto t\pi + (1-t)\pi'$ is continuous on $[0, 1]$. So, $\Sigma(\delta)$ is path-connected and hence connected under weak topology. □

A.3. Proofs in Section 5

A.3.1. Proof of Theorem 8. Note that the proof of Lemma S.4 in the Online Supplement implies that if $\mathcal{I}_{\mathrm{D}}(\delta)$ is finite for some $\delta > 0$, then $\mathcal{I}_{\mathrm{D}}(\delta)$ is finite for all $\delta > 0$ because $\mathcal{I}(\delta)$ is concave.

Lemma A.12. *Suppose that Assumptions 2 and 8 hold. Then for any $\delta = (\delta_1, \delta_2) \in \mathbb{R}^2_+$, we have*

$$0 \le \mathcal{I}_{\mathrm{D}}(\delta_1, \delta_2) - \mathcal{I}_{\mathrm{D}}(0,0) \le \Psi(\delta_1, \delta_2).$$

Moreover, $\mathcal{I}_{\mathrm{D}}$ is continuous on $(0,0)$.

Proof of Lemma A.12. Fix any $\tilde{\gamma} \in \Sigma_{\mathrm{D}}(\delta)$ and any $\epsilon > 0$. We can construct random variables $\tilde{V} = (\tilde{S}_1, \tilde{S}_2) \in \mathcal{V}$ with $\tilde{\gamma} = \mathrm{Law}(\tilde{V})$ and write $\tilde{\gamma}_j = \mathrm{Law}(\tilde{S}_j)$ for $j \in [2]$. Let $K = \{K_1, K_2, K_3\}$ with $K_1 = \{1,3\}$, $K_2 = \{2,4\}$, and $K_3 = \{3,4\}$. It is easy to see K is decomposable, and Proposition S.1 in the Online Supplement implies that there are random variables $(V, \tilde{V}) = (S_1, S_2, \tilde{S}_1, \tilde{S}_2) \in \mathcal{V} \times \mathcal{V}$ such that $\mu_1 = \mathrm{Law}(S_1)$, $\mu_2 = \mathrm{Law}(S_2)$, and $\mathbb{E}[c_j(S_j, \tilde{S}_j)] \le \boldsymbol{K}_j(\mu_j, \tilde{\gamma}_j) + \epsilon \le \delta_j + \epsilon$ for $j \in [2]$. Let π denote the law of $(V, \tilde{V})$. Therefore, with $\gamma = \mathrm{Law}(S_1, S_2) \in \Sigma_{\mathrm{D}}(0)$, we have

$$\begin{aligned}\int_{\mathcal{V}} g\, d\tilde{\gamma} - \mathcal{I}_{\mathrm{D}}(0,0) &\le \int_{\mathcal{V}} g\, d\tilde{\gamma} - \int_{\mathcal{V}} g\, d\gamma = \int_{\mathcal{V}\times\mathcal{V}} [g(v) - g(\tilde{v})]\, d\pi(v, \tilde{v}) \\ &= \mathbb{E}[g(V) - g(\tilde{V})] \le \mathbb{E}[\Psi(c_1(S_1, \tilde{S}_1), c_2(S_2, \tilde{S}_2))] \\ &\le \Psi(\mathbb{E}[c_1(S_1, \tilde{S}_1)], \mathbb{E}[c_2(S_2, \tilde{S}_2)]) \\ &\le \Psi(\delta_1 + \epsilon, \delta_2 + \epsilon).\end{aligned}$$

Because the measure $\tilde{\gamma} \in \Sigma_{\mathrm{D}}(\delta)$ is arbitrary, we must have

$$\mathcal{I}_{\mathrm{D}}(\delta_1, \delta_2) - \mathcal{I}_{\mathrm{D}}(0,0) = \sup_{\tilde{\gamma} \in \Sigma_{\mathrm{D}}(\delta)} \int_{\mathcal{V}} g\, d\tilde{\gamma} - \mathcal{I}_{\mathrm{D}}(0,0) \le \Psi(\delta_1 + \epsilon, \delta_2 + \epsilon).$$

Because Ψ is continuous and $\epsilon > 0$ is arbitrary, then $\mathcal{I}_{\mathrm{D}}(\delta_1, \delta_2) - \mathcal{I}_{\mathrm{D}}(0,0) \le \Psi(\delta_1, \delta_2)$. The monotonicity of $\mathcal{I}_{\mathrm{D}}$ implies $\mathcal{I}_{\mathrm{D}}(\delta_1, \delta_2) \ge \mathcal{I}_{\mathrm{D}}(0,0)$. In addition, the continuity of $\mathcal{I}_{\mathrm{D}}$ at $(0,0)$ follows from the continuity of Ψ at $(0,0)$ and letting $(\delta_1, \delta_2) \to (0,0)$. □

In fact, Lemma S.3(i) in the Online Supplement and the proof of Lemma A.12 imply the effective domain of $\mathcal{I}_{\mathrm{D}}$ is either $\mathbb{R}^2_+$ or $\emptyset$ because $\mathcal{I}_{\mathrm{D}}$ is nondecreasing and concave.

Lemma A.13. *Suppose that Assumptions 2 and 8 hold, and $\mathcal{I}_{\mathrm{D}}(\delta)$ is finite for some $\delta \in \mathbb{R}^2_{++}$. If $\eta_0 > \eta \ge 0$ and $\delta \ge 0$, one has*

$$0 \le \mathcal{I}_{\mathrm{D}}(\eta_0, \delta) - \mathcal{I}_{\mathrm{D}}(\eta, \delta) \le \Psi(\eta_0 - \eta, 0)$$

and

$$0 \le \mathcal{I}_{\mathrm{D}}(\delta, \eta_0) - \mathcal{I}_{\mathrm{D}}(\delta, \eta) \le \Psi(0, \eta_0 - \eta).$$

Proof of Lemma A.13. We assume that for all $\eta,\delta\geq 0$, there exists $\gamma^{\eta,\delta}\in\Sigma_{\mathrm{D}}(\eta,\delta)$ such that $\mathcal{I}_{\mathrm{D}}(\eta,\delta)=\int g\,d\gamma^{\eta,\delta}$. Otherwise, because of the continuity of Ψ on $\mathbb{R}^2_+$, we can repeat the proof with ϵ-approximation optimizer and let $\epsilon\downarrow 0$. In addition, because $\mathcal{I}_{\mathrm{D}}(\delta)<\infty$ for some $\delta\in\mathbb{R}^2_+$, the $\mathcal{I}_{\mathrm{D}}(\delta)<\infty$ for all $\delta\in\mathbb{R}^2_+$.

Let $\gamma_\ell^{\eta,\delta}$ denote the marginal of $\gamma^{\eta_0,\delta}$ on $\mathcal{S}_\ell$. Fix $\gamma^{\eta_0,\delta}\in\mathcal{P}(\mathcal{S}_1\times\mathcal{S}_2)$. Define a probability measure $\gamma_1^\star$ on $\mathcal{S}_1$ as

$$\gamma_1^\star=\left(\frac{\eta}{\eta_0}\right)\gamma_1^{\eta_0,\delta}+\left(\frac{\eta_0-\eta}{\eta_0}\right)\mu_1.$$

By definition, $\boldsymbol{K}_1(\gamma_1^{\eta_0,\delta},\mu_1)\leq\eta_0$ and $\boldsymbol{K}_2(\gamma_2^{\eta_0,\delta},\mu_2)\leq\delta$. By convexity of $\nu\longmapsto\boldsymbol{K}_1(\nu,\mu_1)$, we have $\boldsymbol{K}_1(\gamma_1^\star,\mu_1)\leq\eta$ and $\boldsymbol{K}_1(\gamma_1^\star,\gamma_1^{\eta_0,\delta})\leq\eta_0-\eta$. Without loss of generality, suppose there is an optimal coupling $\nu\in\Pi(\gamma_1^{\eta,\delta},\gamma_1^\star)$ such that

$$\boldsymbol{K}_1(\gamma_1^{\eta_0,\delta},\gamma_1^\star)=\int_{\mathcal{S}_1\times\mathcal{S}_1}c_1\,d\nu.$$

By the gluing lemma, we can construct random variables $(S_1,S_2,\tilde{S}_1)\in\mathcal{V}\times\mathcal{S}_1$ with the law $\hat{\pi}\equiv\mathrm{Law}(S_1,S_2,\tilde{S}_1)$ such that

$$\hat{\pi}_{1,2}=\mathrm{Law}(S_1,S_2)=\gamma^{\eta_0,\delta},\quad \hat{\pi}_{1,3}=\mathrm{Law}(S_1,\tilde{S}_1)=\nu\in\Pi(\gamma_1^{\eta,\delta},\gamma_1^\star),$$

and

$$\boldsymbol{K}_1(\gamma_1,\gamma_1^{\eta_0,\delta})=\mathbb{E}[c_1(S_1,\tilde{S}_1)]\leq\eta_0-\eta.$$

Let $\gamma=\mathrm{Law}(\tilde{S}_1,S_2)\in\mathcal{P}(\mathcal{V})$, and it is obvious that $\tilde{\gamma}_1\in\Sigma_{\mathrm{D}}(\eta,\delta)$. Next, consider the following derivation:

$$\begin{aligned}\mathcal{I}_{\mathrm{D}}(\eta_0,\delta)-\mathcal{I}_{\mathrm{D}}(\eta,\delta)&\leq\int g(v)\,d\gamma^{\eta_0,\delta}(v)-\int g(v)\,d\gamma(v)\\&=\int_{\mathcal{V}\times\mathcal{V}}[g(s_1,s_2)-g(\tilde{s}_1,s_2)]d\hat{\pi}(s_1,s_2,\tilde{s}_1)\\&=\mathbb{E}[g(S_1,S_2)-g(\tilde{S}_1,S_2)]\leq\mathbb{E}[\Psi(c_1(S_1,\tilde{S}_1),0)]\\&\leq\Psi(\mathbb{E}[c_1(S_1,\tilde{S}_1)],0)\leq\Psi(\eta_0-\eta,0).\end{aligned}$$

Using the same argument, we can show $\mathcal{I}_{\mathrm{D}}(\delta,\eta_0)-\mathcal{I}_{\mathrm{D}}(\delta,\eta)\leq\Psi(0,\eta_0-\eta)$. □

Now we present the proof of Theorem 8.

Proof of Theorem 8. Because $\mathcal{I}_{\mathrm{D}}$ is concave on $\mathbb{R}^2_+$, then $\mathcal{I}_{\mathrm{D}}$ is continuous on $\mathbb{R}^2_{++}$. By Lemma A.12, $\mathcal{I}_{\mathrm{D}}$ is continuous at $(0,0)$. Let $E_0=\{(x,0)\in\mathbb{R}^2_+:x>0\}$ and $E_1=\{(0,y)\in\mathbb{R}^2_+:y>0\}$. To complete the proof, it suffices to show $\mathcal{I}_{\mathrm{D}}$ is continuous at all $\delta\in E_0\cup E_1$.

Fix any $(\eta,0)\in E_0$. For any $\eta_0\geq\eta$ and any $\delta>0$, we have

$$\begin{aligned}\mathcal{I}_{\mathrm{D}}(\eta_0,\delta)-\mathcal{I}_{\mathrm{D}}(\eta,0)&=\mathcal{I}_{\mathrm{D}}(\eta_0,\delta)-\mathcal{I}_{\mathrm{D}}(\eta,\delta)+\mathcal{I}_{\mathrm{D}}(\eta,\delta)-\mathcal{I}_{\mathrm{D}}(\eta,0)\\&\leq\Psi(\eta_0-\eta,0)+\Psi(0,\delta)=\Psi(|\eta_0-\eta|,0)+\Psi(0,\delta).\end{aligned}$$

Similarly, for any $\eta_0<\eta$ and $\delta>0$,

$$\mathcal{I}_{\mathrm{D}}(\eta,\delta)-\mathcal{I}_{\mathrm{D}}(\eta_0,0)\leq\Psi(|\eta_0-\eta|,0)+\Psi(0,\delta).$$

This shows that for all η,η_0 and δ in $(0,\infty)$, one has

$$|\mathcal{I}_{\mathrm{D}}(\eta_0,\delta)-\mathcal{I}_{\mathrm{D}}(\eta,0)|\leq\Psi(|\eta_0-\eta|,0)+\Psi(0,\delta).$$

The continuity of $\mathcal{I}_{\mathrm{D}}$ at $(\eta,0)$ follows from the continuity of Ψ at $(0,0)$ and letting $(\eta_0,\delta)\to(\eta,0)$. Because $(\eta,0)\in E_0$ is arbitrary, $\mathcal{I}_{\mathrm{D}}$ is continuous at all $x\in E_0$. Using the same argument, we can show $\mathcal{I}_{\mathrm{D}}$ is continuous at all $x\in E_1$. The desired result follows. □

A.3.2. Proof of Theorem 9. Note that the proof of Lemma S.4 in the Online Supplement implies that If $\mathcal{I}(\delta)$ is finite for some $\delta\in\mathbb{R}^2_{++}$, then $\mathcal{I}(\delta)$ is finite for all $\delta\in\mathbb{R}^2_{++}$ because $\mathcal{I}(\delta)$ is concave. Based on this, we give the following lemma that is used to show the continuity of $\mathcal{I}$.

Lemma A.14. *Let $\delta\geq 0$, $\eta_0>\eta\geq 0$. Suppose that $\mathcal{I}(\delta)<\infty$ for some $\delta\in\mathbb{R}^2_{++}$. Under Assumptions 3, 5(ii), 6, 9, and 10, there is a constant $M>0$ such that*

$$\mathcal{I}(\eta_0,\delta)-\mathcal{I}(\eta,\delta)\leq\Psi_1(\eta_0-\eta,M(1-\eta/\eta_0)),$$

and

$$\mathcal{I}(\delta,\eta_0)-\mathcal{I}(\delta,\eta)\leq\Psi_2(M(1-\eta/\eta_0),\eta_0-\eta).$$

Proof of Lemma A.14. For simplicity, assume that for any $\eta,\delta\geq 0$, one has $\gamma^{\eta,\delta}=\operatorname{argmax}_{\gamma\in\Sigma(\eta,\delta)}\int_{\mathcal{S}} f d\gamma$, equivalently, $\mathcal{I}(\eta,\delta)=\int_{\mathcal{S}} f d\gamma^{\eta,\delta}$. Otherwise, because of the global continuity of Ψ_j, we can repeat the proof with an ϵ-approximation argument and let $\epsilon\downarrow 0$.

For fixed $\eta_0>0$ and $\delta>0$, we have $\boldsymbol{K}_1(\gamma_{1,3}^{\eta_0,\delta},\mu_1)\leq\eta_0$ and $\boldsymbol{K}_2(\gamma_{2,3}^{\eta_0,\delta},\mu_2)\leq\delta$ by the definition of $\gamma^{\eta_0,\delta}$. Let $K_1=\{1,2,3\}$, $K_2=\{1,3,4,6\}$, and $K_3=\{5,6\}$, and it is easy to verify the collection $\{K_1,K_2,K_3\}$ is decomposable. As a result, by Proposition S.1 in the Online Supplement, we can construct random variables

$$(S,\tilde{S})\equiv(Y_1,Y_2,X,\tilde{Y}_1,\tilde{Y}_2,\tilde{X})\in\mathcal{S}\times\mathcal{S},$$

such that

$$\operatorname{Law}(Y_1,Y_2,X)=\gamma^{\eta_0,\delta},\quad \operatorname{Law}(\tilde{Y}_1,\tilde{X})=\mu_1,\quad \operatorname{Law}(\tilde{Y}_2,\tilde{X})=\mu_2,$$

and

$$\boldsymbol{K}_1(\gamma_{1,3}^{\eta_1,\delta},\mu_1)=\mathbb{E}[c_1(S_1,\tilde{S}_1)]\leq\eta_0,\quad \text{where } S_1=(Y_1,X) \text{ and } \tilde{S}_1=(\tilde{Y}_1,\tilde{X}).$$

Let ε be a Bernoulli random variable that is independent of $(S,\tilde{S})$ with $\mathbb{P}(\varepsilon=1)=\eta/\eta_0$. Define new random variables

$$\hat{S}\equiv(\hat{Y}_1,\hat{Y}_2,\hat{X})=\varepsilon(Y_1,Y_2,X)+(1-\varepsilon)(\tilde{Y}_1,\tilde{Y}_2,\tilde{X}),$$

and let $\hat{\gamma}=\operatorname{Law}(\hat{Y}_1,\hat{Y}_2,\hat{X})$. For any measurable set $A\in\mathcal{B}_{\mathcal{S}}$, we have

$$\begin{aligned}\hat{\gamma}(A)&=\mathbb{P}(\hat{S}\in A)=\mathbb{E}[\mathbb{P}(\hat{S}\in A|\varepsilon)]\\&=(\eta/\eta_0)\mathbb{P}(S\in A)+(1-\eta/\eta_0)\mathbb{P}(\tilde{S}\in A).\end{aligned}$$

This shows

$$\hat{\gamma}=(\eta/\eta_0)\gamma^{\eta_0,\delta}+(1-\eta/\eta_0)\tilde{\gamma},\quad \text{where } \tilde{\gamma}=\operatorname{Law}(\tilde{Y}_1,\tilde{Y}_2,\tilde{X}).$$

Next, we verify $\hat{\gamma}\in\Sigma(\eta,\delta)$. Because $\nu\longmapsto\boldsymbol{K}_1(\nu,\mu_1)$ is convex and $\tilde{\gamma}_{1,3}=\operatorname{Law}(\tilde{Y}_1,\tilde{X})=\mu_1$, we have

$$\boldsymbol{K}_1(\hat{\gamma}_{1,3},\mu_1)\ \leq\left(\frac{\eta}{\eta_0}\right)\boldsymbol{K}_1(\gamma_{1,3}^{\eta_1,\delta},\mu_1)+\left(1-\frac{\eta}{\eta_0}\right)\boldsymbol{K}_1(\tilde{\gamma}_{1,3},\mu_1)\leq\eta.$$

Similarly, we have $\boldsymbol{K}_2(\hat{\gamma}_{2,3},\mu_2)\leq\delta$. As a result, we verify $\hat{\gamma}\in\Sigma(\eta,\delta)$. Next, it is easy to see

$$\mathbb{E}[c_1((\hat{Y}_1,\hat{X}),(Y_1,X))]\leq\left(1-\frac{\eta}{\eta_0}\right)\mathbb{E}[c_1((\tilde{Y}_1,\tilde{X}),(Y_1,X))]\ \leq(\eta-\eta_0).$$

Because $\operatorname{Law}(Y_2,X)=\gamma_2^{\eta_0,\delta}$, $\operatorname{Law}(\tilde{Y}_2,\tilde{X})=\mu_2$, and $\boldsymbol{K}_2(\gamma_{2,3}^{\eta_0,\delta},\mu_2)\leq\delta$, that is, $\boldsymbol{W}_{p_2}(\gamma_{2,3}^{\eta_0,\delta},\mu_2)\leq\delta^{1/p_2}$, by the triangle inequality, we have

$$\boldsymbol{W}_{p_2}(\gamma_{2,3}^{\eta_1,\delta},\delta_{s_2})\leq\boldsymbol{W}_{p_2}(\gamma_{2,3}^{\eta_1,\delta},\mu_2)+\boldsymbol{W}_{p_2}(\mu_2,\delta_{s_2})\leq\delta^{1/p_2}+\boldsymbol{W}_{p_2}(\mu_2,\delta_{s_2}),$$

where δ_{s_2} denotes the Dirac measure at $\{s_2\}$ and $s_2\in\mathcal{S}_2$ is arbitrary. Further, Assumption 9(ii) implies $\rho_2(y_2',y_2)\leq 1+\boldsymbol{d}_{\mathcal{S}_2}(s_2',s_2)^{p_2}$ for all $s_2=(y_2,x)$ and $s_2'=(y_2',x')$,

$$\mathbb{E}[\rho_2(Y_2,y_2)]-1\ \leq\mathbb{E}[\boldsymbol{d}_{\mathcal{S}_2}(S_2,s_2)^{p_2}]=\boldsymbol{W}_{p_2}(\gamma_{2,3}^{\eta_1,\delta},\delta_{s_2})^{p_2}\leq[\delta^{1/p_2}+\boldsymbol{W}_{p_2}(\mu_2,\delta_{s_2})]^{p_2},$$

and

$$\mathbb{E}[\rho_2(\tilde{Y}_2,y_2)]-1\leq\mathbb{E}[\boldsymbol{d}_{\mathcal{S}_2}(\tilde{S}_2,s_2)^{p_2}]=\boldsymbol{W}_{p_2}(\mu_2,\delta_{s_2})^{p_2}.$$

As a result, by Assumption 9(iii),

$$\begin{aligned}\mathbb{E}[\rho_2(Y_2,\hat{Y}_2)]&=(\eta/\eta_0)\underbrace{\mathbb{E}[\rho_2(Y_2,Y_2)|\varepsilon=0]}_{=0}+(1-\eta/\eta_0)\mathbb{E}[\rho_2(Y_2,\tilde{Y}_2)|\varepsilon=1]\\&\leq(1-\eta/\eta_0)\mathbb{E}[\rho_2(Y_2,\tilde{Y}_2)]\ \leq(1-\eta/\eta_0)N(\mathbb{E}[\rho_2(Y_2,y_2)]+\mathbb{E}[\rho_2(y_2,\tilde{Y}_2)])\\&\leq M(1-\eta/\eta_0),\end{aligned}$$

where

$$M=N\boldsymbol{W}_{p_2}(\mu_2,\delta_{s_2})^{p_2}+N[\delta^{1/p_2}+\boldsymbol{W}_{p_2}(\mu_2,\delta_{s_2})]^{p_2}<\infty.$$

Therefore, by Assumption 10, we have

$$\begin{aligned}\mathcal{I}(\eta_0,\delta)-\mathcal{I}(\eta,\delta) &\le \mathbb{E}[f(Y_1,Y_2,X)]-\mathbb{E}[f(\hat{Y}_1,\hat{Y}_2,\hat{X})]\\ &\le \mathbb{E}[\Psi(c_1((Y_1,X),(\hat{Y}_1,\hat{X})),\rho_2(Y_2,\hat{Y}_2))]\\ &\le \Psi(\mathbb{E}[c_1(S_1,\hat{S}_1)],\mathbb{E}[\rho_2(Y_2,\hat{Y}_2)])\\ &\le \Psi(\eta_0-\eta,M(1-\eta/\eta_0)).\end{aligned}$$

The rest of the proof can be completed using the same reasoning. □

Now, we give the proof of Theorem 9.

Proof of Theorem 9. If $\eta_0>\eta\ge 0$, Lemma A.14 implies

$$\begin{aligned}0\le \mathcal{I}(\eta_0,\delta)-\mathcal{I}(\eta,0) &= \mathcal{I}(\eta_0,\delta)-\mathcal{I}(\eta,\delta)+\mathcal{I}(\eta,\delta)-\mathcal{I}(\eta,0)\\ &\le \Psi_1(\eta_0-\eta,M(1-\eta/\eta_0))+\Psi_2(M\delta,\delta).\end{aligned}$$

If $\eta\ge\eta_0$, by monotonicity of $\eta\mapsto\mathcal{I}(\eta,0)$ and Lemma A.14, we have

$$\mathcal{I}(\eta_0,\delta)-\mathcal{I}(\eta,0)\le\mathcal{I}(\eta_0,\delta)-\mathcal{I}(\eta_0,0)\le\Psi_2(M\delta,\delta),$$

and

$$\mathcal{I}(\eta_0,\delta)-\mathcal{I}(\eta,0)\ge\mathcal{I}(\eta,\delta)-\mathcal{I}(\eta,0)\ge 0.$$

As a result, we must have for all η_0,η and δ in $[0,\infty)$

$$0\le\mathcal{I}(\eta_0,\delta)-\mathcal{I}(\eta,0)\le\Psi_1(|\eta_0-\eta|,M|1-\eta/\eta_0|)+\Psi_2(M\delta,\delta).$$

The continuity of $\mathcal{I}$ at $(\eta,0)$ follows from the continuity of Ψ_1 and Ψ_2, and letting $(\eta_0,\delta)\to(\eta,0)$. Using a similar argument, we can show $\mathcal{I}$ is continuous at $(0,\eta)$. □

A.4. Proofs in Section 6

A.4.1. Proof of Proposition 2.

By some simple algebra and Theorem 2, we have

$$\begin{aligned}\mathcal{I}_{\mathrm{D}}(\delta) &= \inf_{\lambda\in\mathbb{R}_+^2}\left\{\langle\lambda,\delta\rangle+\sup_{\gamma\in\Pi(\mu_1,\mu_2)}\int_{\mathcal{S}}[(f_1)_{\lambda_1}(y_1)+(f_2)_{\lambda_2}(y_2)]\,d\gamma(y_1,y_2)\right\}\\ &= \inf_{\lambda_1\ge0}\left[\lambda_1\delta_1+\int_{\mathcal{Y}_1}(f_1)_{\lambda_1}\,d\mu_1\right]+\inf_{\lambda_2\ge0}\left[\lambda_2\delta_2+\int_{\mathcal{Y}_2}(f_2)_{\lambda_2}\,d\mu_2\right],\end{aligned}$$

where the last step holds because $(f_\ell)_\lambda\ge f_\ell$ and the right-hand side is well-defined because $f_\ell\in L^1(\mu_\ell)$. Next, we show $\mathcal{I}(\delta)=\mathcal{I}_{\mathrm{D}}(\delta)$. Theorem 3 implies

$$\mathcal{I}(\delta)=\inf_{\lambda\in\mathbb{R}_+^2}\left\{\langle\lambda,\delta\rangle+\sup_{\pi\in\Pi(\mu_{13},\mu_{23})}\int_{\mathcal{S}_1\times\mathcal{S}_2}(f_{\mathcal{S}})_\lambda\,d\pi\right\},$$

where $(f_{\mathcal{S}})_\lambda:\mathcal{S}_1\times\mathcal{S}_2\to\mathbb{R}$ is given by

$$(f_{\mathcal{S}})_\lambda(s_1,s_2)=\sup_{(y_1',y_2',x')\in\mathcal{S}}\left\{f_1(y_1')+f_2(y_2')-\sum_{1\le\ell\le2}\lambda_\ell c_\ell((y_\ell,x_\ell),(y_\ell',x'))\right\}.$$

In fact, Assumption 12 implies that for all $s_\ell=(y_\ell,x_\ell)\in\mathcal{S}_\ell$ and $s_\ell'=(y_\ell',x_\ell')\in\mathcal{S}_\ell$, one has

$$c_{Y_\ell}(y_\ell,y_\ell')=\inf_{x_\ell,x_\ell'\in\mathcal{X}}c_\ell((y_\ell,x_\ell),(y_\ell',x_\ell'))\le c_\ell((y_\ell,x_\ell),(y_\ell',x_\ell')).$$

Recall $(f_{\mathcal{S}})_\lambda:(s_1,s_2)\mapsto(f_{\mathcal{S}})_\lambda(s_1,s_2)$ is a function from $\mathcal{S}_1\times\mathcal{S}_2\to\mathbb{R}$ with $s_\ell=(y_\ell,x_\ell)\in\mathcal{S}_\ell$. As a result, for all $s_1\in\mathcal{S}_1$ and $s_2\in\mathcal{S}_2$

$$\begin{aligned}(f_{\mathcal{S}})_\lambda(s_1,s_2) &\le \sup_{(y_1',y_2',x')\in\mathcal{S}}\left\{f_1(y_1')+f_2(y_2')-\sum_{1\le\ell\le2}\lambda_\ell c_{Y_\ell}(y_\ell,y_\ell')\right\}\\ &= (f_1)_{\lambda_1}(y_1)+(f_2)_{\lambda_2}(y_2).\end{aligned}$$

This shows that for all $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2_+$, one has

$$\sup_{\pi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{S}_1 \times \mathcal{S}_2} (f_S)_\lambda \, d\pi \le \sup_{\gamma \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{Y}_1 \times \mathcal{Y}_2} [(f_1)_{\lambda_1}(y_1) + (f_2)_{\lambda_2}(y_2)] \, d\gamma(y_1, y_2),$$

and hence $\mathcal{I}(\delta) \le \mathcal{I}_{\mathrm{D}}(\delta)$. We end the proof by showing

$$\sup_{\pi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{S}_1 \times \mathcal{S}_2} (f_S)_\lambda d\pi \ge \int_{\mathcal{Y}_1} (f_1)_{\lambda_1} d\mu_1 + \int_{\mathcal{Y}_2} (f_2)_{\lambda_2} d\mu_2.$$

It suffices to show that there is $\pi \in \Pi(\mu_{13}, \mu_{23})$ such that $(f_S)_\lambda(s_1, s_2) \ge (f_1)_{\lambda_1}(y_1) + (f_2)_{\lambda_2}(y_2)$, π-a.e. In fact, we note that if $x_1 = x_2$, then $(f_S)_\lambda((y_1, x_1), (y_2, x_2)) = (f_1)_{\lambda_1}(y_1) + (f_2)_{\lambda_2}(y_2)$ under Assumption 12. Consider a probability measure $\pi^\star = \mathrm{Law}(Y_1, X, Y_2, X)$, where $\mu_{\ell,3} = \mathrm{Law}(Y_\ell, X)$ for $\ell = 1, 2$. As a result,

$$\begin{aligned}\sup_{\pi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathcal{S}_1 \times \mathcal{S}_1} (f_S)_\lambda \, d\pi &\ge \int_{\mathcal{S}_1 \times \mathcal{S}_2} (f_S)_\lambda \, d\pi^\star = \int_{\mathcal{S}_1 \times \mathcal{S}_2} [(f_1)_{\lambda_1} + (f_2)_{\lambda_2}] \, d\pi^\star \\ &= \int_{\mathcal{Y}_1} (f_1)_{\lambda_1} \, d\mu_1 + \int_{\mathcal{Y}_2} (f_2)_{\lambda_2} \, d\mu_2. \quad \square\end{aligned}$$

A.4.2. Proof of Proposition 3. Because $c_{Y_\ell}(y_\ell, y'_\ell) = \inf_{x_\ell, x'_\ell \in \mathcal{X}_\ell} c_\ell(s_\ell, s'_\ell)$, the proof of Proposition 2 implies $\mathcal{I}(\delta) \le \mathcal{I}_{\mathrm{D}}(\delta)$. $\square$

A.4.3. Proof of Proposition 4(i). The proof consists of two steps. In Step 1, we derive the dual form of $\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ for $\delta \in \mathbb{R}^2_{++}$. In Step 2, we derive the dual reformulations of $\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ for $\delta \in \mathbb{R}^2_+ \setminus \mathbb{R}^2_{++}$.

Step 1. We derive the expressions of $\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ for $\delta \in \mathbb{R}^2_{++}$. First, recall $c_{Y_\ell}(y_\ell, y'_\ell) = V^{-1}_{\ell,YY}(y_\ell - y'_\ell)^2$. Theorem 2 implies

$$\mathcal{I}_{\mathrm{D}}(\delta) = \inf_{\lambda \in \mathbb{R}^2_+} \left[\langle \lambda, \delta \rangle + \sup_{\varpi \in \Pi(\mu_{Y_1}, \mu_{Y_2})} \int_{\mathbb{R}^2} (f_Y)_\lambda(y_1, y_2) \, d\varpi(y_1, y_2) \right],$$

where $(f_Y)_\lambda : (y_1, y_2) \longmapsto (f_Y)_\lambda(y_1, y_2)$ from $\mathbb{R}^2$ to $\mathbb{R}$ is given by

$$(f_Y)_\lambda(y_1, y_2) = y_2 - y_1 + \frac{V_{1,YY}}{4\lambda_1} + \frac{V_{2,YY}}{4\lambda_2}.$$

Because $V_{\ell,YY} > 0$ for $\ell \in [2]$, by some simple algebra, we have for all $\delta \in \mathbb{R}^2_{++}$

$$\mathcal{I}_{\mathrm{D}}(\delta) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + V^{1/2}_{1,YY} \delta_1^{1/2} + V^{1/2}_{2,YY} \delta_2^{1/2}.$$

Next, we derive the expression of $\mathcal{I}(\delta)$ for $\delta \in \mathbb{R}^2_{++}$. Let $Q_\ell \in \mathbb{R}^{(d+1)\times(d+1)}$ be the inverse of V_ℓ, that is,

$$Q_\ell = \begin{bmatrix} Q_{\ell,YY} & Q_{\ell,YX} \\ Q_{\ell,XY} & Q_{\ell,XX} \end{bmatrix} = \begin{bmatrix} (V_\ell/V_{\ell,XX})^{-1} & -(V_\ell/V_{\ell,XX})^{-1} V_{\ell,YX} V^{-1}_{\ell,XX} \\ -V^{-1}_{\ell,XX} V_{\ell,XY} (V_\ell/V_{\ell,XX})^{-1} & (V_\ell/V_{\ell,YY})^{-1} \end{bmatrix},$$

where $V_\ell/V_{\ell,XX} = V_{\ell,YY} - V_{\ell,YX} V^{-1}_{\ell,XX} V_{\ell,XY}$ and $V_\ell/V_{\ell,YY} = V_{\ell,XX} - V_{\ell,XY} V^{-1}_{\ell,YY} V_{\ell,YX}$. Conversely,

$$\begin{bmatrix} V_{\ell,YY} & V_{\ell,YX} \\ V_{\ell,XY} & V_{\ell,XX} \end{bmatrix} = \begin{bmatrix} (Q_\ell/Q_{\ell,XX})^{-1} & -Q^{-1}_{\ell,YY} Q_{\ell,YX} (Q_\ell/Q_{\ell,YY})^{-1} \\ -(Q_\ell/Q_{\ell,YY})^{-1} Q_{\ell,XY} Q^{-1}_{\ell,YY} & (Q_\ell/Q_{\ell,YY})^{-1} \end{bmatrix},$$

where $Q_\ell/Q_{\ell,XX} = Q_{\ell,YY} - Q_{\ell,YX} Q^{-1}_{\ell,XX} Q_{\ell,XY}$ and $Q_\ell/Q_{\ell,YY} = Q_{\ell,XX} - Q_{\ell,XY} Q^{-1}_{\ell,YY} Q_{\ell,YX}$. Next, we evaluate the function $(f_S)_\lambda(s_1, s_2)$ that appears in the dual reformulation. For simplicity, we write $a_1 = -1$ and $a_2 = 1$. Consider the following derivation:

$$\begin{aligned}(f_S)_\lambda(s_1, s_2) &:= \sup_{y'_1, y'_2, x'} \left\{ y'_2 - y'_1 - \sum_{\ell=1,2} \lambda_\ell c_\ell((y'_\ell, x'), (y_\ell, x_\ell)) \right\} \\ &= \sup_{y'_1, y'_2, x'} \left\{ \sum_{1 \le \ell \le 2} \left(a_\ell y_\ell - \lambda_\ell \begin{bmatrix} y'_\ell - y_\ell \\ x' - x_\ell \end{bmatrix}^\top Q_\ell \begin{bmatrix} y'_\ell - y_\ell \\ x' - x_\ell \end{bmatrix} \right) \right\} \\ &=_{(1)} y_2 - y_1 + \sup_{z'_1, z'_2, x'} \left\{ \sum_{1 \le \ell \le 2} \left(a_\ell z'_\ell - \lambda_\ell \begin{bmatrix} z'_\ell \\ x' - x_\ell \end{bmatrix}^\top Q_\ell \begin{bmatrix} z'_\ell \\ x' - x_\ell \end{bmatrix} \right) \right\} \\ &= y_2 - y_1 + \sup_{x' \in \mathbb{R}^d} \left\{ \sum_{1 \le \ell \le 2} \sup_{z'_\ell \in \mathbb{R}} \left(a_\ell z'_\ell - \lambda_\ell \begin{bmatrix} z'_\ell \\ x' - x_\ell \end{bmatrix}^\top Q_\ell \begin{bmatrix} z'_\ell \\ x' - x_\ell \end{bmatrix} \right) \right\},\end{aligned}$$

where Equation (1) follows from the change of variables $z'_\ell = y'_\ell - y_\ell$. So, to evaluate $(f_S)_\lambda(s_1, s_2)$, it suffices to maximize $(z'_1, z'_2, x') \mapsto \phi_1(z'_1, x'; x_1) + \phi_2(z'_2, x'; x_2)$, where

$$\phi_\ell(z'_\ell, x'; x_\ell) = a_\ell z'_\ell - \lambda_\ell \begin{bmatrix} z'_\ell \\ x' - x_\ell \end{bmatrix}^\top Q_\ell \begin{bmatrix} z'_\ell \\ x' - x_\ell \end{bmatrix}.$$

We first consider $\sup_{z'_\ell \in \mathbb{R}} \phi_\ell(z'_\ell, x'; x_\ell)$. The first-order conditions imply that the optimal solution is

$$z'_\ell = (\lambda_\ell Q_{\ell,YY})^{-1} \left[\frac{a_\ell}{2} - \lambda_\ell Q_{\ell,YX}(x' - x_\ell) \right].$$

By some simple algebra, $\sup_{z'_\ell \in \mathbb{R}} \phi_\ell(z'_\ell, x', x_\ell) = \varphi_\ell(x' - x_\ell, \lambda)$, where $\varphi_\ell : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is given by

$$\varphi_\ell(x, \lambda_\ell) = \frac{Q^{-1}_{\ell,YY}}{4\lambda_\ell} + a_\ell x^\top V^{-1}_{\ell,XX} V_{\ell,XY} - \lambda_\ell x^\top V^{-1}_{\ell,XX} x.$$

As a result,

$$(f_S)_\lambda(s_1, s_2) = \sup_{x' \in \mathbb{R}^d} [\varphi_1(x' - x_1, \lambda_1) + \varphi_2(x' - x_2, \lambda_2)].$$

Now, we consider the optimization above. The first-order conditions imply the optimal solution x' takes the form of $x' - x_\ell = B_\ell(x_2 - x_1) + b_\ell$ for some $B_\ell \in \mathbb{R}^{d \times d}$ and $b_\ell \in \mathbb{R}^d$ that depend on λ_ℓ. So, we have

$$\sup_{x' \in \mathbb{R}^d} [\varphi_1(x', x_1) + \varphi_2(x', x_2)] = b + B(x_1 - x_2) - (x_1 - x_2)^\top W (x_1 - x_2)$$

for some positive definite matrix $W \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}$ that depend on $\lambda_1, \lambda_2, x_1$ and x_2. Here, the constant b will be determined below. For any $\pi \in \Pi(\mu_{13}, \mu_{23})$, we have

$$\begin{aligned} \int_{\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}} (f_S)_\lambda \, d\pi &= \frac{1}{4\lambda_1} Q^{-1}_{1,YY} + \frac{1}{4\lambda_2} Q^{-1}_{2,YY} + \underbrace{\int_{\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}} B(x_1 - x_2) \, d\pi}_{=0} + \int_{\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}} (x_1 - x_2)^\top W (x_1 - x_2) \, d\pi(s_1, s_2) + b \\ &= \frac{1}{4\lambda_1} Q^{-1}_{1,YY} + \frac{1}{4\lambda_2} Q^{-1}_{2,YY} - \int (x_1 - x_2)^\top W (x_1 - x_2) \, d\pi + b. \end{aligned}$$

Now, let us consider $\sup_{\pi \in \Pi(\mu_{13}, \mu_{23})} \int (f_S)_\lambda \, d\pi$. To maximize $\int (f_S)_\lambda \, d\pi$, it suffices to consider

$$\inf_{\pi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathbb{R}^{d+1} \times \mathbb{R}^{d+1}} (x_1 - x_2)^\top W (x_1 - x_2) \, d\pi(s_1, s_2).$$

Because $(x_1 - x_2)^\top W (x_1 - x_2)$ for all $x_1, x_2 \in \mathbb{R}^d$, the probability measure $\pi = \mathrm{Law}(Y_1, X, Y_2, X)$ with $\mathrm{Law}(Y_\ell, X) = \mu_{\ell,3}$ for $\ell = 1, 2$ is a solution and the optimal value is zero. We denote by Π the set of all probability measures on $\mathcal{S}_1 \times \mathcal{S}_2$ that takes forms of $\pi = \mathrm{Law}(Y_1, X, Y_2, X)$. As a consequence,

$$\sup_{\pi \in \Pi(\mu_{13}, \mu_{23})} \int_{\mathbb{R}^{2d+2}} (f_S)_\lambda \, d\pi = \frac{1}{4\lambda_1} Q^{-1}_{1,YY} + \frac{1}{4\lambda_2} Q^{-1}_{2,YY} + b,$$

where $b = \frac{1}{4} V_o^\top (\lambda_1 V^{-1}_{1,XX} + \lambda_2 V^{-1}_{2,XX})^{-1} V_o$ with $V_o = V^{-1}_{2,XX} V_{2,XY} - V^{-1}_{1,XX} V_{1,XY}$. As a result, the dual reformulation of $\mathcal{I}_{\mathrm{D}}(\delta)$ is given by

$$\mathcal{I}(\delta) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + \inf_{\lambda \in \mathbb{R}^2_+} \left\{ \lambda_1 \delta_1 + \lambda_2 \delta_2 + \frac{1}{4\lambda_1}(V_1 / V_{1,XX}) + \frac{1}{4\lambda_2}(V_2 / V_{2,XX}) + \frac{1}{4} V_o^\top (\lambda_1 V^{-1}_{1,XX} + \lambda_2 V^{-1}_{2,XX})^{-1} V_o \right\}.$$

Step 2. We derive the dual reformulation of $\mathcal{I}_{\mathrm{D}}(\delta)$ and $\mathcal{I}(\delta)$ for $\delta \in \mathbb{R}^2_+ \setminus \mathbb{R}^2_{++}$. First, we note that $\mathcal{I}_{\mathrm{D}}(0) = \mathcal{I}(0) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1]$. Theorem 2 implies that

$$\mathcal{I}_{\mathrm{D}}(\delta_1, 0) = \inf_{\lambda \in \mathbb{R}^2_+} \left[\lambda_1 \delta_1 + \sup_{\varpi \in \Pi(\mu_{Y_1}, \mu_{Y_2})} \int_{\mathbb{R}^2} (f_Y)_{\lambda,1}(y_1, y_2) \, d\varpi(y_1, y_2) \right],$$

$$\mathcal{I}_{\mathrm{D}}(0, \delta_2) = \inf_{\lambda_2 \in \mathbb{R}^2_+} \left[\lambda_2 \delta_2 + \sup_{\varpi \in \Pi(\mu_{Y_1}, \mu_{Y_2})} \int_{\mathbb{R}^2} (f_Y)_{\lambda,2}(y_1, y_2) \, d\varpi(y_1, y_2) \right],$$

where $(f_Y)_{\lambda,\ell}$, for $\ell = 1, 2$, is given by $(f_Y)_{\lambda,\ell} = y_2 - y_1 + (4\lambda_\ell)^{-1} V_{\ell,YY}$. Because $V_{\ell,YY} > 0$, by simple algebra, we have for all $\delta \in \mathbb{R}^2_{++}$

$$\mathcal{I}_{\mathrm{D}}(\delta_1, 0) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + V^{1/2}_{1,YY} \delta_1^2 \quad \text{and} \quad \mathcal{I}_{\mathrm{D}}(0, \delta_2) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + V^{1/2}_{2,YY} \delta_2^2.$$

Theorem 3 implies that

$$\mathcal{I}(\delta_1,0)=\inf_{\lambda\in\mathbb{R}_+^2}\left[\langle\lambda,\delta\rangle+\sup_{\varpi\in\Pi(\mu_{13},\mu_{23})}\int_{\mathbb{R}^2}(f_S)_{\lambda,1}(y_1,y_2)\,d\varpi(y_1,y_2)\right],$$

$$\mathcal{I}(0,\delta_2)=\inf_{\lambda_1\in\mathbb{R}_+^2}\left[\langle\lambda,\delta\rangle+\sup_{\varpi\in\Pi(\mu_{13},\mu_{23})}\int_{\mathbb{R}^2}(f_S)_{\lambda,2}(y_1,y_2)\,d\varpi(y_1,y_2)\right],$$

where $(f_Y)_{\lambda,\ell}$, for $\ell=1,2$, is given by

$$(f_Y)_{\lambda,1}=\sup_{y_1'}\left\{y_2-y_1'-\lambda_1\begin{bmatrix}y_1'-y_1\\x_2-x_1\end{bmatrix}^\top Q_1\begin{bmatrix}y_1'-y_1\\x_2-x_1\end{bmatrix}\right\},$$

$$(f_Y)_{\lambda,2}=\sup_{y_2'}\left\{y_2'-y_1-\lambda_2\begin{bmatrix}y_2'-y_2\\x_1-x_2\end{bmatrix}^\top Q_2\begin{bmatrix}y_2'-y_2\\x_1-x_2\end{bmatrix}\right\}.$$

With similar calculations as in Step 1, the functions $(f_Y)_{\lambda,1}$ and $(f_Y)_{\lambda,2}$ can be written as

$$(f_Y)_{\lambda,1}=y_2-y_1+\frac{V_1/V_{1,XX}}{4\lambda_1}-(x_2-x_1)^\top V_{1,XX}^{-1}V_{1,XY}-\lambda_1(x_2-x_1)^\top V_{1,XX}^{-1}(x_2-x_1),$$

$$(f_Y)_{\lambda,2}=y_2-y_1+\frac{V_2/V_{2,XX}}{4\lambda_2}+(x_1-x_2)^\top V_{2XX}^{-1}V_{2,XY}-\lambda_2(x_1-x_2)^\top V_{2,XX}^{-1}(x_1-x_2).$$

With the same reasoning as in Step 1, we have

$$\sup_{\varpi\in\Pi(\mu_{13},\mu_{23})}\int(f_S)_{\lambda,\ell}\,d\varpi=\mathbb{E}[Y_2]-\mathbb{E}[Y_1]+\frac{V_\ell/V_{\ell,XX}}{4\lambda_\ell},\quad\text{for }\ell\in[2].$$

Therefore,

$$\mathcal{I}(\delta_1,0)=\mathbb{E}[Y_2]-\mathbb{E}[Y_1]+(V_1/V_{1,XX})^{1/2}\delta_1^{1/2}=\mathcal{I}_{\mathrm{D}}(\delta_1,0),$$
$$\mathcal{I}(0,\delta_2)=\mathbb{E}[Y_2]-\mathbb{E}[Y_1]+(V_2/V_{2,XX})^{1/2}\delta_2^{1/2}=\mathcal{I}_{\mathrm{D}}(0,\delta_2).\quad\square$$

A.4.4. Proof of Proposition 4(ii). Recalling the proof of Proposition 4(i), we have

$$\mathcal{I}(\delta)=\inf_{\lambda\in\mathbb{R}_+^2}\left\{\langle\lambda,\delta\rangle+\sup_{\pi\in\tilde{\Pi}}\int_{\mathbb{R}^{2d+2}}(f_S)_\lambda\,d\pi\right\},$$

where $\tilde{\Pi}$ is the set of all probability measures such that their supports $\mathrm{Supp}(\pi)$ are in $\{(y_1,x_1,y_2,x_2)\in\mathbb{R}^{2d+2}:x_1=x_2\}$. By the definition of $\tilde{\Pi}$, to evaluate $\mathcal{I}(\delta)$, it suffices to restrict the domain of $(f_S)_\lambda$ on $\mathrm{Supp}(\pi)$. For any $(s_1,s_2)\in\mathrm{Supp}(\pi)$, we have $x_1=x_2$

$$\begin{aligned}(f_S)_\lambda(s_1,s_2)&=(y_2-y_1)+\sup_{x'\in\mathbb{R}^d}[\varphi_1(x'-x_1,\lambda_1)+\varphi_2(x'-x_2,\lambda_2)]\\&=(y_2-y_1)+\underbrace{\sup_{x'\in\mathbb{R}^d}\left\{\sum_{1\le\ell\le2}\frac{Q_{\ell,YY}^{-1}}{4\lambda_\ell}+x'^\top V_{\ell,XX}^{-1}V_{\ell,XY}a_\ell-\lambda_\ell x'^\top V_{\ell,XX}^{-1}x'\right\}}_{=\mathcal{H}(\lambda,\delta)}.\end{aligned}$$

As a consequence, $(f_S)_\lambda(s_1,s_2)$ is independent of x_1 and x_2 for all $(s_1,s_2)\in\mathrm{Supp}(\pi)$, and hence for all $\pi\in\tilde{\Pi}$, we have

$$\int_{\mathbb{R}^{2d+2}}(f_S)_\lambda d\pi=\mathbb{E}[Y_2]-\mathbb{E}[Y_1]+\mathrm{R}(\lambda,\delta),$$

where $\mathrm{R}(\lambda,\delta)=\mathcal{H}(\lambda,\delta)+\langle\lambda,\delta\rangle$ and $\mathrm{Law}(Y_\ell,X)=\mu_{\ell3}$ for $\ell=1,2$. So, $\mathcal{I}(\delta)=\mathbb{E}[Y_2]-\mathbb{E}[Y_1]+\inf_{\lambda\in\mathbb{R}_+^2}\mathrm{R}(\lambda,\delta)$. Moreover,

$$\mathcal{I}_{\mathrm{D}}(\delta)=\mathbb{E}[Y_2]-\mathbb{E}[Y_1]+\inf_{\lambda\in\mathbb{R}_+^2}\mathrm{R}_{\mathrm{D}}(\lambda,\delta),$$

where

$$\mathrm{R}_{\mathrm{D}}(\lambda,\delta)=\langle\lambda,\delta\rangle+\frac{V_{1,YY}}{4\lambda_1}+\frac{V_{2,YY}}{4\lambda_2}.$$

The rest of the proof is divided into the following two steps.

Step 1. We show that $\mathcal{I}_{\mathrm{D}}(\delta) = \mathcal{I}(\delta)$ implies that the following holds:

$$\delta_1^{1/2} V_{1,YY}^{-1/2} V_{1,XY} + \delta_2^{1/2} V_{2,YY}^{-1/2} V_{2,XY} = 0. \tag{A.4}$$

Because $Q_{\ell,YY} \geq V_{\ell,YY}^{-1}$ by definition, then $Q_{\ell,YY}^{-1} \leq V_{\ell,YY}$ and $\mathrm{R}(\lambda,\delta) \leq \mathrm{R}_{\mathrm{D}}(\lambda,\delta)$. Let $\lambda_{\mathrm{D}}^\star = (\delta_1^{-1/2} V_{1,YY}^{1/2}, \delta_2^{-1/2} V_{2,YY}^{1/2})$. It is easy to see $\inf_{\lambda\in\mathbb{R}_+^2} \mathrm{R}_{\mathrm{D}}(\lambda,\delta) = \mathrm{R}_{\mathrm{D}}(\lambda_{\mathrm{D}}^\star,\delta) \geq \mathrm{R}(\lambda_{\mathrm{D}}^\star,\delta)$ and hence

$$\mathcal{I}(\delta) \leq \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + \mathrm{R}(\lambda_{\mathrm{D}}^\star,\delta) \leq \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + \mathrm{R}_{\mathrm{D}}(\lambda_{\mathrm{D}}^\star,\delta) = \mathcal{I}_{\mathrm{D}}(\delta).$$

Thus, $\mathcal{I}(\delta) = \mathcal{I}_{\mathrm{D}}(\delta)$ implies $\mathrm{R}_{\mathrm{D}}(\lambda_{\mathrm{D}}^\star,\delta) = \mathrm{R}(\lambda_{\mathrm{D}}^\star,\delta)$. In fact, we note that

$$\mathrm{R}_{\mathrm{D}}(\lambda,\delta) = \langle\lambda,\delta\rangle + \sup_{x'\in\mathbb{R}^d}\left[\sum_{1\leq\ell\leq 2}\varphi_\ell(x',\lambda_\ell)\right] \quad\text{and}\quad \mathrm{R}_{\mathrm{D}}(\lambda,\delta) = \langle\lambda,\delta\rangle + \sum_{1\leq\ell\leq 2}\sup_{x'\in\mathbb{R}^d}\varphi_\ell(x',\lambda_\ell).$$

Because $x' \longmapsto \varphi_\ell(x',\lambda_\ell)$ is strictly concave, it admits a unique maximizer and hence $\mathrm{R}_{\mathrm{D}}(\lambda_{\mathrm{D}}^\star,\delta) = \mathrm{R}(\lambda_{\mathrm{D}}^\star,\delta)$ implies for $\ell = 1,2$

$$\operatorname*{argmax}_{x'\in\mathbb{R}^d}\left[\sum_{1\leq\ell\leq 2}\varphi_\ell(x',\lambda_{\mathrm{D},\ell}^\star)\right] = \operatorname*{argmax}_{x'\in\mathbb{R}^d}\varphi_\ell(x',\lambda_{\mathrm{D},\ell}^\star).$$

The first-order conditions imply

$$\operatorname*{argmax}_{x'\in\mathbb{R}^d}\left[\sum_{1\leq\ell\leq 2}\varphi_\ell(x',\lambda_\ell)\right] = \left(\sum_{1\leq\ell\leq 2}\lambda_\ell V_{\ell,XX}^{-1}\right)^{-1}\left(\sum_{1\leq\ell\leq 2}a_\ell V_{\ell,XX}^{-1}V_{\ell,XY}\right),$$

and

$$\operatorname*{argmax}_{x'\in\mathbb{R}^d}\varphi_\ell(x',\lambda_\ell) = \frac{1}{2}a_\ell {\lambda_{\mathrm{D},\ell}^\star}^{-1} V_{2,XY}, \qquad \text{for } \ell = 1,2.$$

So, recalling $\lambda_{\mathrm{D},\ell}^\star = \delta_\ell^{-1/2} V_{\ell,YY}^{1/2}$, $a_1 = -1$, and $a_2 = 1$, we have

$$\delta_1^{1/2} V_{1,YY}^{-1/2} V_{1,XY} + \delta_2^{1/2} V_{2,YY}^{-1/2} V_{2,XY} = 0.$$

Step 2. We show $\delta_1^{1/2} V_{1,YY}^{-1/2} V_{1,XY} + \delta_2^{1/2} V_{2,YY}^{-1/2} V_{2,XY} = 0$ implies $\mathcal{I}_{\mathrm{D}}(\delta) = \mathcal{I}(\delta)$. We note $\lambda \longmapsto \mathrm{R}_{\mathrm{D}}(\lambda,\delta)$ is convex because it is supremum of a set of affine functions. It can be written as

$$\mathrm{R}_{\mathrm{D}}(\lambda,\delta) = \langle\lambda,\delta\rangle + \sum_{1\leq\ell\leq 2}\frac{V_\ell/V_{\ell,XX}}{4\lambda_\ell} + \frac{1}{4}V_o^\top \underbrace{(\lambda_1 V_{1,XX}^{-1} + \lambda_2 V_{2,XX}^{-1})}_{=\Lambda_\lambda}{}^{-1} V_o.$$

Taking derivatives with respect to λ_ℓ yields

$$\frac{\partial \mathrm{R}_{\mathrm{D}}(\lambda,\delta)}{\partial\lambda_\ell} = \delta_\ell - \frac{V_\ell/V_{\ell,XX}}{4\lambda_\ell^2} - \frac{1}{4}V_o^\top \Lambda_\lambda^{-1} V_{\ell,XX}^{-1} \Lambda_\lambda^{-1} V_o.$$

By some algebra and under $\delta_1^{1/2} V_{1,YY}^{-1/2} V_{1,XY} + \delta_2^{1/2} V_{2,YY}^{-1/2} V_{2,XY} = 0$, we can show

$$\frac{\partial \mathrm{R}_{\mathrm{D}}(\lambda_{\mathrm{D}}^\star,\delta)}{\partial\lambda_\ell} = 0.$$

As a result, $\mathrm{R}_{\mathrm{D}}(\lambda_{\mathrm{D}}^\star,\delta) = \inf_{\lambda\in\mathbb{R}_+^2}\mathrm{R}_{\mathrm{D}}(\lambda,\delta) = \mathrm{R}(\lambda_{\mathrm{D}}^\star,\delta) = \inf_{\lambda\in\mathbb{R}_+^2}\mathrm{R}(\lambda,\delta)$ and

$$\mathcal{I}(\delta) = \mathbb{E}[Y_2] - \mathbb{E}[Y_1] + \inf_{\lambda\in\mathbb{R}_+^2}\mathrm{R}_{\mathrm{D}}(\lambda,\delta) = \mathcal{I}_{\mathrm{D}}(\delta).$$

Step 3. We show that Equation (A.4) incorporates the case when $\delta_1 = 0$ or $\delta_2 = 0$. From Proposition 4(ii), we know the following statements hold:

- When $\delta_1 > 0$ and $\delta_2 = 0$, $\mathcal{I}_{\mathrm{D}}(\delta) = \mathcal{I}(\delta)$ if and only if $V_{1,XY} = 0$.
- When $\delta_1 = 0$ and $\delta_2 > 0$, $\mathcal{I}_{\mathrm{D}}(\delta) = \mathcal{I}(\delta)$ if and only if $V_{2,XY} = 0$.
- When $\delta_1 = \delta_2 = 0$, $\mathcal{I}_{\mathrm{D}}(\delta) = \mathcal{I}(\delta) = \mathcal{I}_{\mathrm{D},0}$.

We see that Equation (A.4) incorporates all these cases:

- When $\delta_1 > 0$ and $\delta_2 = 0$, Equation (A.4) is equivalent to $V_{1,XY} = 0$.
- When $\delta_1 = 0$ and $\delta_2 > 0$, Equation (A.4) is equivalent to $V_{2,XY} = 0$.
- When $\delta_1 = \delta_2 = 0$, Equation (A.4) is satisfied always.

This completes the proof. □

A.4.5. Proof of Proposition 4(iii). The continuity of $\mathcal{I}_{\mathrm{D}}$ can be seen from Proposition 4(i) or Theorem 8. Next, we show $\mathcal{I}$ is continuous on $\mathbb{R}_+^2$ by verifying the conditions of Theorem 9. Obviously, $\boldsymbol{d}_{\mathcal{S}_\ell}(s_\ell,s_\ell') = \sqrt{c_\ell(s_\ell,s_\ell')}$ defines a norm on

$\mathcal{S}_\ell = \mathbb{R}^{q+1}$. Define a function $\rho_\ell : \mathcal{Y}_\ell \times \mathcal{Y}_\ell \to \mathbb{R}_+$ as

$$\rho_\ell(y_\ell, y'_\ell) = (y_\ell - y'_\ell)^\top V_{\ell,YY}^{-1}(y_\ell - y'_\ell).$$

In fact, it is not difficult to see

$$\rho_\ell(y_\ell, y'_\ell) = \min_{(x_\ell, x'_\ell)\in\mathcal{X}_\ell\times\mathcal{X}_\ell} (s_\ell - s'_\ell)^\top V_\ell^{-1}(s_\ell - s'_\ell) \le c_\ell(s_\ell, s'_\ell), \quad \forall s_\ell, s'_\ell \in \mathcal{S}_\ell.$$

Moreover, $\rho_\ell^{1/2}$ is a norm on $\mathcal{Y}_\ell$ and the triangle inequality implies

$$\rho_\ell^{1/2}(y_\ell, y'_\ell) \le \rho_\ell^{1/2}(y_\ell, y^\star_\ell) + \rho_\ell^{1/2}(y^\star_\ell, y'_\ell), \quad \forall y_\ell, y'_\ell, y^\star_\ell \in \mathcal{Y}_\ell.$$

As a result, we must have

$$\rho(y_\ell, y'_\ell) \le 2[\rho_\ell(y_\ell, y^\star_\ell) + \rho_\ell(y^\star_\ell, y'_\ell)], \quad \forall y_\ell, y'_\ell, y^\star_\ell \in \mathcal{Y}_\ell.$$

We verified that the functions ρ_1 and ρ_2 satisfy Assumption 9 with respect to Mahalanobis distances. Recall $f(y_1, y_2, x) = y_1 - y_2$ and define a concave function $\Psi : \mathbb{R}^2 \to \mathbb{R}_+$ as

$$\Psi : (a_1, a_2) \longmapsto V_{1,YY}^{1/2} a^{1/2} + V_{2,YY}^{1/2} a_2^{1/2}.$$

Because $|y_\ell - y'_\ell|^2 = V_{\ell,YY}\rho_\ell(y_\ell, y'_\ell)$ and $\rho_\ell \le c_\ell$, then

$$\begin{aligned} f(y_1, y_2, x) - f(y'_1, y'_2, x') &\le |y_1 - y'_1| + |y_2 - y'_2| \\ &\le \sum_{\ell=1}^{2} V_{\ell,YY}^{1/2}\rho_\ell^{1/2}(y_\ell, y'_\ell) = \Psi(\rho_1(y_1, y'_1), \rho_2(y_2, y'_2)) \\ &\le \Psi(c_1(s_1, s'_1), \rho_2(y_2, y'_2)). \end{aligned}$$

Similarly, we can show

$$f(y_1, y_2, x) - f(y'_1, y'_2, x') \le \Psi(\rho_1(y_1, y'_1), c_2(s_2, s'_2)).$$

Theorem 9 implies the continuity of $\mathcal{I}$ on $\mathbb{R}_+^2$. □

A.5. Proofs in Section 6.2

A.5.1. Proof of Proposition 6. We prove Proposition 6(i) using a technique similar to Adjaho and Christensen [1]. For any $s_\ell = (y_\ell, x_\ell) \in \mathcal{S}_\ell$, we have

$$\begin{aligned} (f_S)_\lambda(s_1, s_2) &= \sup_{x'\in\mathcal{X}} \sup_{(y'_1, y'_2)\in\mathcal{Y}_1\times\mathcal{Y}_2} \left\{ -y'_2 d(x') - y'_1[1 - d(x')] - \sum_{1\le\ell\le 2} \lambda_\ell[|y_\ell - y_{\ell'}| + \|x_\ell - x'\|_2] \right\} \\ &= \sup_{x'\in\mathcal{X}} \left\{ \left[\sup_{y'_2\in\mathcal{Y}_2} \{-y'_2 d(x') - \lambda_2|y_2 - y_{y'_2}|\} + \sup_{y'_1\in\mathcal{Y}_1} \{-y'_1(1 - d(x')) - \lambda_1|y_1 - y'_1|\} \right] - \sum_{1\le\ell\le 2} \lambda_\ell\|x_\ell - x'\| \right\}. \end{aligned}$$

We note that

$$\sup_{y'_2\in\mathcal{Y}_2} \{-y'_2 d(x') - \lambda_2|y_2 - y'_2|\} = \begin{cases} \infty & \text{if } 0 \le \lambda_2 < 1 \\ -y_2 d(x') & \text{if } \lambda_2 \ge 1, \end{cases}$$

and

$$\sup_{y'_1\in\mathcal{Y}_1} \{-y'_1(1 - d(x')) - \lambda_1|y_1 - y'_1|\} = \begin{cases} \infty & \text{if } 0 \le \lambda_1 < 1 \\ -y_1(1 - d(x')) & \text{if } \lambda_1 \ge 1. \end{cases}$$

Therefore, we have for $\lambda_1 \ge 1$ and $\lambda_2 \ge 1$

$$\begin{aligned} (f_S)_\lambda(s_1, s_2) &= \sup_{x'\in\mathcal{X}} \left\{ -y_2 d(x') - y_1(1 - d(x')) - \sum_{1\le\ell\le 2} \lambda_\ell\|x_\ell - x'\| \right\} \\ &= -\min\{y_2 + \varphi_{\lambda,1}(x_1, x_2), y_1 + \varphi_{\lambda,0}(x_1, x_2)\}, \end{aligned}$$

where

$$\varphi_{\lambda,d}(x_1, x_2) = \min_{u\in\mathcal{X}:d(u)=d} \sum_{1\le\ell\le 2} \lambda_\ell\|x_\ell - u\|_2,$$

for $d \in \{0,1\}$. If $\lambda_1 < 1$ or $\lambda_2 < 1$, then $(f_S)_\lambda(s_1,s_2) = \infty$. As a result, we have

$$\begin{aligned}\mathrm{RW}(d) &= \inf_{\gamma\in\Sigma(\delta)} \mathbb{E}[Y_2 d(X) + Y_1(1-d(X))] = -\inf_{\lambda\in\mathbb{R}^2_+}\left[\langle\lambda,\delta\rangle + \sup_{\pi\in\Pi(\mu_{13},\mu_{23})}\int_{\mathcal{V}} (f_S)_\lambda \, d\pi\right]\\ &= -\inf_{\lambda\in[1,\infty)^2}\left[\langle\lambda,\delta\rangle + \sup_{\pi\in\Pi(\mu_{13},\mu_{23})}\int_{\mathcal{V}} -\min\{y_2+\varphi_{\lambda,1}(x_1,x_2), y_1+\varphi_{\lambda,0}(x_1,x_2)\}\, d\pi(v)\right]\\ &= \sup_{\lambda\in[1,\infty)^2}\left[\inf_{\pi\in\Pi(\mu_{13},\mu_{23})}\int_{\mathcal{V}} \min\{y_2+\varphi_{\lambda,1}(x_1,x_2), y_1+\varphi_{\lambda,0}(x_1,x_2)\}\, d\pi(v) - \langle\lambda,\delta\rangle\right].\end{aligned}$$

Next, we show Proposition 6(ii). Recall the set $\tilde{\Pi}$ defined in the proof of Proposition 4(ii). Here, $\tilde{\Pi}$ is the set of all the probability measures concentrated on $\{(y_1,x_1,y_2,x_2)\in\mathbb{R}^{2d+2} : x_1 = x_2\}$. Consider the following derivation:

$$\begin{aligned}\mathrm{RW}(d) &= \sup_{\lambda_0\ge 1,\lambda_2\ge 1}\left[\inf_{\pi\in\Pi(\mu_{13},\mu_{23})}\int_{\mathcal{V}} \min\{y_2+\varphi_{\lambda,1}(x_1,x_2), y_1+\varphi_{\lambda,0}(x_1,x_2)\}\, d\pi(v) - (\lambda_1+\lambda_2)\delta_0\right]\\ &\le \sup_{\lambda_1\ge 1,\lambda_2\ge 1}\left[\inf_{\pi\in\tilde{\Pi}}\int_{\mathcal{V}} \min\{y_2+\varphi_{\lambda,1}(x_1,x_2), y_1+\varphi_{\lambda,0}(x_1,x_2)\}\, d\pi(v) - (\lambda_1+\lambda_2)\delta_0\right].\end{aligned}$$

Recalling the functions h_0 and h_1 defined in Proposition 5, we notice that for all $(y_1,x_1,y_2,x_2)\in\tilde{\Pi}$,

$$\varphi_{\lambda,\ell}(x_1,x_2) = (\lambda_1+\lambda_2)h_\ell(x_1), \quad \forall \ell = 1,2.$$

As a result, we have

$$\begin{aligned}\mathrm{RW}(d) &\le \sup_{\lambda_1\ge 1,\lambda_2\ge 1}\left[\inf_{\pi\in\mathcal{F}(\mu_{13},\mu_{23})}\int_{\mathcal{S}} \min\{y_2+\varphi_{\lambda,1}(x), y_1+\varphi_{\lambda,0}(x)\}\, d\pi(s) - (\lambda_1+\lambda_2)\delta_0\right]\\ &= \sup_{\eta\ge 2}\left[\inf_{\pi\in\mathcal{F}(\mu_{13},\mu_{23})}\int_{\mathcal{S}} \min\{y_2+\eta h_1(x), y_1+\eta h_0(x)\}\, d\pi(s) - \eta\delta_0\right]\\ &\le \sup_{\eta\ge 1}\left[\inf_{\pi\in\mathcal{F}(\mu_{13},\mu_{23})}\int_{\mathcal{S}} \min\{y_2+\eta h_1(x), y_1+\eta h_0(x)\}\, d\pi(s) - \eta\delta_0\right]\\ &=_{(1)} \sup_{\eta\ge 1}\left[\inf_{\pi\in\mathcal{F}(\mu_{13},\mu_{23})}\mathbb{E}_X[\mathbb{E}(\min\{Y_2 - Y_1 + \eta h_1(X), \eta h_0(X)\}|X)] + \mathbb{E}(Y_1) - \eta\delta_0\right]\\ &= \sup_{\eta\ge 1}\left[\int_{\mathcal{S}} \min\{y_2+\eta h_1(x), y_1+\eta h_0(x)\}\, d\pi^*(s) - \eta\delta_0\right]\\ &= \mathrm{RW}_0(d),\end{aligned}$$

where Equation (1) follows from proposition 2.17 in Santambrogio [49] and the concavity of $y \longmapsto \min\{y+\eta h_1(x), \eta h_0(x)\}$ (see also section 4.3.1 in Adjaho and Christensen [1]). □

A.6. Proofs in Section 7

We provide a brief sketch of proofs in Section 7.

A.6.1. Proof of Theorem 10. Similarly to the proof of Theorem 2, it is sufficient to derive the dual reformulation of $\mathcal{I}_{\mathrm{D}}(\delta)$ for $\delta\in\mathbb{R}^L_{++}$. Let $\mathcal{P}_{\mathrm{D}}$ denote the set of $\gamma\in\mathcal{P}(\mathcal{V})$ that satisfies $\boldsymbol{K}_\ell(\mu_\ell,\gamma_\ell) < \infty$ for all $\ell\in[L]$ and $\int_{\mathcal{V}} g\, d\gamma > -\infty$. Taking the Legendre transform on $\mathcal{I}_{\mathrm{D}}$ yields that any $\lambda\in\mathbb{R}^2_+$

$$\begin{aligned}\mathcal{I}^\star_{\mathrm{D}}(\lambda) &:= \sup_{\delta\in\mathbb{R}^L_+}\{\mathcal{I}_{\mathrm{D}}(\delta) - \langle\lambda,\delta\rangle\} = \sup_{\delta\in\mathbb{R}^L_+}\sup_{\gamma\in\Sigma_{\mathrm{D}}(\delta)}\left\{\int_{\mathcal{V}} g\, d\gamma - \langle\lambda,\delta\rangle\right\}\\ &= \sup_{\gamma\in\mathcal{P}_{\mathrm{D}}}\underbrace{\left\{\int_{\mathcal{V}} g\, d\gamma - \sum_{\ell\in[L]}\lambda_\ell \mathbf{K}_\ell(\mu_\ell,\gamma_\ell)\right\}}_{:=I_{\mathrm{D},\lambda}[\gamma]} = \sup_{\gamma\in\mathcal{P}_{\mathrm{D}}} I_{\mathrm{D},\lambda}[\gamma].\end{aligned}$$

Using Lemma S.7 in the Online Supplement and similar reasoning as in the proof of Theorem 2, we can show

$$\mathcal{I}_{\mathrm{D}}^{\star}(\lambda) = \sup_{\gamma\in\mathcal{P}_{\mathrm{D}}} I_{\mathrm{D},\lambda}[\gamma] = \sup_{\pi\in\Gamma(\Pi,\varphi_\lambda)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda \, d\pi = \sup_{\pi\in\Pi(\mu_1,\dots,\mu_L)} \int_{\mathcal{V}} g_\lambda d\pi.$$

The desired result follows from Lemma S.4 in the Online Supplement. □

A.6.2. Proof of Theorem 11. Similarly to the proof of Theorem 3, it is sufficient to derive the dual reformulation of $\mathcal{I}(\delta)$ for $\delta\in\mathbb{R}_{++}^{L}$. Let $\overline{\mathcal{P}}$ denote the set of $\gamma\in\mathcal{P}(\mathcal{S})$ that satisfies $\boldsymbol{K}_\ell(\mu_{\ell,L},\gamma_{\ell,L})<\infty$ for all $\ell\in[L]$ and $\int_{\mathcal{S}} f d\gamma > -\infty$. Taking the Legendre transform on $\mathcal{I}$ yields that any $\lambda\in\mathbb{R}_{+}^{2}$

$$\begin{aligned}\mathcal{I}^{\star}(\lambda) &:= \sup_{\delta\in\mathbb{R}_{+}^{L}}\{\mathcal{I}(\delta)-\langle\lambda,\delta\rangle\} = \sup_{\delta\in\mathbb{R}_{+}^{L}}\sup_{\gamma\in\Sigma(\delta)}\left\{\int_{\mathcal{V}} f d\gamma - \langle\lambda,\delta\rangle\right\}\\ &= \sup_{\gamma\in\overline{\mathcal{P}}}\underbrace{\left\{\int_{\mathcal{V}} g d\gamma - \sum_{\ell\in[L]}\lambda_\ell \mathbf{K}_\ell(\mu_\ell,\gamma_\ell)\right\}}_{:=I_\lambda[\gamma]} = \sup_{\gamma\in\overline{\mathcal{P}}} I_\lambda[\gamma].\end{aligned}$$

For notational simplicity, we write $\Pi := \Pi(\mu_{1,L+1},\dots,\mu_{L,L+1})$. Using Lemma S.8 in the Online Supplement and similar reasoning as in the proof of Theorem 3, we can show

$$\mathcal{I}^{\star}(\lambda) = \sup_{\gamma\in\overline{\mathcal{P}}} I_\lambda[\gamma] = \sup_{\pi\in\Gamma(\Pi,\phi_\lambda)} \int_{\mathcal{V}\times\mathcal{V}} \varphi_\lambda d\pi = \sup_{\pi\in\Pi} \int_{\mathcal{V}} f_\lambda d\pi.$$

The desired result follows from Lemma S.4 in the Online Supplement. □

A.6.3. Proof of Proposition 7. The proof is identical to that of Proposition 6.

Endnotes

[1] When the marginals are univariate, optimal transport problem can be conveniently expressed in terms of copulas. Fan and Park [19], Fan and Park [20], Fan and Wu [21], Fan et al. [22], Ridder and Moffitt [46], and Firpo and Ridder [23] explicitly use copula tools.

[2] See Graham et al. [27] and Chen et al. [7] for general data combination problems.

[3] Section 2.3.3 provides a detailed comparison of our set-up and Awasthi et al. [2].

[4] By convention, we call all uncertainty sets based on optimal transport costs as Wasserstein uncertainty sets.

[5] The strong duality result in Zhang et al. [57] allows for general space $\mathcal{X}$.

[6] Because $\inf_{\gamma\in\Sigma(\delta)}\int_{\mathcal{S}} f(y_1,y_2)d\gamma(s)$ can be rewritten as $-\sup_{\gamma\in\Sigma(\delta)}\int_{\mathcal{S}}[-f(y_1,y_2)]d\gamma(s)$, we also refer to the lower limit as W-DMR-MP.

[7] During the revision of our paper, we learned that chapter 4 of Kent [32] presents a similar duality for nonoverlapping marginals and element-wise general penalty function discussed in Remark 6(ii).

[8] For multimarginals, the collection of given marginals can be more complicated than the nonoverlapping and overlapping marginals (see Rüschendorf [48], Embrechts and Puccetti [12], and Doan et al. [10]); we leave a complete treatment of the W-DMR with multimarginals to future work.

[9] Kido [33] mentions the possibility of allowing for covariate shift by incorporating uncertainty sets in, for example, Mo et al. [38] and Zhao et al. [58] for the distribution of the covariate in future work.

[10] To be more precise, $\pi((A_1\times\mathcal{S}_2)\times\mathcal{V}) = \mu_1(A_1)$ and $\pi((\mathcal{S}_1\times A_2)\times\mathcal{V}) = \mu_2(A_2)$ for all sets $A_1\in\mathcal{B}_{\mathcal{S}_1}$ and $A_2\in\mathcal{B}_{\mathcal{S}_2}$.

[11] To be more precise, the measure $\pi_n(\mathcal{V}\times\cdot)$ is in $\mathcal{P}_{\mathrm{D}}$.

[12] To be more precise, $\pi((A_1\times\mathcal{S}_2)\times\mathcal{S}) = \mu_{13}(A_1)$ and $\pi((\mathcal{S}_1\times A_2)\times\mathcal{S}) = \mu_{23}(A_2)$ for all Borel sets $A_1\in\mathcal{B}_{\mathcal{S}_1}$ and $A_2\in\mathcal{B}_{\mathcal{S}_2}$.

References

[1] Adjaho C, Christensen T (2022) Externally valid policy choice. Preprint, submitted May 11, https://arxiv.org/abs/2205.05561.

[2] Awasthi P, Jung C, Morgenstern J (2022) Distributionally robust data join. Preprint, submitted February 11, https://arxiv.org/abs/2202.05797.

[3] Bartl D, Drapeau S, Tangpi L (2020) Computational aspects of robust optimized certainty equivalents and option pricing. *Math. Finance* 30(1):287–309.

[4] Bertsekas DP, Shreve SE (1978) *Stochastic Optimal Control. The Discrete-Time Case*, Optimization and Neural Computation Series (Athena Scientific, Belmont, MA).

[5] Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* 44(2):565–600.

[6] Blanchet J, Murthy K, Nguyen VA (2021) Statistical analysis of Wasserstein distributionally robust estimators. Carlsson JG, Shier D, Greenberg HJ, eds. *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications* (INFORMS, Catonsville, MD), 227–254.

[7] Chen X, Hong H, Tarozzi A (2008) Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.* 36(2):808–843.

[8] Chen M, Du W, Tang Y, Jin Y, Yen GG (2022) A decomposition method for both additively and non-additively separable problems. *IEEE Trans. Evolutionary Computat.* 27(6):1720–1734.

[9] Cheridito P, Eckstein S (2023) Optimal transport and Wasserstein distances for causal models. Preprint, submitted March 24, https://arxiv.org/abs/2303.14085.
[10] Doan XV, Li X, Natarajan K (2015) Robustness to dependency in portfolio optimization using overlapping marginals. *Oper. Res.* 63(6):1468–1488.
[11] Eckstein S, Kupper M, Pohl M (2020) Robust risk aggregation with neural networks. *Math. Finance* 30(4):1229–1272.
[12] Embrechts P, Puccetti G (2010) Bounds for the sum of dependent risks having overlapping marginals. *J. Multivariate Anal.* 101(1):177–190.
[13] Embrechts P, Höing A, Juri A (2003) Using copulae to bound the Value-at-Risk for functions of dependent risks. *Finance Stochastics* 7(2):145–167.
[14] Embrechts P, Höing A, Puccetti G (2005) Worst VaR scenarios. *Insurance Math. Econom.* 37(1):115–134.
[15] Embrechts P, Puccetti G, Rüschendorf L (2013) Model uncertainty and VaR aggregation. *J. Banking Finance* 37(8):2750–2764.
[16] Ennaji H, Mérigot Q, Nenna L, Pass B (2022) Robust risk management via multi-marginal optimal transport. Preprint, submitted November 14, https://arxiv.org/abs/2211.07694.
[17] Fan K (1953) Minimax theorems. *Proc. Natl. Acad. Sci. USA* 39(1):42–47.
[18] Fan Y, Park SS (2009) Partial identification of the distribution of treatment effects and its confidence sets. Li Q, Racine JS, eds. *Advances in Econometrics*, vol. 25 (Emerald Group Publishing Limited, Leeds, UK), 3–70.
[19] Fan Y, Park SS (2010) Sharp bounds on the distribution of treatment effects and their statistical inference. *Econom. Theory* 26(3):931–951.
[20] Fan Y, Park SS (2012) Confidence intervals for the quantile of treatment effects in randomized experiments. *J. Econometrics* 167(2):330–344.
[21] Fan Y, Wu J (2009) Partial identification of the distribution of treatment effects in switching regime models and its confidence sets. *Rev. Econom. Stud.* 77(3):1002–1041.
[22] Fan Y, Guerre E, Zhu D (2017) Partial identification of functionals of the joint distribution of "potential outcomes." *J. Econometrics* 197(1):42–59.
[23] Firpo S, Ridder G (2019) Partial identification of the treatment effect distribution and its functionals. *J. Econometrics* 213(1):210–234.
[24] Frank MJ, Nelsen RB, Schweizer B (1987) Best-possible bounds for the distribution of a sum — A problem of Kolmogorov. *Probab. Theory Related Fields* 74(2):199–211.
[25] Gao R, Kleywegt A (2022) Distributionally robust stochastic optimization with Wasserstein distance. *Math. Oper. Res.* 48(2):603–655.
[26] Ghossoub M, Hall J, Saunders D (2023) Maximum spectral measures of risk with given risk factor marginal distributions. *Math. Oper. Res.* 48(2):1158–1182.
[27] Graham BS, de Xavier Pinto CC, Egel D (2016) Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *J. Bus. Econom. Statist.* 34(2):288–301.
[28] Jiang Y (2024) Duality of causal distributionally robust optimization: The discrete-time case. Preprint, submitted January 29, https://arxiv.org/abs/2401.16556.
[29] Kallus N, Mao X, Zhou A (2022) Assessing algorithmic fairness with unobserved protected class using data combination. *Management Sci.* 68(3):1959–1981.
[30] Kellerer HG (1964) Verteilungsfunktionen mit gegebenen marginalverteilungen. *Z. Wahrscheinlichkeitstheorie Verw. Gebiete* 3(3):247–270.
[31] Kellerer HG (1984) Duality theorems for marginal problems. *Z Wahrscheinlichkeitstheorie Verw. Gebiete* 67(4):399–432.
[32] Kent CR (2021) Optimization in the space of measures: New techniques from optimal transport. PhD dissertation, Stanford University, Stanford, CA.
[33] Kido D (2022) Distributionally robust policy learning with Wasserstein distance. Preprint, submitted May 10, https://arxiv.org/abs/2205.04637.
[34] Kitagawa T, Tetenov A (2018) Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616.
[35] Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. Netessine S, Shier D, Greenberg HJ, eds. *Operations Research and Management Science in the Age of Analytics* (INFORMS, Catonsville, MD), 130–166.
[36] Makarov GD (1982) Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory Probab. Appl.* 26(4):803–806.
[37] Mehta R, Kline J, Lokhande VS, Fung G, Singh V (2023) Efficient discrete multi marginal optimal transport regularization. *11th Internat. Conf. Learn. Representations* (ICLR, Appleton, WI).
[38] Mo W, Qi Z, Liu Y (2020) Learning optimal distributionally robust individualized treatment rules. *J. Amer. Statist. Assoc.* 116(534):659–674.
[39] Nenna L, Pass B (2022) An ODE characterisation of multi-marginal optimal transport. Preprint, submitted December 23, https://arxiv.org/abs/2212.12492.
[40] Pass B (2010) Uniqueness and Monge solutions in the multi-marginal optimal transportation problem. Preprint, submitted July 2, https://arxiv.org/abs/1007.0424.
[41] Pass B (2012) Multi-marginal optimal transport and multi-agent matching problems: Uniqueness and structure of solutions. Preprint, submitted October 27, https://arxiv.org/abs/1210.7372.
[42] Pass B (2015) Multi-marginal optimal transport: Theory and applications. *ESAIM Math. Model. Numer. Anal.* 49(6):1771–1790.
[43] Peyré G, Cuturi M (2018) Computational optimal transport. Preprint, submitted March 1, https://arxiv.org/abs/1803.00567.
[44] Puccetti G, Rüschendorf L (2012) Bounds for joint portfolios of dependent risks. *Statist. Risk Model.* 29(2):107–132.
[45] Rachev ST, Rüschendorf L (1998) *Mass Transportation Problems: Volume 1: Theory* (Springer Science & Business Media, New York).
[46] Ridder G, Moffitt R (2007) Chapter 75 the econometrics of data combination. Heckman JJ, Leamer EE, eds. *Handbook of Econometrics*, vol. 6, part B (Elsevier, Amsterdam), 5469–5547.
[47] Rüschendorf L (1982) Random variables with maximum sums. *Adv. Appl. Probab.* 14(3):623–632.
[48] Rüschendorf L (1991) Bounds for distributions with multivariate marginals. Mosler K, Scarsini M, eds. *Stochastic Orders and Decision under Risk*, IMS Lecture Notes Monograph Series, vol. 19 (Institute of Mathematical Statistics, Muenster, Germany), 285–310.
[49] Santambrogio F (2015) *Optimal Transport for Applied Mathematicians* (Springer International Publishing, Cham, Switzerland).

[50] Shortt RM (1983) Combinatorial methods in the study of marginal problems over separable spaces. *J. Math. Anal. Appl.* 97(2):462–479.
[51] Sinha A, Namkoong H, Volpi R, Duchi J (2017) Certifying some distributional robustness with principled adversarial training. Preprint, submitted October 29, https://arxiv.org/abs/1710.10571.
[52] Villani C (2009) *Optimal Transport: Old and New* (Springer, Berlin).
[53] Villani C (2021) *Topics in Optimal Transportation* (American Mathematical Society, Providence, RI).
[54] von Lindheim J (2022) Approximative algorithms for multi-marginal optimal transport and free-support Wasserstein Barycenters. Preprint, submitted February 2, https://arxiv.org/abs/2202.00954.
[55] Vorob'ev NN (1962) Consistent families of measures and their extensions. *Theory Probab. Appl.* 7(2):147–163.
[56] Yue MC, Kuhn D, Wiesemann W (2022) On linear optimization over Wasserstein balls. *Math. Programming* 195(1–2):1107–1122.
[57] Zhang L, Yang J, Gao R (2022) A simple duality proof for Wasserstein distributionally robust optimization. Preprint, submitted April 30, https://arxiv.org/abs/2205.00362.
[58] Zhao YQ, Zeng D, Tangen CM, Leblanc ML (2019) Robustifying trial-derived optimal treatment rules for a target population. *Electronic J. Statist.* 13(1):1717–1743.