
Bridging Multiple Worlds: Multi-marginal Optimal Transport for Causal Partial-identification Problem

Zijun Gao
USC Marshall

Shu Ge
MIT EECS

Jian Qian
MIT EECS

Abstract

Under the prevalent potential outcome model in causal inference, each unit is associated with multiple potential outcomes but at most one of which is observed, leading to many causal quantities being only partially identified. The inherent missing data issue echoes the multi-marginal optimal transport (MOT) problem, where marginal distributions are known, but how the marginals couple to form the joint distribution is unavailable. In this paper, we cast the causal partial identification problem in the framework of MOT with K margins and d -dimensional outcomes and obtain the exact partial identified set. In order to estimate the partial identified set via MOT, statistically, we establish a convergence rate of the plug-in MOT estimator for the ℓ_2 cost function stemming from the variance minimization problem and prove it is minimax optimal for arbitrary K and $d \leq 4$. We also extend the convergence result to general quadratic objective functions. Numerically, we demonstrate the efficacy of our method over synthetic datasets and several real-world datasets where our proposal consistently outperforms the baseline by a significant margin (over 70%). In addition, we provide efficient off-the-shelf implementations of MOT with general objective functions.

1 INTRODUCTION

The potential outcome model (Rubin 1974) has been extensively used to perform causal inference. Suppose there are K possible treatment levels, and each unit

i is associated with K potential outcomes $\mathbf{Y}_i(\cdot) = (Y_i(1), \dots, Y_i(K))$, where each potential outcome $Y_i(k) = (Y_i^1(k), \dots, Y_i^d(k))$ is a d dimensional vector drawn independently¹ from an unknown joint distribution μ_\star . Causal questions focus on comparing these potential outcomes to evaluate the treatment effect.

The fundamental challenge of causal inference is that we can only observe the potential outcome $Y_i = Y_i(W_i)$ at the realized treatment level $W_i \in \{1, \dots, K\}$, and it is impossible to observe the entire set of K potential outcomes. As a consequence, only the marginal distribution $\mu_\star(k)$ of $Y_i(k)$, $1 \leq k \leq K$ can be identified and how the marginals $\mu_\star(k)$ couple to form the full joint distribution μ_\star is not identifiable. This renders a large class of causal estimands² relying on the joint distribution μ_\star to be only partially identifiable, meaning that the causal estimands can not be precisely determined but can only be confined to a range of plausible values based on the available data. For partially-identified causal estimands, the statistical task is determining the identified set: the full set of values that are compatible with the observed data.

In the following, we provide an example of a causal quantity that could be only partially identified and whose partially identified set hasn't been fully characterized in the literature.

1.1 Motivating Example

To illustrate the concept of partial identifiability, we revisit the classic Neymanian confidence interval (Splawa-Neyman 1923). Consider $d = 1$ and the average contrast effect estimand $\tau_\beta := \sum_{k=1}^K \beta_k \bar{Y}(k)$, with $\bar{Y}(k) = \sum_{i=1}^n Y_i(k)/n$ and pre-fixed coefficients $\beta_k \in \mathbf{R}$. Suppose W_i are drawn from a completely randomized experiment³ with n_k units at the treat-

¹In addition to the super-population scenario, we discuss finite sample analysis in Section 1.1.

²Only causal estimands that can be expressed as the expectation of the sum of univariate functions of a single potential outcome can be identified.

³A completely randomized experiment involves a random sample of size n_k chosen without replacement from a

ment level k , then the Neymanian confidence interval for τ_β takes the form (Lu 2016),

$$\begin{aligned} V_\beta &:= \sum_{k=1}^K \frac{\beta_k^2 S_k^2}{(n/K)} - \frac{S_\tau^2}{n}, \\ S_k^2 &= \frac{\sum_{i=1}^n (Y_i(k) - \bar{Y}(k))^2}{n-1}, \\ S_\tau^2 &= \frac{\sum_{i=1}^n (\sum_{k=1}^K \beta_k Y_i(k) - \tau_\beta)^2}{n-1}. \end{aligned} \quad (1)$$

The variances S_k^2 can be estimated using the empirical variance of units with $W_i = k$. However, the variance S_τ^2 , remains unidentifiable because the outcomes $Y_i(k)$ for different k are not simultaneously observable for the same individual. The conventional estimator of the variance V_β substitutes S_k^2 with the empirical variance and replaces S_τ^2 by a trivial lower bound zero. The conventional variance estimator is upward-biased and thus induces an over-conservative confidence interval.

For cases where $K = 2$ (one treatment level and one control level), Aronow, Green, and Lee 2014 provide the sharp lower bound for S_τ^2 . Our proposed method below can be used to generalize the result to arbitrary K levels to offer a sharp lower bound of S_τ^2 as well as the narrowest confidence intervals. Practically, a narrower confidence interval often indicates that a smaller sample size, such as less patients to involve in a clinical trial, could be sufficient to make a discovery with statistical significance, thereby potentially reducing the associated costs.

1.2 Our Proposal

This paper addresses the partial identification problem under the potential outcome model using Multi-margin Optimal Transport (MOT). Optimal transport, particularly MOT, is a fast-growing research area partly due to the emerging applications in machine learning including networks (GANs) (Choi et al. 2018), domain adaptation (Hui et al. 2018), and Wasserstein barycenters (Agueh and Carlier 2011). The MOT problem considers a cost (objective) function of random variables with known marginal distributions and seeks to find the coupling of the marginal distributions that minimizes the expected cost (Pass 2015). In addition, MOT outputs the exact lower bound of the expected cost among all joint distributions obeying the marginal distributions, even if the optimal coupling is not unique.

There is a natural connection between the partially identified set and MOT: the marginal distribution of each potential outcome is identifiable but the joint finite sample of size n .

distribution can not be uniquely determined. For a causal estimand, there is an MOT problem whose optimal objective value corresponds to the lower (upper) limit of the causal quantity that can be achieved by some joint distribution respecting the identifiable marginals. Particularly, let Γ represent the collection of joint probabilities respecting the marginal distributions $\mu_\star(k)$, $1 \leq k \leq K$. For a causal estimand of the form $\mathbb{E}_{\mu_\star}[\ell(\mathbf{Y}_i(\cdot))]$, where ℓ can depend on all potential outcomes, the identified set can be expressed as

$$\{\mathbb{E}_\gamma[\ell(\mathbf{Y}_i(\cdot))] : \gamma \in \Gamma\}.$$

Since Γ is convex by definition and the estimand $\mathbb{E}_\gamma[\ell(\mathbf{Y}_i(\cdot))]$ is linear in γ , the identified set is always an interval⁴

$$\left[\inf_{\gamma \in \Gamma} \mathbb{E}_\gamma[\ell(\mathbf{Y}_i(\cdot))], \sup_{\gamma \in \Gamma} \mathbb{E}_\gamma[\ell(\mathbf{Y}_i(\cdot))] \right]. \quad (2)$$

The task of determining the identified set reduces to calculating the lower and upper limits, which can be obtained by solving two MOT problems. See Figure 1 for visualization.

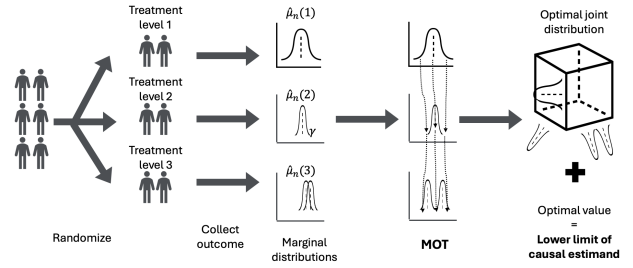


Figure 1: Diagram of solving partial identification problem via MOT. A treatment with multiple levels is applied to a group of individuals. The outcomes of each treatment group are collected to form the empirical marginal distribution denoted by $\hat{\mu}_n(k)$, $1 \leq k \leq K$. An MOT problem with the causal estimand as the objective function and empirical marginal distributions as the constraints are solved to obtain the lower limit of the partially identified set. The upper limit can be obtained similarly by flipping the sign of the MOT objective function.

We illustrate the connection between MOT and the partial identification problem using the motivating example in Section 1.1. For the partially identified S_τ^2 , the loss function associated is $\ell(\mathbf{Y}_i(\cdot)) := (n/(n-1))(\sum_{k=1}^K Y_i(k))^2 - \tau_\beta^2$. Here τ_β is identifiable, and

⁴We allow the interval to include $-\infty$ and ∞ . We also allow the interval to be degenerate (only containing a single point), or even empty.

the family of joint distributions Γ includes all the distributions whose k -th marginal distribution equals the distribution of $Y_i(k)$. Determining the lower bound of S_τ^2 is thus equivalent to minimizing $\mathbb{E}_\gamma[\ell(\mathbf{Y}_i(\cdot))]$ for $\gamma \in \Gamma$, which becomes a standard MOT problem. Theoretically, for this specific case, we show that the plug-in estimator (which computes the MOT value of the empirical marginal distributions) for the MOT optimal objective value converges at a rate of $O(n^{-1/2})$. Empirically, we apply our proposal to the Exptaxial layer growth dataset (Wu and Hamada 2011) in Section 4. Our variance estimator using the sharp lower bound of S_τ^2 sees a reduction of 8.16 percent compared to the conventional variance estimator (lower bounding S_τ^2 by zero), effectively decreasing the number of units or repeats required by the same percentage.

1.3 Contributions

1. Methodologically, our proposal recovers the exact partially identified set given the true marginal distributions. Our proposal can accommodate various scenarios including arbitrary objectives (beyond quadratic), multiple treatment levels, multiple treatments, and multiple outcomes (see Section 2 for examples). Concretely, we demonstrate that our proposal can be used to detect contrast treatment effect and interaction effect, analyze the covariance between different treatment effects, and shorten Neymanian confidence intervals (Equation (1)).
2. Theoretically, we analyze the plug-in estimator associated with the MOT problem where the true marginal distributions are replaced by the empirical marginal distributions. In particular, we focus on the ℓ_2 norm of the average potential outcomes as the objective. We show the plug-in estimator converges at the parametric rate $O(n^{-1/2})$ up to logarithmic terms for $d \leq 4$, which is surprisingly independent of the number of margins K . This is complemented with a matching lower bound, demonstrating that the plug-in estimator is min-max optimal for the particular case. The proof techniques involved, as elaborated in Section 3, could be of independent interest. The convergence of general quadratic objectives is also provided as an extension of Chizat et al. 2020.
3. Implementation-wise, we provide several solvers⁵ of MOT applicable to arbitrary objectives (not limited to quadratic objectives), addressing the lack of efficient off-the-shelf MOT packages. A notable feature is that our algorithm always

provides valid, albeit conservative, partially identified sets. On synthetic datasets, we compare our method to a case where the closed-form solution of the optimal values is known, showing that our estimates closely match the true optimal values, and in a scenario where the exact solution is unknown, we demonstrate that our method’s bounds are valid for the true value of the causal estimand (see details in Section 4). On real datasets, we validate our proposal on the aforementioned applications and showcase that our method achieves an improvement of approximately 70% or more compared to a baseline approach⁶ detailed in Section 2.

Notations and related works See notations in section 6 and related works in section 7.

2 MOT FORMULATION OF PARTIAL IDENTIFICATION PROBLEM

We recall briefly the relevant notations. We denote by $\mathbf{Y}_i(\cdot)$ the potential outcomes of unit i . At treatment level k , the potential outcome is denoted by $Y_i(k) = (Y_i^1(k), \dots, Y_i^d(k)) \in \mathbf{R}^d$. Let μ_\star be the joint distribution that are only partially identifiable and $\mu_\star(k)$ be its k -th margin for $k \in [K]$. Without loss of generality, we suppose the marginals $\mu_\star(k)$ for any $k \in [K]$ are supported on $B_1^d(0) \subset \mathbf{R}^d$ which is the d -dimensional unit ball.

We explore three types of objectives in the increasing order of specificity, namely, arbitrary loss functions, arbitrary quadratic functions, and average squared norm. The MOT algorithm we implement provides a lower bound for arbitrary loss functions and provably converges for arbitrary quadratic functions. The convergence of the empirical MOT value to the true MOT value holds for arbitrary quadratic functions with lower bounds proven for the average squared norm.

2.1 Arbitrary Function

In Section 1, we described the causal partial identification problem where the parameter of interest is the expectation of an arbitrary objective function ℓ . For the two-margin optimal transport with general objective function, the convergence of the empirical optimal transport cost has been established

⁵The implementation can be accessed at https://github.com/ShuGe-MIT/mot_project.

⁶The causal estimands in our applications under no model assumptions have not been thoroughly investigated in the literature and the baseline bound used is the most relevant counterpart to compare with to the authors’ knowledge.

for smooth objective functions with compact supports (Hundrieser, Staudt, and Munk 2022), objectives of the form $\ell(y(2) - y(1))$ (Manole and Niles-Weed 2024), or under bounded moment assumptions (Staudt and Hundrieser 2023). To the authors' knowledge, the convergence results for general MOT cost has not been explicitly stated. For the rest of the paper, we will focus on quadratic objective functions that possess favorable statistical properties and admit various applications.

2.2 Arbitrary Quadratic Function

Quadratic objective functions are practically useful (as demonstrated in the examples below) and the estimation error analysis can be helped by the quadratic structure. Let $A \in \mathbf{R}^{K \times K}$ be a matrix and recall that $\Gamma(\mu_\star) := \Gamma(\mu_\star(1), \dots, \mu_\star(K))$ denotes the set of all possible couplings of $\mu_\star(1), \dots, \mu_\star(K)$, i.e., distributions γ where the k -th margin $\gamma(k)$ is equal to $\mu_\star(k)$ for all $k \in [K]$. We define the Quadratic Multi-marginal Wasserstein (QMW) loss parametrized by A as

$$\begin{aligned} \text{QMW}(\mu_\star; A) &= \text{QMW}(\mu_\star(1), \dots, \mu_\star(K); A) \\ &:= \inf_{\gamma \in \Gamma(\mu_\star)} \int y^\top A y \gamma(dy), \end{aligned} \quad (3)$$

where $y^\top A y := \sum_{1 \leq i, j \leq K} a_{ij} y(i)^\top y(j)$.

2.3 Average Squared ℓ_2 Norm

Specifically, we focus on the following quadratic form which is flexible enough to encompass most of our examples of interest (Section 1.1 and example 1) as well as objectives in existing literature detailed below. Concretely, we define Multi-marginal Wasserstein (MW) loss as

$$\begin{aligned} \text{MW}_2^2(\mu_\star) &= \text{MW}_2^2(\mu_\star(1), \dots, \mu_\star(K)) \\ &:= \inf_{\gamma \in \Gamma(\mu_\star)} \int \left\| \frac{\sum_{k=1}^K y(k)}{K} \right\|_2^2 \gamma(dy), \end{aligned} \quad (4)$$

where the objective function in Eq. (4) corresponds to the integral of the squared ℓ_2 norm of the average potential outcome $(\sum_{k=1}^K y(k))/K$. This objective is a special case of in (3) with $A = (a_{ij} = 1/K^2)_{1 \leq i, j \leq K}$.

The quantity $\text{MW}_2^2(\mu_\star)$ has multiple connections to the existing literature. For instance, the expected sum of pairwise squared Wasserstein distance $\sum_{k \neq k'} \mathbb{E}_{\mu_\star}[(Y_i(k) - Y_i(k'))^2]$ is related to the repulsive harmonic cost arising from the weak interaction regime in Quantum Mechanics (Di Marino, Gerolin, and Nenna 2015) and can be written as $K^2 \text{MW}_2^2(\mu_\star)$ subtracted from the estimable quantity $2K \sum_{k=1}^K \mathbb{E}_{\mu_\star(k)}[(Y_i(k))^2]$. Another example is variance minimization problem (Rüschendorf 2013),

where the minimum variance can be written as $K^2 \text{MW}_2^2(\mu_\star)$ subtracting the estimable quantity $(\sum_{k=1}^K \mathbb{E}_{\mu_\star(k)}[Y_i(k)])^2$.

We describe a non-tight lower bound of the objective Eq. (4), which serves as the baseline to compare with. The objective Eq. (4) admits the following decomposition (Lemma 4),

$$\text{MW}_2^2(\mu_\star) = \frac{\left\| \sum_{k=1}^K \mathbb{E}_{\mu_\star(k)}[Y_i(k)] \right\|_2^2}{K^2} + \text{MW}_2^2(\bar{\mu}_\star), \quad (5)$$

where $\bar{\mu}_\star$ denotes the centered counterpart of μ_\star , that is, $\bar{\mu}_\star(k)$ is a translation of $\mu_\star(k)$ by $-\mathbb{E}_{\mu_\star(k)}[Y_i(k)]$. The first term on the right-hand side of Eq. (5) depends solely on the first moment of each margin, and the second term contains the information of higher-order moments. Since $\text{MW}_2^2(\bar{\mu}_\star)$ is non-negative, the first term is a straightforward lower bound of $\text{MW}_2^2(\mu_\star)$, which we refer to as the baseline lower bound. When $\text{MW}_2^2(\bar{\mu}_\star)$ is zero, i.e., $\mu_\star(k)$ satisfying the joint mixability property (Wang, Peng, and Yang 2013), the baseline lower bound is adequate. When the marginal distributions differ in second or higher-order moments, the baseline is not sufficient and our method offers a potentially significantly tighter lower bound (see Section 4.3.1 for real data demonstrations).

2.4 Examples

We provide examples of causal partial-identification problems that can be formulated in terms of the objective in Eq. (4) or (3).

Example 1 (Detection of interaction treatment effect). *Suppose there are two binary treatments. We use $Y_i(0, 0)$ to denote the potential outcome under control of both treatments, and similarly for $Y_i(0, 1)$, $Y_i(1, 0)$, $Y_i(1, 1)$, and the potential outcomes $Y_i(\cdot) = (Y_i(1, 1), -Y_i(1, 0), -Y_i(0, 1), Y_i(0, 0))$ possess four margins in total. Identifying a non-zero interaction treatment effect, i.e., $Y_i(1, 1) - Y_i(1, 0) - Y_i(0, 1) + Y_i(0, 0) \neq 0$, is crucial as it can facilitate the detection of drug synergy (whether different drugs' combined effect is greater than the sum of their individual effects). The tight lower bound of $\mathbb{E}_{\mu_\star}[(Y_i(1, 1) - Y_i(1, 0) - Y_i(0, 1) + Y_i(0, 0))^2]$ can be solved for as in Eq. (4), and a significantly positive lower bound suggests the existence of interaction effect.*

Example 2 (Detection of contrast effect). *The method in Example 1 can also be applied to detect the contrast effect of a treatment with multiple levels. Suppose there is a treatment with K levels, and the corresponding potential outcomes are associated with K margins. Unlike the case of a binary treatment, there are various ways to define an estimand for treatments*

of multiple levels, and a popular choice is considering a contrast vector $\beta \in \mathbf{R}^K$, $\sum_{k=1}^K \beta_k = 0$ and the associated contrast treatment effect $\sum_{k=1}^K \beta_k Y_i(k)$. Similar to the interaction effect, a significantly positive lower bound of $\mathbb{E}_{\mu_*}[(\sum_{k=1}^K \beta_k Y_i(k))^2]$ is a sign to reject the null hypothesis of no contrast effect.

We remark that standard methods for testing the existence of treatment effects construct a confidence interval for the treatment effect estimator $\sum_{i=1}^n (n/K)^{-1} \sum_{k=1}^K \beta_k \mathbb{1}_{\{W_i=k\}} Y_i$ (Imbens and Rubin 2015). This is testing whether the average contrast treatment effect $\mathbb{E}_{\mu_*}[\sum_{k=1}^K \beta_k Y_i(k)]$ is zero but we are testing whether all the individual contrast treatment effect is zero, i.e., $\sum_{k=1}^K \beta_k Y_i(k) = 0$ for all i .

Example 3 (Analysis of variance (ANOVA)). *Analysis of variance (ANOVA) is a standard topic in statistics to determine the importance of different variables, and recent research (Hines, Diaz-Ordaz, and Vansteelandt 2022) has focused on using ANOVA to define the importance of various treatments. Particularly, Hines, Diaz-Ordaz, and Vansteelandt 2022 use variance to quantify the importance, which is a special case of the quadratic form (with the identifiable mean squared subtracted). Our work can be used to provide lower and upper bounds of the importance of a treatment.*

Example 4 (Covariance between treatment effects). *The covariance between treatment effects of different dimensions of the outcomes $\text{Cov}_{\mu_*}(\tau_i^1, \tau_i^2)$ with $\tau_i^j = Y_i^j(2) - Y_i^j(1)$, $j = 1, 2$ takes the form of Eq. (3) and its upper and lower bounds can be obtained by solving the corresponding MOT problems.*

By investigating the exact lower and upper bounds of the covariance, we provide a model-free way to detect positively or negatively associated treatment effects (Section 14.1). There are many relevant applications in economics, e.g., the effects of a promotion scheme on the sale of complementary goods shall be positively correlated, and social sciences, e.g., the effects of a subsidy program on the full-time employment and part-time employment rate, might be negatively correlated.

3 THEORETICAL CHARACTERIZATION OF THE ESTIMATION ERROR

In this section, we consider the minimax estimation error of the value $\text{MW}_2^2(\mu_*)$ and $\text{QMW}(\mu_*; A)$ with n empirical samples. Concretely, suppose we have n i.i.d. samples from the distribution μ_* and assigned n/K samples to each treatment level k . For simplicity, we assume n/K is an integer. Specifically, suppose we have samples $\{Y_i(k)\}_{i \in [n/K], k \in [K]}$. Let the empirical distribution be $\hat{\mu}_n = (\hat{\mu}_n(1), \dots, \hat{\mu}_n(K))$,

where $\hat{\mu}_n(k) = \text{Unif}(\{Y_i(k)\}_{i \in [n/K]})$. In Section 3.1, we establish the parametric rate of convergence for $\text{MW}_2^2(\hat{\mu}_n)$ to $\text{MW}_2^2(\mu_*)$ for $d \leq 4$ and convergence result for $\text{QMW}(\hat{\mu}_n; A)$ to $\text{QMW}(\mu_*; A)$, respectively. Then we present a matching minimax lower bound for the estimation error of $\text{MW}_2^2(\mu_*)$ when $d \leq 4$ in Section 3.2.

3.1 Upper Bound

In this section, we prove finite sample upper bounds for the convergence of $\text{MW}_2^2(\hat{\mu}_n)$ and $\text{QMW}(\hat{\mu}_n; A)$ to $\text{MW}_2^2(\mu_*)$ and $\text{QMW}(\mu_*; A)$, respectively. Similar convergence results have been shown for optimal transport (Chizat et al. 2020). For the convergence of $\text{MW}_2^2(\hat{\mu}_n)$, we first apply the techniques from Chizat et al. 2020 to obtain a suboptimal (sometimes vacuous) convergence rate and then refine their analysis using more detailed properties of the MOT. Concretely, for $d = 1$, applying techniques from Chizat et al. 2020 yields a convergence rate of \sqrt{K}/\sqrt{n} , where K is the number of margins and n is the number of samples. This rate is not ideal as K can range from 0 to n . For $K = \sqrt{n}$, the rate becomes $1/n^{1/4}$, and if K/n is constant, the bound is vacuous. Our key technical contribution is removing this dependence on K through a subtle two-step approach, uncovering a surprising property of the optimal coupling for MOT. We subsequently extend the results of Chizat et al. 2020 to achieve the convergence of $\text{QMW}(\mu_*; A)$.

Theorem 1. *For any n, d, K , we have*

$$\begin{aligned} \mathbb{E} [|\text{MW}_2^2(\hat{\mu}_n) - \text{MW}_2^2(\mu_*)|] &\lesssim n^{-1/2} \\ &+ (1/K) \wedge \begin{cases} K^{-1/2}(n/K)^{-2/d}, & \text{if } d > 4, \\ n^{-1/2} \log(n/K), & \text{if } d = 4, \\ n^{-1/2}, & \text{if } d < 4, \end{cases} \end{aligned} \quad (6)$$

where the notation \lesssim hides constants that only depend on the dimension d . In particular, for $d \leq 4$, we have

$$\mathbb{E} [|\text{MW}_2^2(\hat{\mu}_n) - \text{MW}_2^2(\mu_*)|] \lesssim \tilde{O}(n^{-1/2}).$$

Proof sketch. The full proof is deferred to section 10.1. According to the techniques from Chizat et al. 2020, to achieve our stated convergence rate, the main challenge is to control $\sup_y \|\sum_k y(k)\|$ for any point $y = (y(1), \dots, y(K))$ in the support of the optimal coupling. This is difficult when $\mu_*(1), \dots, \mu_*(K)$ are arbitrary distributions with densities. The optimal coupling will also have a density, and adding any coupled point y with positive density (but zero probability) will not affect the ℓ_2 cost but will make $\sup_y \|\sum_k y(k)\|$ uncontrollable. We address this difficulty by (1) controlling $\sup_y \|\sum_k y(k)\|$ in the support of the optimal coupling

for discretely supported distributions, and (2) generalizing the convergence rate by bypassing the techniques from Chizat et al. 2020.

For discretely supported distributions, the probability mass of any support of the optimal coupling is positive, avoiding issues seen with continuous distributions. If two coupled points y_1 and y_2 are in the support, any exchange in their arguments will not decrease the coupled ℓ_2 cost. This is the analogue of the argument for optimal transport, where a change in the optimal transport map will not reduce the transport value. This surprisingly provides control over $\sup_y \|\sum_k y(k)\|$ for y in the support of the optimal coupling, as shown in lemma 7. Thus, somewhat unexpectedly, the ℓ_2 norm minimization forces the optimal coupling to achieve point-wise ℓ_2 norm guarantees for discretely supported distributions, despite being an average minimization.

The second step is to generalize the convergence to arbitrary distributions, which is challenging as we can't control $\sup_y \|\sum_k y(k)\|$ generally. To show convergence in total ℓ_2 cost, we use an imaginary sampling process. Suppose we sample enough points from $\mu_*(1), \dots, \mu_*(K)$ to form empirical distributions $\mu_N(1), \dots, \mu_N(K)$, with N arbitrarily large. The convergence of $\text{MW}_2^2(\mu_N)$ to $\text{MW}_2^2(\mu_*)$ holds with a suboptimal rate by directly applying Chizat et al. 2020, but since N is imaginary, we choose it large enough to eliminate the bad dependence on K . These empirical distributions are discretely supported, allowing us to achieve the desired rate by sampling from them, albeit with replacement. When N is large enough, sampling n points with replacement approximates sampling n points without replacement (where the latter is equivalent to sampling n points from μ_*). Thus concluding our proof. \square

General quadratic functions The convergence result can be obtained for arbitrary quadratic functions by generalizing the approach from Chizat et al. 2020. Concretely, we have the following theorem.

Proposition 1. *For any symmetric matrix $A \in \mathbb{R}^{K \times K}$, we have*

$$\begin{aligned} \mathbb{E} [|\text{QMW}(\hat{\mu}_n; A) - \text{QMW}(\mu_*; A)|] &\lesssim K \sqrt{\sum_{k,l} a_{kl}^2} \cdot n^{-1/2} \\ &+ \left(\sum_{k=1}^K |a_{kk}| \right) \wedge \|A\|_{1,\infty} K(n/K)^{-2/(d \vee 4)}, \end{aligned}$$

where the notation \lesssim hides constants that depend on the dimension d and log terms that depend on K, n , and where $\|A\|_{1,\infty} := \sup_k (\sum_l |a_{kl}|)$.

3.2 Lower Bound

In this section, we show that the minimax lower bound on the convergence for any estimator is $\Omega(n^{-1/2})$ for estimating the MOT problem (Eq. (4)). This, together with the upper bound shown in Theorem 1, asserts that when $d \leq 4$, the plug-in estimation is minimax optimal.

Proposition 2. *For any $K \geq 2$, any $d \geq 1$, and any estimator $F(\{Y_i(k)\}_{i \in [n/K], k \in [K]})$, we have*

$$\begin{aligned} &\sup_{\mu_*} \mathbb{E} [|F(\{Y_i(k)\}_{i,k}) - \text{MW}_2^2(\mu_*)|] \\ &\gtrsim \begin{cases} n^{-1/2} & \text{if } d \leq 4, \\ K^{-2}(n/K \log(n/K))^{-2/d} \wedge n^{-1/2} & \text{if } d > 4, \end{cases} \end{aligned}$$

Proof sketch. The full proof is deferred to Section 10.2. This proof uses Le Cam's two-point method (Lemma 1 of Yu 1997), where we consider $\mu_*(1) = \dots = \mu_*(K) = \text{Ber}(1/2)$ or $\mu_*(1) = \dots = \mu_*(K) = \text{Ber}(1/2 + \varepsilon)$. When $\varepsilon = \Theta(n^{-1/2})$ is small enough, then there is a constant probability any estimator would not be able to differentiate the two cases. Here $\text{Ber}(p)$ denotes the Bernoulli distribution with probability parameter p . Meanwhile, the divergence between the target values in these two cases scales with ε . Furthermore, by the lower bound of Wasserstein distance estimation from theorem 22 of Manole and Niles-Weed 2024, we obtain a lower bound of $K^{-2}(n/K)^{-2/d}$ when $d > 4$. \square

4 NUMERICAL EXPERIMENTS

4.1 Implementation

There are several algorithms proposed to solve the MOT problem in the literature. For example, Sinkhorn's algorithm (Peyré, Cuturi, et al. 2019, Section 10.1), Multi-marginal Sinkhorn (Lin et al. 2022), Primal-Dual Accelerated Alternating Minimization (PD-AAM) (Tupitsa et al. 2020), etc. In addition, we also consider the Multi-marginal Greenkhorn algorithm which is extended from Altschuler, Weed, and Rigollet 2018. The detail of the Multi-marginal Greenkhorn algorithm and Multi-marginal Sinkhorn algorithm is deferred to Appendix 11. We provide implementations for all four algorithms, while our numerical results are obtained from MOT Sinkhorn for the following three advantages: (1) This algorithm has a convergence guarantee, in which case we can use the dual objective as a lower bound. We obtained the lower bound for the Epitaxialgrowth dataset in Section 4.3.1 and Section 4.3.2 using this bound. (2) The algorithm is compatible with the celebrated log sum exponential trick in our implementation to avoid instability due to floating point errors caused by finite

machine representation. (3) Even when the algorithm does not converge, during the iteration lower bounds for the target value can be obtained. In particular, the lower bounds in Section 4.3.1 and that in Section 14.1 are obtained using the lower bounds obtained during the iteration without convergence.

4.2 Synthetic Dataset

We consider 1-d Gaussian distribution in this section, and include another synthetic data analysis of 2-d Gaussian distribution in Section 12.1. We pick $K = 3$ margins, each following a Gaussian distribution with different standard deviations. Here, we generate 200 data points in each margin (600 data points in total), with the standard deviations of the margins being 2, 0.3, 0.1, respectively. The objective function chosen here is $(Y(1) + Y(2) + Y(3))^2/9$. For any $k \in [K]$, if the marginal distributions are Gaussian, i.e. $\mu(k) = \mathcal{N}(0, \sigma_k^2)$ for $\sigma_k > 0$, then we have the following explicit form for the theoretical optimal multi-marginal Wasserstein loss.

Lemma 1 (Xiao and Yao 2020). *Let $\mu(1) = \mathcal{N}(0, \sigma_1^2), \dots, \mu(K) = \mathcal{N}(0, \sigma_K^2)$ be K Gaussian distributions. Then we have*

$$\text{MW}_2^2(\mu) = \frac{1}{K^2} \left(\left(2 \max_{k \in [K]} \sigma_k - \sum_{l=1}^K \sigma_l \right) \vee 0 \right)^2.$$

Using the provided standard deviation values from the data generation process, the corresponding theoretical optimal loss can be calculated as 0.28. The empirical MOT lower bound we get by running our Sinkhorn algorithm is 0.24. We note that the empirical optimal value we obtained serves as a lower bound for the theoretical optimal value.

4.3 Real Data Applications

4.3.1 Detection Of Treatment Effect

Continued from Example 1, we compute tight lower bounds for causal estimands taking the form $\mathbb{E}_{\mu_*}[(\sum_{k \in [K]} \beta_k Y_i(k))^2]$ on three real datasets (results are summarized in Table 1). We explain the three experiments in detail below. For details of the datasets, please refer to Section 13.

We first consider the estimand $\mathbb{E}_{\mu_*}[(Y_i(1, 1) - Y_i(1, 0) - Y_i(0, 1) + Y_i(0, 0))^2]$ related to interaction effects. We use the Epitaxial Layer Growth data (Wu and Hamada 2011), which investigates the impact of experimental factors on the thickness of growing an epitaxial layer on polished silicon wafers. In particular, we focus on the interaction of two binary treatments: susceptor’s rotation and nozzle position, and there are 6 data

points per treatment combination. Using the Multi-marginal Sinkhorn algorithm with error margin $\epsilon = 0.001$, the algorithm converges, and we get the lower bound to be 0.0460. We remark that the lower bound we get is 102.8% higher than the baseline lower bound.

Next, we consider the estimand $\mathbb{E}_{\mu_*}[\|Y_i(1, 0)/2 + Y_i(0, 1)/2 - Y_i(0, 0)\|_2^2]$ related to contrast effects. We use the Helpfulness data simulated from Jia 2022, which investigates factors influencing individuals’ empathy. In particular, we focus on two binary treatments: whether the individual has diverse experiences (i.e., has experienced both experimental conditions analogous to “wealth” and “poverty”) and whether the individual was provided with motivations for being empathetic. We consider the margin (0, 0) as the control level (141 units), (0, 1), (1, 0) as two treatment levels (96, 41 units, respectively), and adopt the contrast effect $Y_i(1, 0)/2 + Y_i(0, 1)/2 - Y_i(0, 0)$. Using the Multimarginal Sinkhorn algorithm with error margin $\epsilon = 0.001$, the algorithm converges, and we get the lower bound of the variance to be 0.432. We remark that the lower bound we get is also 79.5% higher than the baseline lower bound.

Finally, we consider the estimand $\mathbb{E}_{\mu_*}[\|Y_i(1, 0)/2 + Y_i(0, 1)/2 - Y_i(0, 0)\|_2^2]$ for the Education data⁷ from the Student Achievement and Retention (STAR) Demonstration Project (Angrist, Lang, and Oreopoulos 2009). Here researchers investigate the impact of two scholarship incentive programs, the Student Support Program (SSP) and the Student Fellowship Program (SFP), to the academic performance. We consider the margin (0, 0) as the control level (161 units), (0, 1), (1, 0) as two treatment levels (37, 36 units, respectively), and adopt the contrast effect $Y_i(1, 0)/2 + Y_i(0, 1)/2 - Y_i(0, 0)$. Using the Multimarginal Sinkhorn algorithm with error margin $\epsilon = 0.001$, the algorithm converges, and we get the lower bound of the variance to be 0.0654. We remark that the lower bound we get is also 72.3% higher than the baseline lower bound.

Table 1: Lower bounds for treatment effect detection.

Dataset	Baseline lower bound	MOT lower bound	Percentage improvement
Epitaxial Layer Growth	0.0227	0.0460	102.8%
Helpfulness	0.241	0.432	79.5%
Education	0.0380	0.0654	72.3%

⁷We focus on students whose baseline GPA falls within the lower 20%.

4.3.2 Neymanian Confidence Interval

Continued from Section 1.1, we apply the proposed method to tighten the conventional estimator of the variance (1). We return to the Expatial layer growth dataset. In particular, we focus on the three treatment levels (0, 1), (1, 0), (1, 1) and consider the contrast $(Y_i(1, 1) - Y_i(1, 0)/2 - Y_i(0, 1)/2)$. Our lower bound of S_τ^2 in (1) is noticeably greater than zero, indicating that bounding it by zero is overly conservative. As a result, our estimated variance has decreased the conventional counterpart by a significant 8.16 percent. For sample size calculation, a reduction of 8.16 percent in variance equates to a saving of 8.16 percent in the number of units/repeats required. The results are shown in Table 2.

Table 2: Neymanian confidence intervals. The contrast effect estimate is 0.0533.

Method	Variance estimate	95% Confidence interval for τ_N
Convention	0.113	(−0.609, 0.711)
Our proposal	0.104	(−0.581, 0.684)

5 DISCUSSION

In this work, we advocate for using MOT to obtain identified sets for causal estimands that are only partially identifiable. We perform theoretical analysis of the plug-in estimator of the identified sets, provide off-the-shelf implementation, and demonstrate the potentials of the approach for various causal tasks across multiple datasets.

We provide several future research directions.

- Covariate-assisted MOT. In the presence of covariates X , the conditional marginal distributions $Y(k) \mid X$ should be respected, a more stringent condition compared to preserving unconditional marginal distributions. We denote the set of joint couplings that adhere to these conditional marginals as Γ_X , then $\Gamma_X \subseteq \Gamma$ and yields larger population lower bounds of the objectives (4) and (3) with Γ replaced by Γ_X . For empirical lower bound, conditioning on more covariates will decrease the estimation accuracy of the plug-in estimator. Striking a balance between the population’s lower bound and the empirical performance by selecting a subset of covariates and solving the associated MOT problem is of interest. See (Lin et al. 2025) for an extension along this direction.
- Approximate joint mixability. For any K cen-

tered distributions $\mu(1), \dots, \mu(K)$ on \mathbb{R}^d , they are jointly mixable if there exists a coupling such that for any $Y(\cdot)$ on the support of the coupling, $\sum_{k=1}^K Y(k)/K = 0$ (Wang and Wang 2016). The objective in Eq. (4) can be used to define a generalization of the joint mixability, where the level of joint mixability is evaluated by the 2-norm. One interesting phenomenon as an implication of Lemma 7 is that the 2-norm approximate joint-mixability implies approximate joint-mixability in infinity norm for discretely supported distributions. Concretely, for any $Y(\cdot)$ on the support of the optimal coupling in terms of objective in Eq.(4), we have $|\sum_{k=1}^K Y(k)/K| \lesssim 1/\sqrt{K}$ almost surely. This posits a promising research direction of approximate joint mixability. Many open problems can be considered, e.g., can we extend Lemma 7 to arbitrary distributions? What are the relationships between the p -norm approximate joint mixability for $p \geq 1$?

- Implementation. There is a tradeoff between precision and the run time/resources. Especially when the accuracy tolerance ε is small, the algorithm might become unstable due to floating point errors. Also, the run time will be much higher. Compared to OT, MOT requires a higher resolution, or equivalently a smaller tolerance parameter ε , due to a larger number (exponential to the number of margins) of cell probabilities, making the acceleration of even higher importance.

Future work could explore potential ways to achieve higher precision in a more efficient way. One possibility is utilizing a low-rank tensor approximation to the quadratic objective function. In addition, in the current implementation, we use greedy Sinkhorn where the parameters are updated every K iterations (after we calculate all K margins). We have seen numerical success in simply updating one margin per iteration in a round-robin fashion, which is computationally economic. However, the convergence of this algorithm remains to be studied.

Acknowledgements

We thank Gabriel Peyré for the useful discussion on the implementation of MOT algorithms. Jian Qian acknowledges support from ARO through award W911NF-21-1-0328, as well as the Simons Foundation and the NSF through awards DMS-2031883 and PHY-2019786. Shu Ge acknowledges the support of ARO through award W911NF-21-1-0328.

References

- Agueh, Martial and Guillaume Carlier (2011). “Barycenters in the Wasserstein space”. In: *SIAM Journal on Mathematical Analysis* 43.2, pp. 904–924.
- Altschuler, Jason, Jonathan Niles-Weed, and Philippe Rigollet (2017). “Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration”. In: *Advances in neural information processing systems* 30.
- Altschuler, Jason, Jonathan Weed, and Philippe Rigollet (2018). *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration*. arXiv: 1705.09634 [cs.DS].
- Angrist, Joshua, Daniel Lang, and Philip Oreopoulos (2009). “Incentives and services for college achievement: Evidence from a randomized trial”. In: *American Economic Journal: Applied Economics* 1.1, pp. 136–163.
- Aronow, Peter M, Donald P Green, and Donald KK Lee (2014). “Sharp bounds on the variance in randomized experiments”. In: *The Annals of Statistics* 42.3, pp. 850–871.
- Balakrishnan, Sivaraman, Edward Kennedy, and Larry Wasserman (2023). “Conservative inference for counterfactuals”. In: *arXiv preprint arXiv:2310.12757*.
- Bickel, Peter J et al. (1993). *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Springer.
- Chizat, Lenaic et al. (2020). “Faster wasserstein distance estimation with the sinkhorn divergence”. In: *Advances in Neural Information Processing Systems* 33, pp. 2257–2269.
- Choi, Yunjeon et al. (2018). “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797.
- Cuturi, Marco (2013). *Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances*. arXiv: 1306.0895 [stat.ML].
- Di Marino, Simone, Augusto Gerolin, and Luca Nenna (2015). “Optimal transportation theory with repulsive costs”. In: *Topological optimization and optimal transport* 17, pp. 204–256.
- Fan, Yanqin and Sang Soo Park (2010). “Sharp bounds on the distribution of treatment effects and their statistical inference”. In: *Econometric Theory* 26.3, pp. 931–951.
- Flamary, Rémi et al. (2021). “Pot: Python optimal transport”. In: *Journal of Machine Learning Research* 22.78, pp. 1–8.
- Galichon, Alfred (2018). *Optimal transport methods in economics*. Princeton University Press.
- Guntuboyina, Adityanand and Bodhisattva Sen (2012). “L1 covering numbers for uniformly bounded convex functions”. In: *Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, pp. 12–1.
- Heckman, James J, Jeffrey Smith, and Nancy Clements (1997). In: *The Review of Economic Studies* 64.4, pp. 487–535.
- Hines, Oliver, Karla Diaz-Ordaz, and Stijn Vansteelandt (2022). “Variable importance measures for heterogeneous causal effects”. In: *arXiv preprint arXiv:2204.06030*.
- Hui, Le et al. (2018). “Unsupervised multi-domain image translation with domain-specific encoders/decoders”. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 2044–2049.
- Hundrieser, Shayan, Thomas Staudt, and Axel Munk (2022). “Empirical optimal transport between different measures adapts to lower complexity”. In: *arXiv preprint arXiv:2202.10434*.
- Imbens, Guido W and Donald B Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.
- Jia, Tingyan (2022). “Empathy, Motivated Reasoning, And Redistribution”. In.
- Kellerer, Hans G (1984). “Duality theorems for marginal problems”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 67, pp. 399–432.
- Kline, Brendan and Elie Tamer (2023). “Recent developments in partial identification”. In: *Annual Review of Economics* 15, pp. 125–150.
- Laan, Mark J and James M Robins (2003). *Unified methods for censored longitudinal data and causality*. Springer.
- Lin, Sirui et al. (2025). “Tightening Causal Bounds via Covariate-Aware Optimal Transport”. In: *arXiv preprint arXiv:2502.01164*.
- Lin, Tianyi, Nhat Ho, and Michael Jordan (2019). “On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms”. In: *International Conference on Machine Learning*. PMLR, pp. 3982–3991.
- Lin, Tianyi et al. (2022). “On the complexity of approximating multimarginal optimal transport”. In: *The Journal of Machine Learning Research* 23.1, pp. 2835–2877.
- Lu, Jiannan (2016). “On randomization-based and regression-based inferences for 2k factorial designs”. In: *Statistics & Probability Letters* 112, pp. 72–78.
- Manole, Tudor and Jonathan Niles-Weed (2024). “Sharp convergence rates for empirical optimal

- transport with smooth costs”. In: *The Annals of Applied Probability* 34.1B, pp. 1108–1135.
- Manski, Charles F (1997). “The mixing problem in programme evaluation”. In: *The Review of Economic Studies* 64.4, pp. 537–553.
- Nelsen, Roger B (2006). *An introduction to copulas*. Springer.
- Pass, Brendan (2015). “Multi-marginal optimal transport: theory and applications”. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 49.6, pp. 1771–1790.
- Peyré, Gabriel, Marco Cuturi, et al. (2019). “Computational optimal transport: With applications to data science”. In: *Foundations and Trends® in Machine Learning* 11.5-6, pp. 355–607.
- Polyanskiy, Yury and Yihong Wu (2022+). “Information Theory: From Coding to Learning”. In: *Cambridge University Press*. URL: <https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf>.
- Rubin, Donald B (1974). “Estimating causal effects of treatments in randomized and nonrandomized studies”. In: *Journal of Educational Psychology* 66.5, pp. 688–701.
- Rüschendorf, Ludger (2013). “Mathematical risk analysis”. In: *Springer Ser. Oper. Res. Financ. Eng. Springer, Heidelberg*.
- Splawa-Neyman, Jerzy (1923). “On the application of probability theory to agricultural experiments. Essay on principles. Section 9.” In: *Statistical Science*, pp. 465–472.
- Staudt, Thomas and Shayan Hundrieser (2023). “Convergence of Empirical Optimal Transport in Unbounded Settings”. In: *arXiv preprint arXiv:2306.11499*.
- Strassen, Volker (1965). “The existence of probability measures with given marginals”. In: *The Annals of Mathematical Statistics* 36.2, pp. 423–439.
- Tupitsa, Nazarii et al. (2020). “Multimarginal optimal transport by accelerated alternating minimization”. In: *2020 59th IEEE conference on decision and control (cdc)*. IEEE, pp. 6132–6137.
- Vaart, Aad W Van der (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.
- Vershynin, Roman (2018). *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.
- Wainwright, Martin J (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press.
- Wang, Bin and Ruodu Wang (2016). “Joint mixability”. In: *Mathematics of Operations Research* 41.3, pp. 808–826.
- Wang, Ruodu, Liang Peng, and Jingping Yang (2013). “Bounds for the sum of dependent risks and worst Value-at-Risk with monotone marginal densities”. In: *Finance and Stochastics* 17.2, pp. 395–417.
- Wu, CF Jeff and Michael S Hamada (2011). *Experiments: planning, analysis, and optimization*. John Wiley & Sons.
- Xiao, Yugu and Jing Yao (2020). “A note on joint mix random vectors”. In: *Communications in Statistics - Theory and Methods* 49.12, pp. 3063–3072. DOI: 10.1080/03610926.2019.1586937. eprint: <https://doi.org/10.1080/03610926.2019.1586937>. URL: <https://doi.org/10.1080/03610926.2019.1586937>.
- Xie, Yujia et al. (2020). “A fast proximal point method for computing exact wasserstein distance”. In: *Uncertainty in artificial intelligence*. PMLR, pp. 433–453.
- Yu, Bin (1997). “Assouad, fano, and le cam”. In: *Festschrift for Lucien Le Cam: research papers in probability and statistics*. Springer, pp. 423–435.

Checklist

- For all models and algorithms presented, check if you include:
 - A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
 - An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
 - (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
- For any theoretical claim, check if you include:
 - Statements of the full set of assumptions of all theoretical results. [Yes]
 - Complete proofs of all theoretical results. [Yes]
 - Clear explanations of any assumptions. [Yes]
- For all figures and tables that present empirical results, check if you include:
 - The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
 - All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
 - A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

- (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
- 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator if your work uses existing assets. [Yes]
 - (b) The license information of the assets, if applicable. [Not Applicable]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
 - (d) Information about consent from data providers/curators. [Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
- 5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

6 NOTATIONS

Table 3: Notation Table

Notation	Description
n	Total number of samples
d	Dimension of the outcome space
K	Number of possible treatment levels
$\mathbf{Y}_i(\cdot)$	Potential outcomes for unit i
$Y_i(k)$	Potential outcome for unit i at treatment level k
$Y_i^j(k)$	j -th dimension of potential outcome for unit i at treatment level k
$Y_i(0, 0), Y_i(0, 1), Y_i(1, 0), Y_i(1, 1)$	Potential outcomes under different binary treatment combinations
$\text{Var}_{\mu_\star}(Y_i(2) - Y_i(1))$	Variance of the treatment effect
$\text{Cov}_{\mu_\star}(Y_i^1(2) - Y_i^1(1), Y_i^2(2) - Y_i^2(1))$	Covariance between individual treatment effects across different dimensions
Γ	Collection of joint probabilities respecting the marginal distributions
μ_\star	Unknown joint distribution of potential outcomes
$\bar{\mu}_\star$	Centered counterpart of μ_\star , that is, $\bar{\mu}_\star(k)$ is μ_\star minus its expectation
$\mu_\star(k)$	Marginal distribution of $Y_i(k)$
$\hat{\mu}_n$	Empirical distribution of μ_\star with n samples
W_i	Realized treatment level for unit i
$\mathbb{E}_{\mu_\star}[\ell(\mathbf{Y}_i(\cdot))]$	Expectation of a causal estimand ℓ under the distribution μ_\star
$\text{QMW}(\cdot; A)$	Quadratic objective parameterized by A
$\text{MW}_2^2(\cdot)$	Multi-Marginal Wasserstein Loss: Average squared ℓ_2 norm of potential outcomes
β	Contrast vector for treatments with multiple levels
τ_β	Average contrast effect estimand
$\hat{\tau}_\beta$	Difference-in-means type estimator
S_k^2	Sample variance of potential outcomes at treatment level k
S_τ^2	Sample variance of the contrast treatment effect
V_β	Variance of the difference-in-means type estimator

7 RELATED WORKS

7.1 Partial-identification

In the literature of causal inference, identifiability is typically the first issue addressed when a new causal estimand is proposed. Under the potential outcome model, causal estimands depending on cross-world potential outcomes are generally not identifiable due to the inherent missingness. In this case, the range of parameter values that are compatible with the data is of interest. A multitude of works have been dedicated to finding causal estimands that can be identified, where a set of assumptions is usually necessary. For instance, the identifiability of the average treatment effect requires the assumption of no unobserved confounders, and using instrumental variables to identify the local treatment effect requires the exogeneity (Imbens and Rubin 2015). However, verifying these assumptions can be challenging or even impossible based solely on the data, and the violation of the assumptions will render the parameter unidentified.

In the literature of econometrics, the partial identification problem has been a topic of heated discussion for decades (see Kline and Tamer 2023 for a comprehensive review and references therein). The works generally begin with a model depending on the parameter of interest and possibly other nuisance parameters or functions. The partially identified sets are specified to be the values that comply with a set of moment inequalities, parameters that maximize a target function, or intervals with specifically defined endpoints. In causal inference, there is a tendency to avoid heavy model assumptions imposed on the potential outcomes to ensure the robustness and generalizability of causal conclusions. Methods therein often involve leveraging non-parametric or semi-parametric methods that make few structural assumptions about the data-generating process (Bickel et al. 1993; Laan and Robins 2003; Vaart 2000).

We detail two particularly relevant threads of works among the studies on partial identification problems. Copula models (Nelsen 2006) are used to describe the dependency between multiple random variables with known marginal distributions, which also aligns with the potential outcome model with marginals accessible and the coupling unknown. The Fréchet–Hoeffding copula bounds can be used to characterize the joint distribution of the potential outcomes (Fan and Park 2010; Heckman, Smith, and Clements 1997; Manski 1997). However, when dealing with more than two margins, one side of the Fréchet–Hoeffding theorem’s bound is only point-wise sharp; moreover, copula models are generally constrained to unidimensional random variables; in addition, the bounds on the joint distribution do not necessarily translate to those of causal estimands, such as the variance of the difference of two potential outcomes. The second thread of related works explicitly employs the optimal transport to address the partial identification problem in econometrics and causal inference. Galichon 2018 considers a model that includes observed variables and latent variables, and the parameter of interest regulates the distribution of the latent variables and the relationship between the observed and unobserved variables, which differ from our problem formulation. Balakrishnan, Kennedy, and Wasserman 2023 introduce the quadratic and differential effects for continuous treatments (infinite levels of treatment) that can be considered as examples of the lower limits of our formulation (2) with a quadratic objective function.

7.2 MOT

MOT and minimax estimation The convergence of the empirical optimal transport cost (with two margins) to its population value has been extensively studied and well-understood (see Chizat et al. 2020; Manole and Niles-Weed 2024; Staudt and Hundrieser 2023 and the references therein). However, the estimation error of the plug-in estimator, which is the empirical optimal transport cost, for general loss functions remains a long-standing open problem that is only partially understood under strong assumptions (Manole and Niles-Weed 2024; Staudt and Hundrieser 2023). Notably, it is known to be minimax optimal for the Wasserstein p -cost. However, to the best of our knowledge, such minimax results have not been established for the MOT problem. In this paper, we generalize the approach by Chizat et al. 2020 to provide a convergence result for the plug-in estimator of the MOT problem for arbitrary quadratic objective functions (Eq. (3)). Furthermore, we refine the analysis for the average squared ℓ_2 norm (Eq. (4)) and show that the plug-in estimator is minimax optimal when the dimension $d \leq 4$. Determining the minimax rates for cases where $d > 4$, and for arbitrary quadratic functions or arbitrary cost functions remains an interesting direction for future research.

Computation for MOT For computing Optimal Transport (OT) distance, the Sinkhorn algorithm (Cuturi 2013) and its accelerated variants (see Altschuler, Niles-Weed, and Rigollet 2017; Lin, Ho, and Jordan 2019; Xie

et al. 2020 and references therein) remain the state-of-the-art approach. Extensions of the Sinkhorn algorithm to the MOT problem have been proposed in the literature (Lin et al. 2022; Peyré, Cuturi, et al. 2019; Tupitsa et al. 2020). Notably, Lin et al. 2022; Tupitsa et al. 2020 provide finite time convergence guarantees for the Multi-marginal Sinkhorn and Primal-Dual Accelerated Alternating Minimization (PD-AAM) algorithms, respectively. While the implementations of OT algorithms are quite advanced (Flamary et al. 2021), those for MOT algorithms are still in the early stages. Specifically, Tupitsa et al. 2020 offers implementations that work with $K = 4$, $n = 60$ and accuracy $\epsilon = 1/80$. Since our tasks require concrete lower bounds for the MOT problem at a considerably larger scale of sample size n and higher accuracy ϵ , we have implemented several algorithms (see Section 4.1 for details), used them to obtain our numerical results, and provided an efficient off-the-shelf MOT package.

8 TECHNICAL LEMMAS

Throughout the appendix, for any $L > 0$, let $\mathcal{F}_{\text{Lip},L}$ denote the L -Lipschitz function class supported on $B_1^d(0)$ such that all functions $\phi \in \mathcal{F}_{\text{Lip},L}$ has $\phi(0) = 0$. Denote by $\gamma_A^*(\mu(1), \dots, \mu(K))$ the optimal coupling between $\mu(1), \dots, \mu(K)$ in terms of the objective (3). Denote by $\gamma^*(\mu(1), \dots, \mu(K))$ the optimal coupling between $\mu(1), \dots, \mu(K)$ in terms of the objective (4).

Lemma 2. *For any μ supported on $B_1^d(0)$, the Rademacher complexity of \mathcal{F}_L under μ is bounded by*

$$\mathcal{R}_n(\mathcal{F}_L, \mu) := \mathbb{E}_{\epsilon, X} \left[\sup_{\phi \in \mathcal{F}_L} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \phi(Y_i) \right| \right] \lesssim \begin{cases} Ln^{-2/d} & \text{if } d > 4, \\ Ln^{-1/2} \log n & \text{if } d = 4, \\ Ln^{-1/2} & \text{if } d < 4, \end{cases}$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)$, $\epsilon_1, \dots, \epsilon_n$ are independent Rademacher variables and Y_1, \dots, Y_n are independent random variables sampled from the distribution μ .

Proof of Lemma 2. We follow the derivation of Chizat et al. 2020, Lemma 4. On $B_1^d(0)$, any function $\phi \in \mathcal{F}_L$ is bounded by $\|\phi\|_\infty \lesssim L$. Thus by Wainwright 2019, Theorem 5.22, we have

$$\mathcal{R}_n(\mathcal{F}_L, \mu) \lesssim \inf_{\delta > 0} \left(\delta + n^{-1/2} \int_\delta^L \sqrt{\log \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, u)} du \right),$$

where $\mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, u)$ denotes the covering number of \mathcal{F}_L in $\|\cdot\|_\infty$ at scale u . Then by Guntuboyina and Sen 2012, Theorem 1 which obtains $\log \mathcal{N}(\mathcal{F}_L, \|\cdot\|_\infty, u) \lesssim (u/L)^{-d/2}$, we have

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}_L, \mu) &\lesssim \inf_{\delta > 0} \left(\delta + n^{-1/2} \int_\delta^L (u/L)^{-d/4} du \right) \\ &= \inf_{\delta > 0} \left(\delta + Ln^{-1/2} \int_{\delta/L}^1 u^{-d/4} du \right) \\ &\lesssim \begin{cases} Ln^{-2/d} & \text{if } d > 4, \\ Ln^{-1/2} \log n & \text{if } d = 4, \\ Ln^{-1/2} & \text{if } d < 4, \end{cases} \end{aligned}$$

where for $d < 4$, we choose $\delta = 0$ and $\delta = Ln^{-1/2}$ for $d = 4$ and $\delta = Ln^{-2/d}$ for $d > 4$. \square

Lemma 3. *For any μ supported on $B_1^d(0)$ and $\hat{\mu}_n$ an empirical distribution of n independent samples, we have*

$$\mathbb{E} \left[\sup_{\phi \in \mathcal{F}_L} \left| \int \phi(y) d(\hat{\mu}_n - \mu) \right| \right] \lesssim 2\mathcal{R}_n(\mathcal{F}_L, \mu) + Ln^{-1/2}.$$

Proof of Lemma 3. By Wainwright 2019, Theorem 4.10, we have for all $\delta > 0$, with probability at least $1 - \exp(-n\delta^2/(2L^2))$,

$$\sup_{\phi \in \mathcal{F}_L} \left| \int \phi(y) d(\hat{\mu}_n - \mu) \right| \leq 2\mathcal{R}_n(\mathcal{F}_L, \mu) + \delta.$$

Then since $\mathbb{E}[X] = \int_0^\infty \mathbf{P}(X \geq \delta) d\delta$ for all positive random variable, we have

$$\begin{aligned}
 & \mathbb{E} \left[\sup_{\phi \in \mathcal{F}_L} \left| \int \phi(y) d(\hat{\mu}_n - \mu) \right| - 2\mathcal{R}_n(\mathcal{F}_L, \mu) \right] \\
 & \leq \mathbb{E} \left[\left(\sup_{\phi \in \mathcal{F}_L} \left| \int \phi(y) d(\hat{\mu}_n - \mu) \right| - 2\mathcal{R}_n(\mathcal{F}_L, \mu) \right)_+ \right] \\
 & = \int_0^\infty \mathbf{P} \left(\left(\sup_{\phi \in \mathcal{F}_L} \left| \int \phi(y) d(\hat{\mu}_n - \mu) \right| - 2\mathcal{R}_n(\mathcal{F}_L, \mu) \right)_+ \geq \delta \right) d\delta \\
 & \leq \int_0^\infty \exp(-n\delta^2/(2L^2)) d\delta \\
 & \lesssim Ln^{-1/2}.
 \end{aligned}$$

□

9 OMITTED PROOFS FROM SECTION 2

Lemma 4. For any K distributions $\mu(1), \dots, \mu(K)$ on \mathbb{R}^d , let $\bar{\mu}(1), \dots, \bar{\mu}(K)$ be the centered counterpart, that is, $\bar{\mu}(k)$ is a translation of $\mu(k)$ by $-\mathbb{E}_{\mu(k)}[X]$ and $Y_1(k) \sim \mu(k)$ for $k \in [K]$. Then we have

$$\text{MW}_2^2(\mu(1), \dots, \mu(K)) = \text{MW}_2^2(\bar{\mu}(1), \dots, \bar{\mu}(K)) + \left\| \frac{\sum_{k=1}^K \mathbb{E}_{\mu(k)}[Y_1(k)]}{K} \right\|_2^2$$

Proof of Lemma 4. Suppose $\gamma^* = \gamma^*(\mu(1), \dots, \mu(K))$ achieves $\text{MW}_2^2(\mu(1), \dots, \mu(K))$, then we have

$$\begin{aligned}
 & \text{MW}_2^2(\mu(1), \dots, \mu(K)) \\
 & = \int_y \left\| \frac{\sum_{k=1}^K y(k)}{K} \right\|^2 \gamma^*(dy) \\
 & = \int_y \left\| \frac{\sum_{k=1}^K (y(k) - \mathbb{E}_{\mu(k)}[Y_1(k)])}{K} \right\|^2 \gamma^*(dy) + \left\| \frac{\sum_{k=1}^K \mathbb{E}_{\mu(k)}[Y_1(k)]}{K} \right\|_2^2 \\
 & \leq \text{MW}_2^2(\bar{\mu}(1), \dots, \bar{\mu}(K)) + \left\| \frac{\sum_{k=1}^K \mathbb{E}_{\mu(k)}[Y_1(k)]}{K} \right\|_2^2.
 \end{aligned}$$

The inequality can be reversed considering the optimal coupling $\bar{\gamma}^* = \gamma^*(\bar{\mu}(1), \dots, \bar{\mu}(K))$, which concludes our proof. □

Lemma 5. For any K distributions $\mu(1), \dots, \mu(K)$ on \mathbb{R}^d , let $\bar{\mu}(1), \dots, \bar{\mu}(K)$ be the centered counterpart, that is, $\bar{\mu}(k)$ is a translation of $\mu(k)$ by $-\mathbb{E}_{\mu(k)}[X]$ and $Y_1(k) \sim \mu(k)$ for $k \in [K]$. Let A be any symmetric matrix. Then we have

$$\text{QMW}(\mu(1), \dots, \mu(K); A) = \text{QMW}(\bar{\mu}(1), \dots, \bar{\mu}(K); A) + \sum_{k,l} a_{kl} \langle \mathbb{E}_{\mu(k)}[Y_1(k)], \mathbb{E}_{\mu(l)}[Y_1(l)] \rangle.$$

Proof of Lemma 5. Suppose $\gamma^* = \gamma^*(\mu(1), \dots, \mu(K); A)$ achieves $\text{QMW}(\mu(1), \dots, \mu(K); A)$, then we have

$$\begin{aligned}
 & \text{QMW}(\mu(1), \dots, \mu(K); A) \\
 &= \int_y \sum_{k,l} a_{kl} \langle y(k), y(l) \rangle \gamma^*(dy) \\
 &= \int_y \sum_{k,l} a_{kl} \langle y(k) - \mathbb{E}_{\mu(k)}[Y_1(k)], y(l) - \mathbb{E}_{\mu(l)}[Y_1(l)] \rangle \gamma^*(dy) + \sum_{k,l} a_{kl} \langle \mathbb{E}_{\mu(k)}[Y_1(k)], \mathbb{E}_{\mu(l)}[Y_1(l)] \rangle \\
 &\leq \text{QMW}(\bar{\mu}(1), \dots, \bar{\mu}(K); A) + \sum_{k,l} a_{kl} \langle \mathbb{E}_{\mu(k)}[Y_1(k)], \mathbb{E}_{\mu(l)}[Y_1(l)] \rangle.
 \end{aligned}$$

The inequality can be reversed considering the optimal coupling $\bar{\gamma}^* = \gamma^*(\bar{\mu}(1), \dots, \bar{\mu}(K))$, which concludes our proof. \square

10 OMITTED PROOFS FROM SECTION 3

10.1 Upper Bound

Lemma 6. *For any matrix $A = (a_{ij})_{1 \leq i, j \leq K}$, we have*

$$\mathbb{E} \left[\left| \sum_{k,l} a_{kl} \langle \mathbb{E}_{\mu_*(k)} Y_1(k), \mathbb{E}_{\mu_*(l)} Y_1(l) \rangle - \sum_{k,l} a_{kl} \langle \mathbb{E}_{\hat{\mu}_n(k)} Y_1(k), \mathbb{E}_{\hat{\mu}_n(l)} Y_1(l) \rangle \right| \right] \lesssim K \sqrt{\sum_{k,l} a_{kl}^2} \cdot n^{-1/2}.$$

Proof of Lemma 6. We first separate the quantity

$$\begin{aligned}
 & \sum_{k,l} a_{kl} \langle \mathbb{E}_{\mu_*(k)} Y_1(k), \mathbb{E}_{\mu_*(l)} Y_1(l) \rangle - \sum_{k,l} a_{kl} \langle \mathbb{E}_{\hat{\mu}_n(k)} Y_1(k), \mathbb{E}_{\hat{\mu}_n(l)} Y_1(l) \rangle \\
 &= \sum_{k=1}^K a_{kk} (\langle \mathbb{E}_{\mu_*(k)} [Y_1(k)], \mathbb{E}_{\mu_*(k)} [Y_1(k)] \rangle - \langle \mathbb{E}_{\hat{\mu}_n(k)} [Y_1(k)], \mathbb{E}_{\hat{\mu}_n(k)} [Y_1(k)] \rangle) \\
 &\quad + \sum_{k \neq l} a_{kl} \langle \mathbb{E}_{\mu_*(k)} [Y_1(k)], \mathbb{E}_{\mu_*(l)} [Y_1(l)] \rangle \\
 &\quad - \sum_{k \neq l} a_{kl} \langle \mathbb{E}_{\hat{\mu}_n(k)} [Y_1(k)], \mathbb{E}_{\hat{\mu}_n(l)} [Y_1(l)] \rangle \\
 &= \sum_{k=1}^K a_{kk} \langle \mathbb{E}_{\mu_*(k)} Y_1(k) - \mathbb{E}_{\hat{\mu}_n(k)} Y_1(k), \mathbb{E}_{\mu_*(k)} Y_1(k) + \mathbb{E}_{\hat{\mu}_n(k)} Y_1(k) \rangle \\
 &\quad + \sum_{k \neq l} a_{kl} \langle \mathbb{E}_{\mu_*(k)} [Y_1(k)], \mathbb{E}_{\mu_*(l) - \hat{\mu}_n(l)} [Y_1(l)] \rangle \\
 &\quad + \sum_{k \neq l} a_{kl} \langle \mathbb{E}_{\mu_*(k) - \hat{\mu}_n(k)} [Y_1(k)], \mathbb{E}_{\hat{\mu}_n(l)} [Y_1(l)] \rangle.
 \end{aligned}$$

Then, by triangle inequality together with Khintchine's inequality (Exercise 2.6.7 of Vershynin 2018), we have

$$\begin{aligned}
 & \mathbb{E} \left[\left| \sum_{k,l} a_{kl} \langle \mathbb{E}_{\mu_\star(k)} Y_1(k), \mathbb{E}_{\mu_\star(l)} Y_1(l) \rangle - \sum_{k,l} a_{kl} \langle \mathbb{E}_{\hat{\mu}_n(k)} Y_1(k), \mathbb{E}_{\hat{\mu}_n(l)} Y_1(l) \rangle \right| \right] \\
 & \leq \mathbb{E} \left[\left| \sum_{k=1}^K a_{kk} \langle \mathbb{E}_{\mu_\star(k)} Y_1(k) - \mathbb{E}_{\hat{\mu}_n(k)} Y_1(k), \mathbb{E}_{\mu_\star(k)} Y_1(k) + \mathbb{E}_{\hat{\mu}_n(k)} Y_1(k) \rangle \right| \right] \\
 & \quad + \sum_{k=1}^K \mathbb{E} \left[\left| \mathbb{E}_{\mu_\star(k)} \left| \mathbb{E}_{(\mu_\star - \hat{\mu}_n)(-k)} \left[\sum_{l \neq k} a_{kl} \langle Y_1(k), Y_1(l) \rangle \right] \right| \right| \right] \\
 & \quad + \sum_{k=1}^K \mathbb{E} \left[\left| \mathbb{E}_{\hat{\mu}_n(k)} \left| \mathbb{E}_{(\mu_\star - \hat{\mu}_n)(-k)} \left[\sum_{l \neq k} a_{kl} \langle Y_1(k), Y_1(l) \rangle \right] \right| \right| \right] \\
 & \lesssim \left(\sqrt{\sum_{k=1}^K a_{kk}^2} + \sum_{k=1}^K \sqrt{\sum_{l \neq k} a_{kl}^2} \right) (n/K)^{-1/2} \\
 & \lesssim \sqrt{K \cdot \sum_{k,l} a_{kl}^2} \cdot (n/K)^{-1/2},
 \end{aligned}$$

where the last inequality is by Cauchy-Schwartz. \square

Proposition 3. For any symmetric matrix $A \in \mathbb{R}^{K \times K}$, we have

$$\begin{aligned}
 & \mathbb{E} [| \text{QMW}(\hat{\mu}_n(1), \dots, \hat{\mu}_n(K); A) - \text{QMW}(\mu_\star(1), \dots, \mu_\star(K); A) |] \\
 & \lesssim K \sqrt{\sum_{k=1}^K a_{kk}^2 \cdot n^{-1/2} + \left(\sum_{k=1}^K |a_{kk}| \right)} \wedge \begin{cases} \eta_{K,d,A} \cdot K^{3/2} (n/K)^{-2/d}, & \text{if } d > 4, \\ \eta_{K,d,A} \cdot K^2 n^{-1/2} \log(n/K), & \text{if } d = 4, \\ \eta_{K,d,A} \cdot K^2 n^{-1/2}, & \text{if } d < 4, \end{cases} \quad (7)
 \end{aligned}$$

where the notation \lesssim hides constants that only depend on the dimension d . Moreover, the coefficient $\eta_{K,d,A}$ is defined as

$$\eta_{K,d,A} := \sup_{\substack{\mu(1), \dots, \mu(K) \\ \text{with zero mean}}} \sup_{\substack{(y(1), \dots, y(K)) \in \\ \text{supp}(\gamma_A^\star(\mu(1), \dots, \mu(K)))}} \sup_{k \in [K]} \left\| \sum_{l \neq k} \frac{a_{kl} y(l)}{\sqrt{K}} \right\|,$$

where $\mu(1), \dots, \mu(K)$ are d -dimensional distributions with zero mean and $(y(1), \dots, y(K))$ in the support of the coupling $\gamma_A^\star(\mu(1), \dots, \mu(K))$.

Proof of Proposition 3. By lemmas 5 and 6, we know that the objective can be separated and the mean of the coupling concentrates. Thus, without loss of generality, in this proof, we only consider the marginals $\mu_\star(1), \dots, \mu_\star(K), \hat{\mu}_n(1), \dots, \hat{\mu}_n(K)$ that have zero means. For any K marginals $\mu(1), \dots, \mu(K)$ supported on $B_1^d(0) \subset \mathbb{R}^d$, we have

$$\begin{aligned}
 & \text{QMW}(\mu(1), \dots, \mu(K); A) \\
 & := \inf_{\gamma \in \Gamma(\mu(1), \dots, \mu(K))} \int \sum_{i,j} a_{ij} y(i)^\top y(j) \gamma(dy) \\
 & \leq \sum_{k=1}^K a_{kk} \int \|y(k)\|^2 \mu(k, dy(k)) + \sum_{i \neq j} a_{ij} \int y(i)^\top y(j) \mu(i, dy(i)) \mu(j, dy(j)) \\
 & = \sum_{k=1}^K a_{kk} \int \|y(k)\|^2 \mu(k, dy(k)) \leq \sum_{k=1}^K |a_{kk}|,
 \end{aligned}$$

where the first inequality is by choosing the independent coupling $\gamma(dy) = \prod_{k=1}^K \mu(k, dy(k))$ and the last inequality is by the boundedness of y . This implies

$$|\text{QMW}(\mu_\star(1), \dots, \mu_\star(K); A) - \text{QMW}(\hat{\mu}_n(1), \dots, \hat{\mu}_n(K); A)| \leq \sum_{k=1}^K |a_{kk}|. \quad (8)$$

Now by the classic duality of Kellerer 1984, we have

$$\text{QMW}(\mu(1), \dots, \mu(K); A) = \sup_{\substack{f_1, \dots, f_K: \\ \sum_{k=1}^K f_k(y(k)) \leq y^\top A y}} \sum_{k=1}^K \int f_k(y(k)) \mu(k, dy(k)),$$

where $\mu(k, dy(k))$ denote the marginal distribution of $\mu(k)$ on $y(k)$. We first separate the square terms out, since they only concern one marginal as

$$\begin{aligned} & \sup_{\substack{f_1, \dots, f_K: \\ \sum_{k=1}^K f_k(y(k)) \leq y^\top A y}} \sum_{k=1}^K \int f_k(y(k)) \mu(k, dy(k)) \\ &= \sum_{k=1}^K \int a_{kk} \|y(k)\|^2 \mu(k, dy(k)) + \sup_{\substack{f_1, \dots, f_K: \\ \sum_{k=1}^K f_k(y(k)) \leq y^\top A y}} \sum_{k=1}^K \int f_k(y(k)) - a_{kk} \|y(k)\|^2 \mu(k, dy(k)) \\ &= \sum_{k=1}^K \int a_{kk} \|y(k)\|^2 \mu(k, dy(k)) + \sup_{\substack{\phi_1, \dots, \phi_K: \\ \sum_{k=1}^K \phi_k(y(k)) \leq \sum_{i \neq j} a_{ij} y^\top(i) y(j)}} \sum_{k=1}^K \int \phi_k(y(k)) \mu(k, dy(k)), \end{aligned} \quad (9)$$

where the second equality is by setting $\phi_k(y(k)) = f_k(y(k)) - a_{kk} \|y(k)\|^2$. Now we are going to show that there exists a sequence of supremum achieving functions $\phi_1^\star, \dots, \phi_K^\star$ that all lie in the space of $\mathcal{F}_{\text{Lip}, 2K}$. To start with, we note again by the classical duality of Kellerer 1984 that there exists $\tilde{\phi}_1, \dots, \tilde{\phi}_K$ that achieves the supremum in eq. (9). For $k = 1, \dots, K$, we iteratively define the function ϕ_k^\star as

$$\phi_k^\star(y) := \inf_{\substack{y(1), \dots, y(K) \\ \in \text{supp}(\gamma_A^\star(\mu(1), \dots, \mu(K)))}} 2 \left\langle y, \sum_{l \neq k} a_{kl} y(l) \right\rangle + \sum_{\substack{i, j: i \neq j, \\ i \neq k, j \neq k}} a_{ij} \langle y(i), y(j) \rangle - \sum_{l=1}^{k-1} \phi_l^\star(y(l)) - \sum_{l=k+1}^K \tilde{\phi}_l(y(l)).$$

And prove by induction that $\phi_1^\star, \dots, \phi_k^\star, \tilde{\phi}_{k+1}, \dots, \tilde{\phi}_K$ are admissible and achieve the supremum in eq. (9). When $k = 0$, the claim is true by definition. Now suppose the claim is true up to $k - 1$. For any k , by definition, we have for any $y(1), \dots, y(K)$

$$\sum_{l=1}^k \phi_l^\star(y(l)) + \sum_{l=k+1}^K \tilde{\phi}_l(y(l)) \leq \sum_{i \neq j} a_{ij} y(i) y(j).$$

Thus $\phi_1^\star, \dots, \phi_k^\star, \tilde{\phi}_{k+1}, \dots, \tilde{\phi}_K$ are admissible function for the supremum in eq. (9). Furthermore, by the induction hypothesis, $\phi_1^\star, \dots, \phi_{k-1}^\star, \tilde{\phi}_k, \dots, \tilde{\phi}_K$ are admissible function and achieves the supremum in eq. (9), we have for any $y(1), \dots, y(k)$

$$\sum_{l=1}^{k-1} \phi_l^\star(y(l)) + \sum_{l=k}^K \tilde{\phi}_l(y(l)) \leq \sum_{i \neq j} a_{ij} y(i) y(j).$$

This implies

$$2 \left\langle y(k), \sum_{l \neq k} a_{kl} y(l) \right\rangle + \sum_{\substack{i, j: i \neq j, \\ i \neq k, j \neq k}} a_{ij} \langle y(i), y(j) \rangle - \sum_{l=1}^{k-1} \phi_l^\star(y(l)) - \sum_{l=k+1}^K \tilde{\phi}_l(y(l)) \geq \tilde{\phi}_k(y(k)).$$

Then take infimum over $y(1), \dots, y(k-1), y(k+1), \dots, y(K)$, we have

$$\phi_k^*(y(k)) \geq \tilde{\phi}_k(y(k)).$$

This implies that $\phi_1^*, \dots, \phi_k^*, \tilde{\phi}_{k+1}, \dots, \tilde{\phi}_K$ achieves the supremum in eq. (9). Thus by induction, we have shown that $\phi_1^*, \dots, \phi_K^*$ are admissible function and achieves the supremum in eq. (9). Furthermore, since for any k , ϕ_k are infimum of linear functions, the Lipschitz constant for ϕ_k is thus upper bounded by

$$\sup_{\substack{y(1), \dots, y(k-1), \\ y(k+1), \dots, y(K) \\ \in \text{supp}(\gamma_A^*(\mu(1), \dots, \mu(K)))}} 2 \left\| \sum_{l \neq k} a_{kl} y(l) \right\| \leq 2\eta_{K,d,A} \sqrt{K}.$$

Therefore,

$$\begin{aligned} & \text{QMW}(\hat{\mu}_n(1), \dots, \hat{\mu}_n(K); A) - \text{QMW}(\mu_*(1), \dots, \mu_*(K); A) \\ &= \sum_{k=1}^K \int a_{kk} \|y(k)\|^2 (\hat{\mu}_n(k, dy(k)) - \mu_*(k, dy(k))) \\ & \quad + \sup_{\substack{\phi_1, \dots, \phi_K: \\ \sum_{k=1}^K \phi_k(y(k)) \leq \sum_{i \neq j} a_{ij} y^\top(i) y(j) \\ \phi_k: 2\sqrt{K}\eta_{K,d,A}\text{-Lip.}}} \sum_{k=1}^K \int \phi_k(y(k)) \hat{\mu}_n(k, dy(k)) \\ & \quad - \sup_{\substack{\phi_1, \dots, \phi_K: \\ \sum_{k=1}^K \phi_k(y(k)) \leq \sum_{i \neq j} a_{ij} y^\top(i) y(j) \\ \phi_k: 2\sqrt{K}\eta_{K,d,A}\text{-Lip.}}} \sum_{k=1}^K \int \phi_k(y(k)) \mu_*(k, dy(k)) \\ & \leq \sum_{k=1}^K \int a_{kk} \|y(k)\|^2 (\hat{\mu}_n(k, dy(k)) - \mu_*(k, dy(k))) \\ & \quad + \sup_{\substack{\phi_1, \dots, \phi_K: \\ \sum_{k=1}^K \phi_k(y(k)) \leq \sum_{i \neq j} a_{ij} y^\top(i) y(j) \\ \phi_k: 2\sqrt{K}\eta_{K,d,A}\text{-Lip.}}} \sum_{k=1}^K \int \phi_k(y(k)) (\hat{\mu}_n(k, dy(k)) - \mu_*(k, dy(k))) \\ & \leq \left| \sum_{k=1}^K \int a_{kk} \|y(k)\|^2 (\hat{\mu}_n(k, dy(k)) - \mu_*(k, dy(k))) \right| \\ & \quad + \sum_{k=1}^K \sup_{\phi_k \in \mathcal{F}_{\text{Lip}, 2\sqrt{K}\eta_{K,d,A}}} \int \phi_k(y(k)) (\hat{\mu}_n(k, dy(k)) - \mu_*(k, dy(k))). \end{aligned}$$

Similarly we can show $-(\text{QMW}(\hat{\mu}_n(1), \dots, \hat{\mu}_n(K); A) - \text{QMW}(\mu_*(1), \dots, \mu_*(K); A))$ satisfies the same upper bound. By Hoeffding's inequality (Proposition 2.5 of Wainwright 2019), we have

$$\mathbb{E} \left| \sum_{k=1}^K \int a_{kk} \|y(k)\|^2 (\hat{\mu}_n(k, dy(k)) - \mu_*(k, dy(k))) \right| \lesssim \left(\sum_{k=1}^K |a_{kk}| \right) \cdot (n/K)^{-1/2}. \quad (10)$$

Finally, by Lemma 3, we obtain

$$\mathbb{E} \left[\sum_{k=1}^K \sup_{\phi_k \in \mathcal{F}_{\text{Lip}, 2\sqrt{K}\eta_{K,d,A}}} \int \phi_k(y(k)) (\hat{\mu}_n(k, dy(k)) - \mu_*(k, dy(k))) \right] \lesssim K^{3/2} \eta_{K,d,A} \cdot \begin{cases} (n/K)^{-2/d}, & \text{if } d > 4, \\ (n/K)^{-1/2} \log(n/K), & \text{if } d = 4, \\ (n/K)^{-1/2}, & \text{if } d < 4, \end{cases} \quad (11)$$

The bound from eq. (10) is lower order compared to the bound of the mean concentration from lemma 6. In all, combine Lemma 6, eq. (8), and eq. (11), we have

$$\begin{aligned} & \mathbb{E} [|\text{QMW}(\hat{\mu}_n(1), \dots, \hat{\mu}_n(K); A) - \text{QMW}(\mu_*(1), \dots, \mu_*(K); A)|] \\ & \lesssim K \sqrt{\sum_{k=1}^K a_{kk}^2} \cdot n^{-1/2} + \left(\sum_{k=1}^K |a_{kk}| \right) \wedge \begin{cases} \eta_{K,d,A} \cdot K^{3/2} (n/K)^{-2/d}, & \text{if } d > 4, \\ \eta_{K,d,A} \cdot K^2 n^{-1/2} \log(n/K), & \text{if } d = 4, \\ \eta_{K,d,A} \cdot K^2 n^{-1/2}, & \text{if } d < 4, \end{cases} \end{aligned}$$

□

Lemma 7. For $n, K \geq 2$ and $d \geq 1$, for any K mean zero marginals $\mu(1), \dots, \mu(K)$ finitely supported on $B_1^d(0) \subset \mathbb{R}^d$, we have

$$\sup_{\substack{(y(1), \dots, y(K)) \in \\ \text{supp}(\gamma^*(\mu(1), \dots, \mu(K)))}} \sup_{l \in [K]} \left\| \sum_{k \neq l} \frac{y(k)}{\sqrt{K}} \right\| \leq 4.$$

$(y(1), \dots, y(K))$ in the support of the optimal coupling $\gamma^*(\mu(1), \dots, \mu(K))$ with respect to the objective in (4).

Proof of Lemma 7. Denote the set $\mathcal{A}_1 = \left\{ y : \left\| \sum_{k=1}^K y(k) \right\| \leq 2\sqrt{K} \right\}$ and $\mathcal{A}_2 = \left\{ y : \left\| \sum_{k=1}^K y(k) \right\| \geq 4\sqrt{K} - 1 \right\}$. We prove by contradiction. Suppose otherwise, then we have $\gamma^*(\mathcal{A}_2) > 0$. On the other hand, we note that the objective value is upper bounded by

$$\begin{aligned} \text{MW}_2^2(\mu(1), \dots, \mu(K)) &= \int \left\| \frac{\sum_{k=1}^K y(k)}{K} \right\|_2^2 \gamma^*(dy) \\ &\leq \int \left\| \frac{\sum_{k=1}^K y(k)}{K} \right\|_2^2 \prod_{k=1}^K \mu(k, dy(k)), \\ &= \frac{1}{K^2} \sum_{k=1}^K \int \|y(k)\|^2 \mu(k, dy(k)) \leq \frac{1}{K}. \end{aligned}$$

Then by Markov's inequality, we have $\gamma^*(\mathcal{A}_1) > 3/4$. Since all the marginals are finitely supported, we can find $y_1 \in \mathcal{A}_1$ and $y_2 \in \mathcal{A}_2$ such that $\gamma^*(y_1), \gamma^*(y_2) > 0$. Let $p = \gamma^*(y_1) \wedge \gamma^*(y_2)$. For any $j \in [K]$, we consider $y_{1,j} = (y_1(1), \dots, y_1(j-1), y_2(j), y_1(j+1), \dots, y_1(K))$ and $y_{2,j} = (y_2(1), \dots, y_2(j-1), y_1(j), y_2(j+1), \dots, y_2(K))$. Then since the K couplings $\gamma^* - p(\delta_{y_1} + \delta_{y_2}) + p(\delta_{y_{1,j}} + \delta_{y_{2,j}})$ for $j \in [K]$ are no better than the optimal coupling in the objective value, we thus have for $j \in [K]$,

$$\left\| \sum_{k=1}^K y_1(k) \right\|^2 + \left\| \sum_{k=1}^K y_2(k) \right\|^2 \leq \left\| y_2(j) + \sum_{k \neq j} y_1(k) \right\|^2 + \left\| y_1(j) + \sum_{k \neq j} y_2(k) \right\|^2.$$

Simplifying the expression, we obtain for any $j \in [K]$,

$$\left\langle y_1(j) - y_2(j), \sum_{k \neq j} (y_1(k) - y_2(k)) \right\rangle \leq 0.$$

This further implies

$$\left\langle y_1(j) - y_2(j), \sum_k (y_1(k) - y_2(k)) \right\rangle \leq \|y_1(j) - y_2(j)\|^2.$$

Summing up over $j \in [K]$, we have

$$\left\| \sum_k (y_1(k) - y_2(k)) \right\|^2 \leq \sum_{k=1}^K \|y_1(k) - y_2(k)\|^2 \leq 2K.$$

However, we have

$$4\sqrt{K} - 1 \leq \left\| \sum_{k=1}^K y_2(k) \right\| \leq \left\| \sum_{k=1}^K y_1(k) \right\| + \left\| \sum_{k=1}^K (y_1(k) - y_2(k)) \right\| \leq (2 + \sqrt{2})\sqrt{K}.$$

This is a contradiction. Thus have $\gamma^*(\mathcal{A}_2) = 0$, which means

$$\sup_{\substack{(y(1), \dots, y(K)) \in \\ \text{supp}(\gamma^*(\mu(1), \dots, \mu(K)))}} \sup_{k \in [K]} \left\| \sum_{l \neq k} \frac{y(l)}{\sqrt{K}} \right\| \leq 4.$$

□

Lemma 8 (Theorem 7.7 of Polyanskiy and Wu 2022+; Strassen 1965). *Suppose P and Q are two distributions on the space \mathcal{X} . Provided that the diagonal $\{(x, x) : x \in \mathcal{X}\}$ is measurable.*

$$\text{TV}(P, Q) = \min_{P_{X,Y}} \{P_{X,Y}[X \neq Y] : P_X = P, P_Y = Q\}$$

where minimization is over the coupling distributions $P_{X,Y}$ with that $P_X = P$ and $P_Y = Q$.

Proof of Theorem 1. By applying proposition 3 with $A = \mathbf{1}\mathbf{1}^\top / K^2$, we first obtain the following convergence rate: For any n, d, K , we have

$$\begin{aligned} \mathbb{E} [|\text{MW}_2^2(\hat{\mu}_n(1), \dots, \hat{\mu}_n(K)) - \text{MW}_2^2(\mu_\star(1), \dots, \mu_\star(K))|] \\ \lesssim n^{-1/2} + (1/K) \wedge \begin{cases} \eta_{K,d} \cdot K^{-1/2} (n/K)^{-2/d}, & \text{if } d > 4, \\ \eta_{K,d} \cdot n^{-1/2} \log(n/K), & \text{if } d = 4, \\ \eta_{K,d} \cdot n^{-1/2}, & \text{if } d < 4, \end{cases} \end{aligned} \quad (12)$$

where the notation \lesssim hides constants that only depend on the dimension d . Moreover, the coefficient $\eta_{K,d}$ is defined as

$$\eta_{K,d} := \sup_{\substack{\mu(1), \dots, \mu(K) \\ \text{with zero mean}}} \sup_{\substack{(y(1), \dots, y(K)) \in \\ \text{supp}(\gamma^*(\mu(1), \dots, \mu(K)))}} \sup_{l \in [K]} \left\| \sum_{k \neq l} \frac{y(k)}{\sqrt{K}} \right\|,$$

where $\mu(1), \dots, \mu(K)$ are d -dimensional distributions with zero mean and $(y(1), \dots, y(K))$ in the support of the optimal coupling $\gamma^*(\mu(1), \dots, \mu(K))$ with respect to the objective in (4). Moreover, since the support is bounded, by lemma 7, we have, $\eta_{K,d} \leq \frac{K-1}{\sqrt{K}} \leq \sqrt{K}$. Let $N \geq n$ be an integer to be specified later. Let

$\hat{\mu}_N = (\hat{\mu}_N(1), \dots, \hat{\mu}_N(K))$ be the empirical distribution sampled from μ_\star where each margin has N/K samples. Then eq. (12) implies

$$\begin{aligned} & \mathbb{E} [|\text{MW}_2^2(\hat{\mu}_N(1), \dots, \hat{\mu}_N(K)) - \text{MW}_2^2(\mu_\star(1), \dots, \mu_\star(K))|] \\ & \lesssim N^{-1/2} + (1/K) \wedge \begin{cases} (N/K)^{-2/d}, & \text{if } d > 4, \\ (N/K)^{-1/2} \log(N/K), & \text{if } d = 4, \\ (N/K)^{-1/2}, & \text{if } d < 4. \end{cases} \end{aligned}$$

Let $\bar{\mu}_n = (\bar{\mu}_n(1), \dots, \bar{\mu}_n(K))$. The distribution $\bar{\mu}_n$ is the empirical distribution sampled from $\hat{\mu}_N$ with replacement where each margin has n/K samples. Since $\hat{\mu}_N$ and $\bar{\mu}_n$ are both discretely supported, then rerun the proof of proposition 3 with lemma 7 gives the following convergence rate due to improved Lipschitz constant

$$\begin{aligned} & \mathbb{E}_{\hat{\mu}_N, \bar{\mu}_n} [|\text{MW}_2^2(\bar{\mu}_n(1), \dots, \bar{\mu}_n(K)) - \text{MW}_2^2(\hat{\mu}_N(1), \dots, \hat{\mu}_N(K))|] \\ & \lesssim n^{-1/2} + (1/K) \wedge \begin{cases} K^{-1/2}(n/K)^{-2/d}, & \text{if } d > 4, \\ n^{-1/2} \log(n/K), & \text{if } d = 4, \\ n^{-1/2}, & \text{if } d < 4, \end{cases} \end{aligned}$$

Finally, we consider the variables $\tilde{\mu}_n = (\tilde{\mu}_n(1), \dots, \tilde{\mu}_n(K))$, which is the empirical distribution sampled from $\hat{\mu}_N$ without replacement where each margin has n/K samples. We couple $\tilde{\mu}_n$ and $\bar{\mu}_n$ conditioned on $\hat{\mu}_N$ by the TV distance achieving coupling $P_{\tilde{\mu}_n, \bar{\mu}_n}$ as defined in lemma 8. Thus in all, we have

$$\begin{aligned} & \mathbb{E}_{\hat{\mu}_N} [|\text{MW}_2^2(\hat{\mu}_N(1), \dots, \hat{\mu}_N(K)) - \text{MW}_2^2(\mu_\star(1), \dots, \mu_\star(K))|] \\ & = \mathbb{E}_{\tilde{\mu}_n} [|\text{MW}_2^2(\tilde{\mu}_n(1), \dots, \tilde{\mu}_n(K)) - \text{MW}_2^2(\mu_\star(1), \dots, \mu_\star(K))|] \\ & \leq \mathbb{E}_{\hat{\mu}_N, \tilde{\mu}_n, \bar{\mu}_n} [|\text{MW}_2^2(\hat{\mu}_N(1), \dots, \hat{\mu}_N(K)) - \text{MW}_2^2(\mu_\star(1), \dots, \mu_\star(K))|] \\ & \quad + \mathbb{E}_{\hat{\mu}_N, \tilde{\mu}_n, \bar{\mu}_n} [|\text{MW}_2^2(\hat{\mu}_N(1), \dots, \hat{\mu}_N(K)) - \text{MW}_2^2(\bar{\mu}_n(1), \dots, \bar{\mu}_n(K))|] \\ & \quad + \mathbb{E}_{\hat{\mu}_N, \tilde{\mu}_n, \bar{\mu}_n} [|\text{MW}_2^2(\bar{\mu}_n(1), \dots, \bar{\mu}_n(K)) - \text{MW}_2^2(\tilde{\mu}_n(1), \dots, \tilde{\mu}_n(K))|] \\ & \lesssim \text{TV}(\bar{\mu}_n, \tilde{\mu}_n) + N^{-1/2} + n^{-1/2} + \begin{cases} (N/K)^{-2/d} + K^{-1/2}(n/K)^{-2/d}, & \text{if } d > 4, \\ (N/K)^{-1/2} \log(N/K) + n^{-1/2} \log(n/K), & \text{if } d = 4, \\ (N/K)^{-1/2} + n^{-1/2}, & \text{if } d < 4, \end{cases} \end{aligned}$$

where the first equality is due to the fact that $\hat{\mu}_n$ and $\tilde{\mu}_n$ is equal in distribution, the second inequality is by triangle inequality and the third inequality is by combing the aforementioned rates. Now we finally show $\text{TV}(\bar{\mu}_n, \tilde{\mu}_n)$ is $o(1)$ as $N \rightarrow \infty$ and then take $N \rightarrow \infty$ to obtain the desired result. Concretely, let event $\mathcal{E}_{N,n,K} = \{\text{no points in } \hat{\mu}_N \text{ are sampled twice when sampling } \bar{\mu}_n \text{ with replacement}\}$. Then we have $\bar{\mu}_n | \mathcal{E}_{N,n,K} = \tilde{\mu}_n$. Thus,

$$\text{TV}(\bar{\mu}_n, \tilde{\mu}_n) \leq 1 - \mathbf{P}_{\bar{\mu}_n}(\mathcal{E}_{N,n,K}) = o(1), \quad \text{as } N \rightarrow \infty.$$

This concludes our proof. □

Proof of Proposition 1. This is a direct corollary of proposition 3, by the upper bound of

$$\eta_{K,d,A} \lesssim \|A\|_{1,\infty} / \sqrt{K}.$$

□

10.2 Lower Bound

Proof of Proposition 2. We only need to prove for $d = 1$. For any $K \geq 2$, if K is even, then consider $\mu(1) = \dots = \mu(K) = \text{Ber}(1/2)$ and $\mu'(1) = \dots = \mu'(K) = \text{Ber}(1/2 + \varepsilon)$ with some $\varepsilon > 0$ to be decided later. If K is

odd, then set the $\mu(K) = \mu'(K)$ to be the point mass distribution on 0 and we reduce back to the even case. We first show that $\text{MW}_2^2(\mu(1), \dots, \mu(K)) = 1/4$ and $\text{MW}_2^2(\mu'(1), \dots, \mu'(K)) \geq 1/4 + \varepsilon$. Indeed, by Lemma 4, we have

$$\text{MW}_2^2(\mu(1), \dots, \mu(K)) \geq \left(\frac{\sum_{k=1}^K \mathbb{E}_{\mu(k)}[X]}{K} \right)^2 = 1/4$$

and

$$\text{MW}_2^2(\mu'(1), \dots, \mu'(K)) \geq \left(\frac{\sum_{k=1}^K \mathbb{E}_{\mu(k)}[X]}{K} \right)^2 \geq 1/4 + \varepsilon.$$

Meanwhile, for $\mu(1), \dots, \mu(K)$ the coupling γ that assigns $Y(1) = Y(3) = \dots = Y(K-1)$ and $Y(2) = Y(4) = \dots = Y(K) = 1 - Y(1)$ is an admissible coupling that achieves

$$\int \left(\sum_{k=1}^K y(K)/K \right)^2 \gamma(dy) = 1/4.$$

Thus we have shown $\text{MW}_2^2(\mu(1), \dots, \mu(K)) = 1/4$ and $\text{MW}_2^2(\mu'(1), \dots, \mu'(K)) \geq 1/4 + \varepsilon$. Let \mathbf{P} be the distribution on the observations $Y_i(1), \dots, Y_i(K)$ with $\mu_\star(1), \dots, \mu_\star(K) = \mu(1), \dots, \mu(K)$ and \mathbf{P}' be the distribution on the observations $Y_i(1), \dots, Y_i(K)$ with $\mu_\star(1), \dots, \mu_\star(K) = \mu'(1), \dots, \mu'(K)$. Then we have

$$\text{TV}(\mathbf{P}, \mathbf{P}') \lesssim \sqrt{\text{KL}(\mathbf{P}, \mathbf{P}')} \lesssim \sqrt{n \cdot \text{KL}(\text{Ber}(1/2), \text{Ber}(1/2 + \varepsilon))} \lesssim \sqrt{n\varepsilon^2}.$$

Thus by choosing $\varepsilon = c \cdot n^{-1/2}$ for constant c small enough, we can make $\text{TV}(\mathbf{P}, \mathbf{P}') \leq 1/2$ while $\text{MW}_2^2(\mu'(1), \dots, \mu'(K)) - \text{MW}_2^2(\mu(1), \dots, \mu(K)) \gtrsim \Omega(n^{-1/2})$. Then we apply Le Cam's two-point method (Lemma 1 of Yu 1997) and obtain a lower bound of $n^{-1/2}$. Furthermore, since our problem has K margins, it is easy to embed the 1-Wasserstein distance estimation problem in the first two margins. Thus, by the lower bound of Wasserstein distance estimation from theorem 22 of Manole and Niles-Weed 2024, we obtain a lower bound of $K^{-2}(n/K)^{-2/d}$ when $d > 4$. \square

11 ALGORITHMS FROM SECTION 4

In this section, we introduce the MOT Greenkhorn algorithm and MOT Sinkhorn algorithm that we implemented. We then specify the following three properties we use regarding the MOT Sinkhorn algorithm: (1) The convergence guarantee. (2) The application of log-exp-sum trick. (3) The lower bounds without convergence.

For any γ a K -marginal tensor, define

$$\text{MTV}(\gamma, \mu(1), \dots, \mu(K)) := \sum_{k=1}^K \int_{y(k)} |\gamma(k, dy(k)) - \mu(k, dy(k))|,$$

where $\gamma(k, dy(k)) = \int_{y(1), \dots, y(k-1), y(k+1), \dots, y(K)} \gamma(dy(1), \dots, dy(K))$ which might be a signed measure in general.

Multi-marginal Greenkhorn We follow Altschuler, Niles-Weed, and Rigollet 2017 and introduce the algorithm of MOT Greenkhorn. The function $\rho(a, b)$ in 4 of Algorithm 1 is defined as $\rho(a, b) = b - a + a \log(a/b)$.

Algorithm 1 Multi-marginal Greenkhorn

Require: Cost matrix C , Multiple marignals $\mu(1), \mu(2), \dots, \mu(K)$, Accuracy $\varepsilon > 0$.

- 1: $\eta \leftarrow 4 \sum_{k=1}^K \log n_k / \varepsilon$, $\varepsilon' \leftarrow \frac{\varepsilon}{8\|C\|_\infty}$, $A \leftarrow \exp(-\eta C)$.
 - 2: $m(k) \leftarrow \mathbf{0} \in \mathbb{R}^{n_k}$ for $k \in [K]$, and $\gamma \leftarrow A/\|A\|_1$.
 - 3: **while** $\text{MTV}(\gamma, \mu(1), \dots, \mu(K)) > \varepsilon'$ **do**
 - 4: $(k^*, i^*) \leftarrow \arg\max_{k \in [K], i \in [n_k]} \rho(\mu(k, i), \gamma(k, i))$. $\triangleright \mu(k, i)$ is $1/n_k$ in our case.
 - 5: $m(k^*, i^*) \leftarrow m(k^*, i^*) + \log \mu(k^*, i^*) - \log \gamma(k^*, i^*)$.
 - 6: $\gamma(i_1, \dots, i_K) = A(i_1, \dots, i_K) \cdot \prod_{k=1}^K \exp(m(k, i_k))$ for all $i_1, \dots, i_K \in [n_1] \times \dots \times [n_K]$.
 - 7: **end while**
 - 8: Let $\hat{\gamma}^{(0)} = \gamma$.
 - 9: **for** $k = 1, \dots, K$ **do**
 - 10: $v(k) \leftarrow (\mu(k)/\hat{\gamma}^{(k-1)}(k)) \wedge \mathbf{1}$, where the division and \wedge are elementwise operation.
 - 11: $\hat{\gamma}^{(k)}(i_1, \dots, i_K) = \hat{\gamma}^{(k-1)}(i_1, \dots, i_K) \cdot v(k, i_k)$ for all $i_1, \dots, i_K \in [n_1] \times \dots \times [n_K]$.
 - 12: **end for**
 - 13: **for** $k = 1, \dots, K$ **do**
 - 14: $\text{err}_k \leftarrow \mu(k) - \hat{\gamma}^{(K)}(k)$.
 - 15: **end for**
 - 16: **return** $\hat{\gamma}(i_1, \dots, i_K) \leftarrow \hat{\gamma}^{(K)}(i_1, \dots, i_K) + \prod_{k=1}^K \text{err}_k(i_k) / \|\text{err}_K\|_1^{K-1}$ for all i_1, \dots, i_K .
-

Multi-marginal Sinkhorn We use the Multi-marginal Sinkhorn algorithm as in Algorithm 2 .

Algorithm 2 Multi-marginal Sinkhorn

- 1: **Parameters:** Cost tensor $C \in \mathbb{R}^{n_1 \times \dots \times n_K}$, Marignals $\mu(1), \dots, \mu(K)$, Targeted accuracy $\varepsilon > 0$.
 - 2: Let $\eta = (4 \sum_{k=1}^K \log n_k) / \varepsilon$, $\varepsilon' = \varepsilon / (8\|C\|_\infty)$, and $A = \exp(-\eta C)$.
 - 3: Let $t = 0$, $m^{(0)}(k) = \mathbf{0} \in \mathbb{R}^{n_k}$ for $k \in [K]$, and $\gamma^{(0)} = A/\|A\|_1$.
 - 4: **for** $t = 1, 2, \dots$ **do**
 - 5: Let $k_t = \arg\max_k \sum_{i_k \in [n_k]} \mu(k, i_k) \log(\mu(k, i_k) / \gamma^{(t-1)}(k, i_k))$.
 - 6: $m^{(t)}(k_t) = m^{(t-1)}(k_t) + \log \mu(k_t) - \log \gamma^{(t-1)}(k_t)$.
 - 7: $\gamma^{(t)}(i_1, \dots, i_K) = A(i_1, \dots, i_K) \cdot \prod_{l=1}^k \exp(m^{(t)}(l, i_l)) \prod_{l=k+1}^K \exp(m^{(t-1)}(l, i_l))$ for all $i_1, \dots, i_K \in [n_1] \times \dots \times [n_K]$.
 - 8: **if** $\text{MTV}(\gamma^{(t)}, \mu(1), \dots, \mu(K)) \leq \varepsilon'$ **then break**
 - 9: **end for**
 - 10: Let $\hat{\gamma}^{(0)} = \gamma^{(t)}$.
 - 11: **for** $k = 1, \dots, K$ **do**
 - 12: $v(k) \leftarrow (\mu(k)/\hat{\gamma}^{(k-1)}(k)) \wedge \mathbf{1}$, where the division and \wedge are elementwise operation.
 - 13: $\hat{\gamma}^{(k)}(i_1, \dots, i_K) = \hat{\gamma}^{(k-1)}(i_1, \dots, i_K) \cdot v(k, i_k)$ for all $i_1, \dots, i_K \in [n_1] \times \dots \times [n_K]$.
 - 14: **end for**
 - 15: **for** $k = 1, \dots, K$ **do**
 - 16: $\text{err}_k \leftarrow \mu(k) - \hat{\gamma}^{(K)}(k)$.
 - 17: **end for**
 - 18: **return** $\hat{\gamma}(i_1, \dots, i_K) \leftarrow \hat{\gamma}^{(K)}(i_1, \dots, i_K) + \prod_{k=1}^K \text{err}_k(i_k) / \|\text{err}_K\|_1^{K-1}$ for all i_1, \dots, i_K .
-

The guarantees that are interesting for us are shown in the following:

11.1 Convergence

Theorem 2 (Theorem 16 of Lin et al. 2022). *When Algorithm 2 converges, we have*

$$\inf_{\gamma \in \Gamma(\mu(1), \dots, \mu(K))} \int_y C(y) \gamma(dy) \geq \int_y C(y) \hat{\gamma}(dy) - \varepsilon.$$

11.2 Application Of Log-sum-exp Trick

Log-sum-exp trick Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be any function with \mathcal{X} being any finite set. The celebrated log-sum-exp trick is used for computing quantities of the type $\log \sum_{x \in \mathcal{X}} \exp(f(x))$. It computes the equivalent form of $\max_x f(x) + \log \sum_x \exp(f(x) - \max_{x'} f(x'))$, which is much more stable computationally since $\sum_x \exp(f(x) - \max_{x'} f(x')) \geq 1$ and $f(x) - \max_{x'} f(x') \leq 0$ for all $x \in \mathcal{X}$.

In the implementation of Algorithm 2, the log-sum-exp trick can be apply to compute the update of $m^{(t)}(k)$: For Algorithm 2 specifically, the computation in 7 is unstable when ε is chosen small because it involves exponential to the power of magnitude $\Omega(\eta) = \Omega(1/\varepsilon)$. We can apply the celebrated log-sum-exp trick to stabilize the algorithm by skipping 7 and updating $m^{(t)}(k_t)$ in 6 for the next timestep by

$$m^{(t)}(k_t) = m^{(t-1)}(k_t) + \log \mu(k_t) - \log \sum_{\substack{i_1, \dots, i_{k_t-1} \\ i_{k_t+1}, \dots, i_K}} \exp \left(-\eta C(i_1, \dots, i_{k_t-1}, \cdot, i_{k_t+1}, \dots, i_K) + \sum_{l \neq k_t} m^{(t-1)}(l, i_l) \right),$$

where the final term is executed using the log-sum-exp trick.

11.3 Lower Bounds Without Convergence

Lemma 9 (Section 4.4 of Peyré, Cuturi, et al. 2019). *For any $t \geq 1$, Algorithm 2 satisfies*

$$\inf_{\gamma \in \Gamma(\mu(1), \dots, \mu(K))} \int_y C(y) \gamma(dy) \geq \frac{1}{\eta} \sum_{k=1}^K \int_{y(k)} m^{(t)}(y(k)) \mu(k, dy(k)).$$

12 ADDITIONAL SYNTHETIC DATASET ANALYSIS

12.1 2-d Gaussian Distributions With Known Data Generating Distribution

We consider 3 margins, each with a 2-dimensional outcome. We generate 100 data points $\{(Y_i^1(1), Y_i^2(1), Y_i^1(2), Y_i^2(2), Y_i^1(3), Y_i^2(3))\}_{i \in [100]}$ from the 6-d Gaussian distribution $\mathcal{N}(\mathbf{0}, \text{diag}(2, 2, 0.3, 0.3, 0.1, 0.1))$. Then, we separate each data point i into three margins, each being 2-dimensional, that is $(Y_i^1(1), Y_i^2(1))$, $(Y_i^1(2), Y_i^2(2))$, and $(Y_i^1(3), Y_i^2(3))$. The objective we choose in this case is $\sum_{j \in \{1, 2\}} (Y^j(1) - Y^j(2)/2 - Y^j(3)/2)^2$.

The exact expression for the theoretical optimal objective function is not accessible for $d > 1$. Therefore, we compare the empirical optimal objective to the known true value to assess whether it serves as a valid lower bound. The theoretical expectation of the objective is calculated to be 4.20, while the empirical optimal objective, obtained by running our Sinkhorn algorithm, is 1.78, confirming that the empirical optimal objective is indeed consistent with the true value of the causal estimand.

13 DESCRIPTION OF DATASETS

13.1 Epitaxial Layer Growth Data

In fabricating integrated circuit (IC) devices, an epitaxial layer is grown on polished silicon wafers mounted on a six-faceted susceptor within a metal bell jar. Chemical vapors are injected and heated until the layer reaches a desired thickness. This dataset originates from a factorial design experiment investigating the impact of various factors to the thickness (outcome). The experiment data we use considers two treatments: susceptor rotation method and nozzle position, resulting in a maximum of four margins.

13.2 Helpfulness Data

The study is a randomized experiment to learn subjects with uniform experience are more susceptible to self-serving motivated reasoning in both their empathy beliefs and redistribution choices. This experiment employs

a 2-by-2 design with two treatments: experience variation—subjects either have varied experience or uniform experience; timing of belief elicitation—time to tell the participant that there is a chance to take on a partner’s workload (corresponding to wealth redistribution). As a result, there are four margins. The outcome is the willingness of sharing the workload/redistributing the wealth.

14 ADDITIONAL REAL DATA ANALYSIS

14.1 Covariance Between Treatment Effects

Continued from Example 4, we investigate the correlation between treatment effects regarding different outcomes. In particular, we return to the Education dataset of the STAR project in Section 4.3.1. We consider the contrast effect in Section 4.3.1 for the responses $Y^j(\cdot)$, $j \in \{1, 2\}$ representing the j -th year GPA at different treatment levels. We estimate the partially identified set for the covariance between the contrasts regarding the first-year GPA and the second-year GPA.

The lower bound of the covariance is -0.322 , and the upper bound of the covariance is 2.256 . Even though we can not rule out the possibility that the treatment effects in the first-year GPA and the second-year GPA are negatively correlated, the asymmetry $|-0.322| \ll |2.256|$ seems to indicate that the treatment effects are positively correlated. To sharpen the bound of the covariance, one can further adjust for baseline covariates like initial GPA and age to remove the variation in the first, and second-year GPAs that are not attributed to the STAR project (discussed in Section 5).

There are five outcomes available in the dataset: the GPA at the end of year 1 (year 2), whether the student is in good standing at the end of year 1 (year 2), and the grade of the first semester in the first year. We also carried out the computation for covariances for other outcome pairs. The results of lower and upper bounds are displayed in Figure 2.

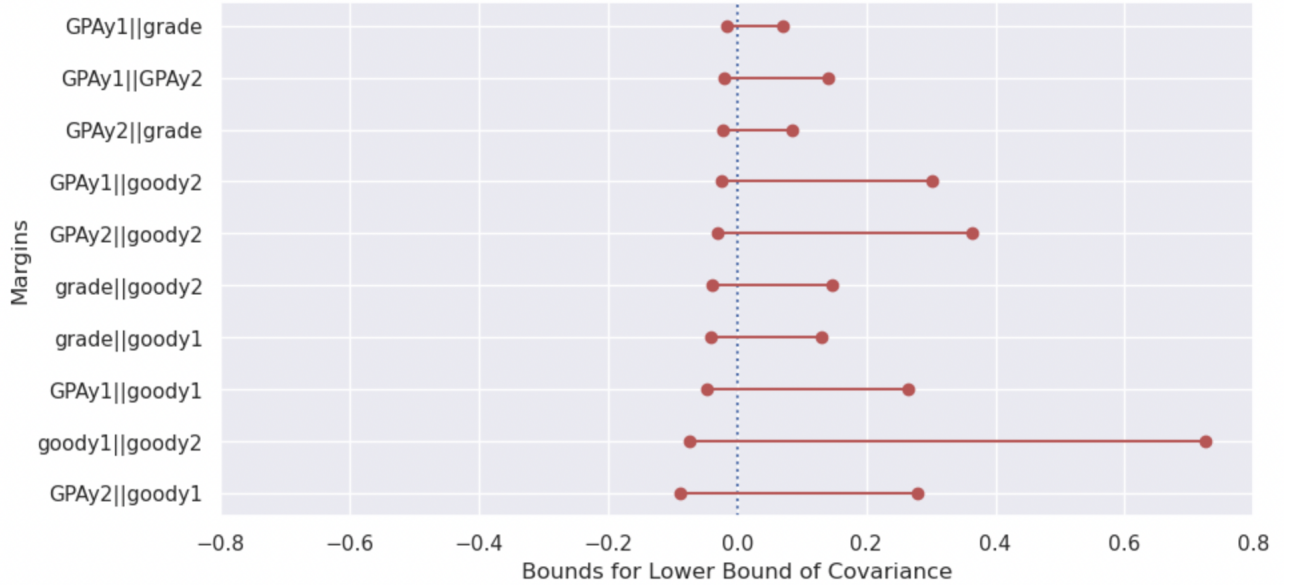


Figure 2: Lower and upper bounds for the covariance between different pairs of outcomes. Here GPAy1 means the GPA at the end of year 1, and similarly for GPAy2; goody1 means whether the student is in good standing at the end of year 1, and similarly for goody2; grade means the grade of the first semester in the first year.