Causal Partial Identification via Conditional Optimal Transport

Sirui Lin[1,*] **Zijun Gao**[2,*] **José Blanchet**[1] **Peter Glynn**[1]
[1]Department of Management Science and Engineering, Stanford University, USA
[2]Marshall School of Business, University of Southern California, USA

Abstract

We study the estimation of causal estimand involving the joint distribution of treatment and control outcomes for a single unit. In typical causal inference settings, it is impossible to observe both outcomes simultaneously, which places our estimation within the domain of partial identification (PI). Pre-treatment covariates can substantially reduce estimation uncertainty by shrinking the partially identified set. Recently, it was shown that covariate-assisted PI sets can be characterized through conditional optimal transport (COT) problems. However, finite-sample estimation of COT poses significant challenges, primarily because the COT functional is discontinuous under the weak topology, rendering the direct plug-in estimator inconsistent. To circumvent this, existing literature relies on relaxations or indirect methods involving the estimation of non-parametric nuisance statistics. In this work, we demonstrate continuity of the COT problem under a stronger topology induced by the adapted Wasserstein distance. Leveraging this result, we propose a direct, consistent, non-parametric estimator for COT that avoids nuisance parameter estimation. We derive the convergence rate for our estimator and validate its effectiveness through comprehensive experiments, demonstrating its improved performance compared to existing techniques.

*Equal contribution. Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1 INTRODUCTION

The potential outcome model (Rubin, 1974; Imbens and Rubin, 2015) is prevalent in causal inference, where each unit is associated with two potential outcomes: one under treatment and one under control. Since the two potential outcomes are never observed simultaneously, their joint distribution is not identified, making the causal estimands that depend on the joint distribution only partially identifiable. Optimal transport (Villani et al., 2009), which minimizes an expected cost over joint distributions respecting the marginals, has been used to recover partial identification (PI) sets of causal estimands (Gao et al., 2025). When pre-treatment covariates are available, they can often be used to reduce uncertainty about the joint distribution of potential outcomes, and produce smaller, more informative PI sets. Conditional optimal transport (COT) has been used to characterize covariate-assisted PI sets (Ji et al., 2023; Fan et al., 2025; Lin et al., 2025; Balakrishnan et al., 2025a) by finding the optimal joint distribution preserving the outcome-covariate distributions.

Despite the potential to yield more meaningful covariate-assisted PI sets, COT poses more substantial statistical challenges compared to OT, especially with continuous covariates. First, it is rare to observe a shared continuous covariate value in both treatment and control groups, leaving the conditional marginal distribution ill-defined at that value for at least one group. Specifically, when conditioning on an exact covariate value, there will typically be only a single unit in the sample that attains that value, with matched or repeated observations being unlikely. Second, even if each covariate value is associated with both a treated and a control unit (such as in twin studies), there is typically only one observation per covariate value in each group. The plug-in estimator of COT value, which uses point masses to estimate the conditional marginal distributions, may not converge to the true COT value as the sample size goes to infinity ((Chemseddine et al., 2025) and Example 2). This issue suggests that the COT functional may not

be continuous under the weak topology, rendering the COT value estimation more challenging than standard OT problems.

In this paper, we demonstrate the continuity of the COT functional under a stronger topology induced by the adapted Wasserstein distance, which requires uniform convergence of the outcome's conditional distribution at each covariate value. Leveraging the continuity result, we propose a direct, consistent, nonparametric estimator for the COT value. Our estimator is constructed using an adapted empirical distribution, which discretizes the covariate space by assigning observed covariates to a finite number of cells. The convergence of the adapted empirical distribution under the adapted Wasserstein distance ensures the consistency of our estimator, and we further establish its finite-sample error bounds. Compared to existing estimators for COT value that rely on relaxations of the original COT formulation (Manupriya et al., 2024; Lin et al., 2025), or utilize indirect approaches that require estimating nuisance functions (Ji et al., 2023), or resort to heuristic procedures without statistical guarantees (Tabak et al., 2021; Bunne et al., 2022; Hosseini et al., 2025), our proposed COT value estimator is direct and statistically consistent.

Contributions.

1. We establish the COT formulation for the covariate-assisted PI sets of a class of causal estimands (Theorem 1). We prove the continuity of the COT functional under the topology induced by the adapted Wasserstein distance (Theorem 2).

2. Building on (a), under Bernoulli design, we propose a direct nonparametric estimator for the covariate-assisted PI set, prove its consistency (Theorem 3), and further establish the finite-sample convergence rate (Theorem 4) together with a robustness guarantee under covariate's distributional shift (Theorem 5). We then extend the result to unconfounded designs with covariate-dependent treatment assignment (Theorem 6). In addition to its relevance for causal inference, our work advances the literature on COT by providing a consistent non-parametric estimator for the COT value with provable convergence rate.

3. We validate the estimator's effectiveness through comprehensive empirical studies (Section 5).

Organization. The rest of the paper is structured as follows. Section 2 introduces the problem setting. Section 3 provides basic properties of COT values, which serve as building blocks for our proposal. Section 4 presents the construction of our COT value estimator and the main theoretical results on its finite-sample performance. In Section 5, we provide simulation studies comparing our method with existing approaches. Technical proofs are deferred to the appendix.

Notation. We use $Z(w)$ to denote the variable Z observed within the control ($w = 0$) and treatment ($w = 1$) groups. We denote $[n] = \{1, ..., n\}, n \wedge m = \min(n, m), n \vee m = \max(n, m)$. We denote by $\mathcal{P}(\Omega)$ the set of probability measures on Ω. We use $\mu_{Y,Z}$ to denote the probability distribution (and its associated measure on $\mathcal{Y} \times \mathcal{Z}$) of (Y, Z) under μ, with μ_Z denoting the marginal distribution of Z, and μ_Y^z the conditional distribution of Y given $Z = z$. By $\mu = \mu_Z \otimes \mu_Y^Z$, we mean that for any measurable function g, $\int g \mathrm{d}\mu = \int_{\mathcal{Z}} \int_{\mathcal{Y}} g(z, y) \mathrm{d}\mu_Y^z(y) \mathrm{d}\mu_Z(z)$. We denote the set of the joint couplings π with marginals $\mu, \nu \in \mathcal{P}(\mathcal{X})$ ($\mathcal{X}$ may be $\mathcal{Y}, \mathcal{Z}$ or $\mathcal{Y} \times \mathcal{Z}$) by

$$\Pi(\mu, \nu) \triangleq \left\{\pi \in \mathcal{P}(\mathcal{X}^2) : \pi_X = \mu,\ \pi_{X'} = \nu\right\}. \quad (1)$$

2 PROBLEM FORMULATION

2.1 Potential Outcome Model

We consider the potential outcome model with a binary treatment (Rubin, 1974). Suppose there are N units. Each unit is associated with two potential outcomes, $Y_i(0)$ and $Y_i(1) \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$, $i \in [N]$, $d_Y \geq 1$, $\mathcal{Y}$ a convex closed set. Here $Y_i(0)$ is the outcome under control and $Y_i(1)$ is the outcome under treatment[1]. Each unit receives a binary treatment $W_i \in \{0, 1\}$, where $W_i = 1$ indicates treatment and $W_i = 0$ indicates control. For unit i, only the outcome $Y_i(W_i)$ is observed.

Often, each unit is also associated with a covariate vector $Z_i \in \mathcal{Z} \subseteq \mathbb{R}^{d_Z}$, $d_Z \geq 1$, $\mathcal{Z}$ a convex closed set. Covariates may often include continuous components, such as age or lab measurements. We denote the probability of receiving treatment conditional on the covariate value, typically referred to as propensity score, by $e(z) := \mathbb{P}(W_i = 1 \mid Z_i = z)$.

We impose the following standard assumptions (Imbens and Rubin, 2015) for the potential outcome model.

Assumption 1 (Stable Unit Treatment Value Assumption (SUTVA))**.** *Unit i's potential outcomes only depend on W_i but not W_j, $j \neq i$.*

Assumption 2 (Unconfoundedness)**.** *Treatment assignment is independent of potential outcomes given covariates, that is $Y_i(w) \perp\!\!\!\perp W_i \mid Z_i,\ w = 0, 1$.*

[1]We allow the potential outcomes to be multivariate to accommodate trials with multiple endpoints, experiments with a primary and secondary outcome, or policy evaluations involving effectiveness and cost.

Assumption 3 (Overlap). *There exists $\delta > 0$ such that for any possible covariate value z, the propensity score satisfies $\delta < e(z) < 1 - \delta$.*

To enable the asymptotic analysis, we adopt the super-population model (Imbens and Rubin, 2015) justified by Assp. 1, 2,

$$(Y_i(0), Y_i(1), Z_i) \overset{\text{i.i.d.}}{\sim} \mu \triangleq \mu_{Y(0),Y(1),Z},$$

where μ is the unknown joint distribution. The following result shows that, under the super-population model with unconfounded treatment assignment, the outcome-covariate distributions $(Y_i(1), Z_i)$ and $(Y_i(0), Z_i)$ are identifiable from the observed data. The outcome-covariate distributions will be later used in the COT formulation of the PI sets.

Proposition 1 (Identifiable marginals). *Under Assp. 1 - 3, for $w = 0, 1$,*

$$\begin{aligned} &\mathbb{P}\left(Y_i(w) = y, Z_i = z\right) \\ &= \tilde{e}_w(z)\ \mathbb{P}\left(Y_i(W_i) = y, Z_i = z \mid W_i = w\right), \end{aligned} \tag{2}$$

where $\tilde{e}_w(z) = \frac{(1-w)-\bar{e}}{(1-w)-e(z)}$, $\bar{e} \triangleq \int e(z)\mathrm{d}\mu_z(z)$.

2.2 PI Sets of Causal Estimands

In this paper, we focus on causal estimands of the form

$$V^* \triangleq \mathbb{E}_\mu[h(Y(0), Y(1))].$$

Here, $h : \mathcal{Y}^2 \to \mathbb{R}$ is a pre-specified objective function.

A fundamental challenge in causal inference is that $Y_i(0)$ and $Y_i(1)$ cannot be observed simultaneously for the same unit, making the joint distribution $\mu_{Y(0),Y(1)}$ inherently unidentifiable. As a result, for causal estimands depending on the joint distribution, their values are not identifiable from observed data[2]. This issue is commonly known as *partial identification*.

Example 1 (Variance of treatment effects). *For $h(y_0, y_1) = (y_0 - y_1)^2$, the corresponding estimand represents the variance of the treatment effect, which can be used to construct a confidence interval for the average treatment effect $\mathbb{E}[Y(1) - Y(0)]$ in the classic Neymanian framework (Splawa-Neyman, 1923). This estimand is only partially identifiable and has been studied in (Aronow et al., 2014; Balakrishnan et al., 2025b).*

For partially identifiable causal estimands, point estimation is infeasible. The object of interest becomes the partial identification (PI) set, that is, the set of values consistent with the marginal distributions of $(Y_i(1), Z_i)$ and $(Y_i(0), Z_i)$. Mathematically, PI set problems can be formulated under the OT framework (Section 2.3). First, define the coupling set as

$$\Pi_{\mathrm{c}} \triangleq \left\{\pi \in \mathcal{P}(\mathcal{Y}^2 \times \mathcal{Z}) : \pi_{Y(w),Z} = \mu_{Y(w),Z}\ w = 0, 1\right\}.$$

Then, the PI set is given by

$$\{\mathbb{E}_\pi[h(Y(0), Y(1))] : \pi \in \Pi_{\mathrm{c}}\} = [V_{\mathrm{c}}, \tilde{V}_{\mathrm{c}}]. \tag{3}$$

Here, we use the fact that the set Π_{c} is convex, the resulting PI set (3) is a convex subset of $\mathbb{R}$, and thus an interval. To construct this interval, it suffices to compute its lower and upper bounds, starting with

$$V_{\mathrm{c}} = \min_{\pi \in \Pi_{\mathrm{c}}} \mathbb{E}_\pi[h(Y(0), Y(1))], \tag{4}$$

and analogously the upper bound $\tilde{V}_{\mathrm{c}}$ via a maximization over the same coupling set.

Note that, for certain objective functions h (including Example 1), Hoeffding–type results yield a closed-form optimizer for (4) (see, e.g., (Tchen, 1980)). In contrast, we develop estimators for (4) that apply to broad classes of smooth functions h.

2.3 Conditional Optimal Transport

We first review optimal transport (OT), followed by a discussion of conditional optimal transport (COT).

OT has a rich literature (Villani et al., 2009) with broad applications in machine learning and statistics. In particular, OT has recently emerged as a powerful tool in causal inference (Black et al., 2020; Charpentier et al., 2023; De Lara et al., 2024; Torous et al., 2024). In an OT problem, the goal is to couple two distributions in a way that minimizes a specified cost function (e.g., h) while preserving their marginal distributions. The OT distance is defined as

$$\mathbb{W}_h(\mu, \nu) \triangleq \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi[h(X, X')],$$

where $\Pi(\mu, \nu)$ is defined in Eq. (1). When $h(x, x') = \|x - x'\|_2$, the OT distance is called Wasserstein 1-distance and denoted by $\mathbb{W}_1$. As discussed in Section 2.2, OT naturally connects to the PI set problem, where $\mathbb{W}_h(\mu_{Y(0)}, \mu_{Y(1)})$ can be interpreted as a lower bound of V^*.

Compared to OT, COT aligns conditional measures across values of the covariates, which has seen various applications, such as conditional sampling (Wang et al., 2025; Hosseini et al., 2025; Baptista et al., 2024), domain adaptation (Rakotomamonjy et al., 2022), and Bayesian flow matching (Kerrigan et al., 2024; Chemseddine et al., 2025). In this paper, we focus on the natural connection between COT and the

[2]Non-identifiability refers to the fact that two different joint distributions can yield different values of the causal estimand while producing the same distribution over observable quantities.

PI problem when covariates are available. Particularly, recall that in the PI set problem with Z_i, it becomes essential to preserve the outcome-covariate distributions of $(Y_i(0), Z_i)$ and $(Y_i(1), Z_i)$, where Z_i is shared across both marginals. The corresponding covariate-assisted lower bound is defined in Eq. (4), which, as shown below, is the optimal objective value of a COT problem.

Proposition 2 (Recursive formulation)**.** *For a measurable function $h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, we have*

$$V_{\mathrm{c}} = \int \mathbb{W}_h \left(\mu^z_{Y(0)}, \mu^z_{Y(1)}\right) \mathrm{d}\mu_Z(z).$$

2.4 Adapted Wasserstein Distance and Its Topology

We conclude this section by introducing the topology that is natural for the COT functional and that motivates our finite-sample estimator in Section 4.

As we will demonstrate in Section 3, a key distinction between OT and COT as functionals lies in their topological properties. While the classical OT functional is continuous with respect to the outcome marginals under the weak topology (metrized by $\mathbb{W}_1$), the COT functional V_{c} is generally discontinuous with respect to the joint outcome–covariate marginals under the weak topology. This discrepancy arises because COT is inherently a *conditional* transport problem: its geometry is driven by transition kernels rather than full joint distributions.

A natural way to formalize this structure is to restrict the set of admissible couplings to those that respect the conditional structure of the marginals, i.e., to (bi-)causal or adapted couplings; see, e.g., (Pflug and Pichler, 2012; Lassalle, 2018; Backhoff-Veraguas et al., 2020). From this perspective, the *adapted* Wasserstein distance can be understood as the optimal transport distance obtained when couplings are constrained to match conditional distributions. This interpretation is particularly natural in causal inference settings, where one aims to compare conditional outcome laws given covariates rather than arbitrary joint couplings of (Y, Z) and (Y', Z').

We therefore equip $\mathcal{P}(\mathcal{Y} \times \mathcal{Z})$ with the topology induced by the adapted Wasserstein distance.

Definition 1 (Adapted Wasserstein distance)**.** For $\mu, \nu \in \mathcal{P}(\mathcal{Y} \times \mathcal{Z})$, the adapted Wasserstein distance between μ and ν is defined as

$$\begin{aligned}&\mathbb{W}_{\mathrm{a}}(\mu, \nu)\\ &= \min_{\pi \in \Pi(\mu_Z, \nu_Z)} \int \|z - z'\|_2 + \mathbb{W}_1(\mu^z_Y, \nu^{z'}_Y)\ \mathrm{d}\pi(z, z').\end{aligned}$$

Compared to the conventional Wasserstein distance, which measures proximity between full joint distributions of (Y, Z) and (Y', Z'), the adapted Wasserstein distance quantifies discrepancies at the level of transition kernels. It therefore captures the transportation cost of conditional outcome laws while optimally aligning covariate marginals. As we show in Theorem 2, under mild regularity conditions, V_{c} becomes continuous with respect to this stronger, causally structured topology.

3 BASIC PROPERTIES OF COT VALUE

The following assumption enables our analysis.

Assumption 4 (Compact covariate domain)**.** *Assume that $\mathcal{Z} \subseteq [0, 1]^{d_Z}$, and $\mathcal{Y} \subseteq \mathbb{R}^{d_Y}$ is compact.*

Assumption 5 (Continuous objective)**.** *Assume that $h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a continuous function.*

Here, we assume that the outcome Y and the covariate Z are supported on a compact set, which is often satisfied in practice—for example, when the outcome includes test scores, earnings, or health indicators and the covariate includes bounded physical measurements like age, height, or weight.

3.1 Optimal Covariate-Assisted PI Sets

In this section, we show that if we only have access to i.i.d. samples drawn from $\mu_{Y(0),Z}$ and $\mu_{Y(1),Z}$, respectively and no additional side information, then the interval $[V_{\mathrm{c}}, \tilde{V}_{\mathrm{c}}]$ defined in (3) is the *optimal* PI set for V^* that can be identified. Note that, in the remaining parts of this work, we will use $((Y_i(w), Z_i(w))$ to denote the outcome and covariate values collected from the sample i within the control group ($w = 0$) and treatment group ($w = 1$). *Note that in this work, $Z_i(w)$ only contains pre-treatment covariates, and the w in $Z_i(w)$ is the group index of the observed pre-treatment covariates.*

Theorem 1 (Optimal PI bounds)**.** *Under Assp. 4-5, suppose that the sample $((Y_i(w), Z_i(w)), i \in [n])$ i.i.d. follows $\mu_{Y(w),Z}$ for $w = 0, 1$. Then, we have (i) $V_{\mathrm{c}}, \tilde{V}_{\mathrm{c}}$ are identifiable from the given sample, (ii) $[V_{\mathrm{c}}, \tilde{V}_{\mathrm{c}}]$ is the interval that exactly contains all possible values of $\mathbb{E}_\pi[h(Y(0), Y(1))]$, where π satisfies: $\pi_{Y(0),Z} = \mu_{Y(0),Z}, \pi_{Y(1),Z} = \mu_{Y(1),Z}$.*

3.2 Stronger Topology for COT Functional

As indicated by Proposition 2, the estimation of V_{c} requires estimating the conditional distributions $\mu^z_{Y(0)}, \mu^z_{Y(1)}$. Thus, a direct plug-in estimator does not

yield consistent estimates of V_c, a fact also observed in (Chemseddine et al., 2025). To further illustrate this issue, we consider an idealized case where the i-th covariate in the treated group is identical to that in the control group, i.e., $Z_i(1) = Z_i(0)\ \forall i$ (e.g. a twin study), presenting an example adapted from (Chemseddine et al., 2025, Example 9).

Example 2 (Inconsistent estimator)**.** *Let $n_0 = n_1$ and $h = |y_0 - y_1|$. Suppose that under μ, $Y(0)$, $Y(1)$, Z are independent, and $\mu_{Y(0),Z} = \mu_{Y(1),Z}$, thus $V_c = 0$. Let $\hat{\mu}_w = \frac{1}{n_w}\sum_{i=1}^{n_w} \delta_{Y_i(w),Z_i(w)}, w = 0, 1$. Further, assume that (i) $Z_i(1) = Z_i(0)\ \forall i$, and (ii) the distribution μ_Z has a density. Then,*

$$\lim_{n_0,n_1\to\infty} \min_{\pi\in\hat{\Pi}_c} \mathbb{E}_\pi[h(Y(0), Y(1))]$$
$$= \mathbb{E}_\mu[|Y(0) - Y(1)|] > 0 = V_c \qquad a.s.,$$

where $\hat{\Pi}_c = \left\{\pi : \pi_{Y(0),Z} = \hat{\mu}_0, \pi_{Y(1),Z} = \hat{\mu}_1\right\}$. This is because we aim for a perfect match in Z while almost surely the values of covariates in the treated group are distinct, i.e., $Z_i(1) \neq Z_j(1)$ for $i \neq j$. Therefore, $\hat{\Pi}_c$ only contains a single coupling: $\frac{1}{n_0}\sum_i \delta_{Y_i(0),Y_i(1),Z_i(0)}$. Thus, the objective simplifies to $\frac{1}{n_0}\sum_i |Y_i(1) - Y_i(0)|$, which converges to the population quantity $\mathbb{E}_\mu[|Y(1) - Y(0)|]$.

From a topological perspective, if V_c is viewed as a functional of $\mu_{Y(0),Z}$ and $\mu_{Y(1),Z}$, Example 2 implies that V_c is discontinuous under the weak topology. However, under mild assumptions, V_c becomes continuous at μ under the stronger topology induced by the adapted Wasserstein distance (see Definition 1).

Theorem 2 (Continuity under $\mathbb{W}_a$)**.** *Let $(\nu_n = \nu_{Y(0),Y(1),Z,n}, n \geq 1)$ be a sequence in $\mathcal{P}(\mathcal{Y}^2 \times \mathcal{Z})$. Suppose $g_w(z) = \mu^z_{Y(w)}$ $w = 0, 1$ are continuous under the weak topology and h is a continuous function. Under Assp. 4-5, if $\lim_{n\to\infty} \mathbb{W}_a(\nu_{Y(w),Z,n}, \mu_{Y(w),Z}) = 0$ for $w = 0, 1$, then*

$$\lim_{n\to\infty} \int \mathbb{W}_h\left(\nu^z_{Y(0),n}, \nu^z_{Y(1),n}\right) \mathrm{d}\nu_{Z,n}(z)$$
$$= \int \mathbb{W}_h\left(\mu^z_{Y(0)}, \mu^z_{Y(1)}\right) \mathrm{d}\mu_Z(z) = V_c.$$

In the next section, we construct an adapted empirical distribution that converges under $\mathbb{W}_a$, which serves as the foundation for our adapted COT value estimator.

4 ESTIMATOR OF COT VALUE

4.1 Bernoulli Design

4.1.1 Adapted COT Value Estimator

In this section, we introduce our adapted COT value estimator in the setting where the treatment is independent of both the observed outcome and covariates. This setting is related to a scenario referred to as the Bernoulli design (and also the randomized controlled trial) in causal inference, where W_i are i.i.d. Bernoulli with a constant probability parameter. In this setting, the sample from the control group, $((Y_i(0), Z_i(0)), i \in [n_0])$, can be viewed as an i.i.d. sample drawn from $\mu_{Y(0),Z}$, with an analogous interpretation for the treated group.

Assumption 6 (Bernoulli design)**.** *Assume that $((Y_i(w), Z_i(w)), i \in [n_w])$ are drawn i.i.d. from $\mu_{Y(w),Z}$ for $w = 0, 1$.*

Inspired by (Backhoff et al., 2022), we introduce the adapted empirical distribution for constructing our COT value estimator. First, we introduce the map below for discretization.

Definition 2 (Cell-center projection)**.** Let $r > 0$. For $n \geq 1$, partition the cube $[0, 1]^{d_Z}$ into a disjoint union of n^{rd_Z} smaller cubes, each with edge length n^{-r}. Define $\varphi^n_r : [0, 1]^{d_Z} \to [0, 1]^{d_Z}$ as the mapping that assigns each point to the center of the small cube it resides in.

Definition 3 (Adapted empirical distribution)**.** The adapted empirical distributions are defined as

$$\hat{\boldsymbol{\mu}}_{Y(w),Z(w),n_w} = \frac{1}{n_w}\sum_{i=1}^{n_w} \delta_{Y_i(w),\varphi^{n_w}_r(Z_i(w))} \qquad (5)$$

for $w = 0, 1$, where φ^n_r is defined in Definition 2.

Compared to (Backhoff et al., 2022), our construction of $\hat{\boldsymbol{\mu}}$ has a key distinction: we discretize only the covariate Z, not the outcome Y, since discretization is required only for the variables being conditioned on.

If $z \mapsto \mu^z_{Y(w)}$ is continuous under the weak topology, the adapted empirical distributions (5) provide consistent approximations to the conditional distributions $\mu^z_{Y(w)}, w = 0, 1$, thereby mitigating the issue that each sample of $Z(0)$ or $Z(1)$ is associated with only a single observed outcome Y when μ_Z has a density. Further, (Backhoff et al., 2022) suggests that $(\hat{\boldsymbol{\mu}}_{Y(w),Z(w),n_w}, w = 0, 1)$ converges to their weak limits under $\mathbb{W}_a$, which is aligned with the conditions introduced in Theorem 2.

Note that larger discretization cells contain more observations, and the resulting local distribution estimator is associated with smaller variance but larger bias. We can choose the cell size to balance the bias-variance trade-off, where the optimal cell size depends on factors including the sample size. As the sample size may differ between treatment and control groups, for example, when units are assigned to treatment with relatively low probability, we discretize each group separately to allow for distinct optimal cell sizes. In addition, in typical causal inference datasets, the covariate

data for the treatment and control groups are not perfectly matched, also leading to a potential mismatch in the supports of the adapted empirical distributions in (5). To address this issue, we match the discretized covariates $Z(0)$ and $Z(1)$ that may differ in value. This matching is formalized as a coupling with marginals $\hat{\boldsymbol{\mu}}_{Z(0),n_0}$ and $\hat{\boldsymbol{\mu}}_{Z(1),n_1}$. Based on this coupling, we define our adapted COT value estimator as follows.

Definition 4 (Adapted COT value estimator)**.** Given a coupling $\hat{\pi}_{n_0,n_1} \in \Pi(\hat{\boldsymbol{\mu}}_{Z(0),n_0}, \hat{\boldsymbol{\mu}}_{Z(1),n_1})$, the COT value estimator is defined as

$$\hat{V}_{n_0,n_1} \triangleq \int \mathbb{W}_h(\hat{\boldsymbol{\mu}}^{z_0}_{Y(0),n_0}, \hat{\boldsymbol{\mu}}^{z_1}_{Y(1),n_1}) \mathrm{d}\hat{\pi}_{n_0,n_1}(z_0, z_1).$$

The COT value estimator achieves consistency as the sample size grows, provided that the coupling $\hat{\pi}_{n_0,n_1}$ converges to a joint distribution that exactly matches the limiting marginals.

Theorem 3 (Consistency)**.** *Under Assp. 4-6, if as $n_0, n_1 \to \infty$, $\hat{\pi}_{n_0,n_1}$ weakly converges to $(\mathrm{id}, \mathrm{id})_{\#}\mu_Z$ almost surely, then $\hat{V}_{n_0,n_1}$ converges to V_{c} almost surely.*

4.1.2 Convergence Rate Analysis

Next, we provide the finite-sample convergence rate for the adapted COT value estimator. We introduce the following assumptions.

Assumption 7 (Smooth objective)**.** *Assume that h is L_h-Lipschitz continuous: for $\forall y_0, y_0', y_1, y_1' \in \mathcal{Y}$,*

$$|h(y_0, y_1) - h(y_0', y_1')| \leq L_h \left(\|y_0 - y_0'\|_2 + \|y_1 - y_1'\|_2 \right).$$

Assumption 8 (Lipschitz kernel)**.** *Assume that there is a constant $L_Z > 0$ such that for μ_Z-a.e. $z_0, z_1 \in \mathcal{Z}$,*

$$\mathbb{W}_1 \left(\mu^{z_0}_{Y(w)}, \mu^{z_1}_{Y(w)} \right) \leq L_Z \|z_0 - z_1\|_2 \quad w = 0, 1.$$

Assp. 7-8 are aligned with standard regularity conditions in nonparametric estimation theory (e.g., (Wasserman, 2006, Theorem 5.65)).

Assumption 9 (Coupling gap)**.** *Assume that $\forall n_0, n_1$: $\mathbb{E}[\int \|z_0 - z_1\|_2 \, \mathrm{d}\hat{\pi}_{n_0,n_1}(z_0, z_1)] \leq C_{\hat{\pi}}(n_0 \wedge n_1)^{-r}$, where $C_{\hat{\pi}}$ is a constant independent of n_0, n_1.*

Assp. 9 enables quantifying the discrepancy between the estimated coupling $\hat{\pi}_{n_0,n_1}$ and the identity coupling $(\mathrm{id}, \mathrm{id})_{\#}\mu_Z$, and to ensure that this error aligns with the granularity of the discretization, as controlled by the parameter r in the map φ^n_r.

Theorem 4 (Finite-sample complexity)**.** *Under Assp. 4-9 with $r = \frac{1}{d_Z + 2 \vee d_Y}$, we have*

$$\mathbb{E}\left[|\hat{V}_{n_0,n_1} - V_{\mathrm{c}}| \right] \leq \bar{C} \gamma_{d_Y, d_Z}(n_0 \wedge n_1),$$

where

$$\gamma_{d_Y, d_Z}(N) \triangleq \begin{cases} N^{-\frac{1}{d_Z + 2 \vee d_Y}} \log(N) & \text{if } d_Y \neq 2, \\ N^{-\frac{1}{d_Z + 2 \vee d_Y}} & \text{if } d_Y = 2. \end{cases}$$

Here, $\bar{C}$ is a constant that depends on d_Z, d_Y, L_h (Assp. 7), and L_Z (Assp. 8).

One natural choice of $\hat{\pi}_{n_0,n_1}$ that satisfies Assp. 9 is an optimal coupling of $\hat{\boldsymbol{\mu}}_{Z(0),n_0}$ and $\hat{\boldsymbol{\mu}}_{Z(1),n_1}$:

$$\hat{\pi}_{n_0,n_1} \in \underset{\pi \in \Pi(\hat{\boldsymbol{\mu}}_{Z(0),n_0}, \hat{\boldsymbol{\mu}}_{Z(1),n_1})}{\operatorname{argmin}} \mathbb{E}_{\pi}[\|Z(0) - Z(1)\|_2]. \quad (6)$$

An additional advantage of the coupling $\hat{\pi}_{n_0,n_1}$ (6) is that it enjoys a robustness guarantee against potential data perturbations to the covariate distributions.

Theorem 5 (Robustness)**.** *Define $\hat{\pi}_{n_0,n_1}$ by (6). Suppose that the sample of $Z(w)$ is i.i.d. drawn from μ_w', satisfying $\mathbb{W}_1(\mu_Z, \mu_w') \leq \epsilon$ for $w = 0, 1$. Then,*

$$\mathbb{E}\left[|\hat{V}_{n_0,n_1} - V_{\mathrm{c}}| \right] \leq 2 L_h L_Z \epsilon + \bar{C} \gamma_{d_Y, d_Z}(n_0 \wedge n_1).$$

4.2 Covariate-Dependent Treatment

Unlike Bernoulli design where treatment assignment is independent of covariates by design, observational studies typically exhibit covariate-dependent treatment assignment. This results in non-constant propensity scores $e(z) = \mathbb{P}(W_i = 1 \mid Z_i = z)$, reflecting heterogeneity in treatment likelihood across covariate profiles. In this section, we discuss how the adapted COT value estimator can be tailored to this setting.

Under the unconfoundedness assumption (Assp. 2), the conditional distributions of potential outcomes given covariates remain identifiable, i.e., consistent with $\mu^z_{Y(w)}$ for $w = 0, 1$. However, since the treatment assignment probability depends on covariates, the marginal distribution of covariates may differ between the treatment group and the control group, inducing a covariate shift. To formalize this setting with covariate shift, we consider the following superpopulation model of $Y(0), Y(1), Z, W$ following the joint distribution $\mu = \mu_{Y(0),Y(1),Z,W}$.

Assumption 10 (Covariate shift)**.** *Assume that $((Z_i(w), Y_i(w)), i \in [n_w])$ are drawn i.i.d. from $\mu_{Z|W=w} \otimes \mu^Z_{Y(w)}$ for $w = 0, 1$.*

Since our estimand of interest V^* relies on the marginal covariate distribution, irrespective of treatment assignment, it is essential to correct for the distributional shift induced by non-constant propensity scores. Specifically, we reweight the samples to align with the marginal covariate distribution. By Proposition 1, when the propensity score $e(z)$ is known, the

appropriate population-level weighting is given by

$$\xi_{w,i} = \frac{\mu_Z(Z_i(w))}{\mu_{Z|W=w}(Z_i(w))} \propto \frac{1}{1-w-e(Z_i(w))},\ W_i = w.$$

On finite samples, when the propensity score $e(z)$ is unknown but estimated by $\hat{e}(z)$, we adopt a group-wise self-normalized weight: for $w = 0, 1$, $\hat{\xi}_{w,G}$ is defined as

$$\frac{(1-w-\hat{e}(\varphi_r^{n_w}(G)))^{-1}}{\sum_G (1-w-\hat{e}(\varphi_r^{n_w}(G)))^{-1}|\{i : Z_i(w) \in G\}|} \quad (7)$$

to ensure the weights sum to one. Here, $G \in \Phi_r^n$ is a cell corresponding to projection map φ_r^n, where $\Phi_r^n \triangleq \left\{(\varphi_r^n)^{-1}(\{x\}) : x \in \varphi_r^n([0,1]^{d_Z})\right\}$, and $\varphi_r^n(G)$ is the cell center of $G \in \Phi_r^n$. Note that the weight (7) is different from the common self-normalized weights that are based on the sample value of Z. This is used to adapt the construction of the adapted COT value estimator with discretization to the setting in Assp. 10. Finally, our reweighted adapted empirical distributions are defined by: for $w = 0, 1$,

$$\hat{\boldsymbol{\mu}}_{Y(w),Z(w),n_w} = \sum_{G \in \Phi_r^{n_w}} \sum_{i:Z_i(w)\in G} \hat{\xi}_{w,G} \delta_{Y_i(w),\varphi_r^{n_w}(Z_i(w))}.$$

We then still employ $\hat{\boldsymbol{\mu}}_{Y(w),Z(w),n_w}, w = 0, 1$ along with $\hat{\pi}_{n_0,n_1}$ defined in Eq. (6), to construct the adapted COT value estimator as formulated in Definition 4.

Specifically, we implement a two-fold cross-fitting procedure to ensure the independence between the estimated propensity score $\hat{e}$ and the adapted empirical distributions $\hat{\boldsymbol{\mu}}$. We randomly split the data into two folds of approximately equal size. For each fold, we construct $\hat{e}$ using data from the other fold, and then construct $\hat{\boldsymbol{\mu}}_{Y(w),Z(w),n_w}$ on the current fold with the estimated propensity score plugged in. This yields two estimates of the lower bound, and we further average them to form our final estimator $\hat{V}_{n_0,n_1}$.

We introduce the assumptions on the estimation error of $\hat{e}(\cdot)$ to enable the convergence analysis of $\hat{V}_{n_0,n_1}$.

Assumption 11 (Lipschitz $\hat{e}$)**.** *Assume that there is a constant L_e, such that $|\hat{e}(z) - \hat{e}(z')| \le L_e \|z - z'\|_2$ for any $z, z' \in \mathcal{Z}^2$.*

Assumption 12 (Bounds of $\hat{e}$)**.** *Assume that there is a constant $\eta \in (0, \frac{1}{2})$, such that $\eta \le \hat{e}(z) \le 1 - \eta$ for any $z \in \mathcal{Z}$.*

Assumption 13 (Average error of propensity score)**.** *Assume that for $n_0, n_1 \ge 1$,*

$$\mathbb{E}\left[\frac{1}{n_w}\sum_{i=1}^{n_w}\left|n_w\hat{\xi}_{w,i} - \frac{\mathrm{d}\mu}{\mathrm{d}\mu_{|W=w}}(Z_i(w))\right|\right] \le C_{\mathrm{w}} n_w^{-r},$$

where C_{w} is a constant independent of n_0, n_1 and

$$\hat{\xi}_{w,i} \triangleq \frac{(1-w-\hat{e}(Z_i(w)))^{-1}}{\sum_{l=1}^{n}(1-w-\hat{e}(Z_l(w)))^{-1}}, w = 0, 1.$$

Assp. 13 characterizes the average estimation error of the estimated propensity score function $\hat{e}$. Notably, Assp. 13 uses the sample-based self-normalized weight to be compatible with the common setting in the literature. Also, we impose a mild convergence rate requirement of order $O(n^{-r})$, where r will be chosen to be less than $1/d_Z$. The rate $O(n^{-1/d_Z})$ aligns with typical nonparametric convergence rates for nuisance parameter estimation. Assp. 11 and 12 impose boundedness and smoothness conditions on $\hat{e}(\cdot)$. Particularly, the smoothness condition is necessary because we evaluate $\hat{e}$ at the centers of discretized cells, as discussed in the following.

Theorem 6 (Finite-sample complexity with covariate shift)**.** *Under Assp. 3-5, 7-13 with $r = \frac{1}{d_Z + 2\vee d_Y}$, then*

$$\mathbb{E}\left[|\hat{V}_{n_0,n_1} - V_{\mathrm{c}}|\right] \le C^{\dagger}\gamma_{d_Y,d_Z}(n_0 \wedge n_1),$$

where $C^{\dagger}$ is a constant that depends on d_Z, d_Y, δ (Assp. 3), L_h (Assp. 7), L_Z (Assp. 8), L_e (Assp. 11), η (Assp. 12) and C_{w} (Assp. 13).

5 SIMULATION

5.1 Selection of Cell Size

In Definition 2, we design the discretization cells with edge length growing in the order of n^{-r}, resulting in $O(n^{rd_Z})$ small cells, in order to balance the bias-variance tradeoff with respect to the sample size n. In practice, the constant in $O(n^{rd_Z})$ may also affect the estimation accuracy. Specifically, in this section we set the number of cells to the nearest integer of cn^{rd_Z}.

We conduct a sensitivity analysis of our method with respect to c under the quadratic location model (see Section 5.2 (b)) with $n_0 = n_1 = 300$. As shown in Table 1, we compute the average relative estimation error (and their standard error), i.e., $\mathbb{E}[|\hat{V}_{n_0,n_1} - V_{\mathrm{c}}|]/V_{\mathrm{c}}$ for 5 choices of c, where $\mathbb{E}$ is approximated through 500 Monte Carlo simulation runs. Compared to the baseline $c = 1$, we can see that when c is relatively small, increasing c improves the estimation accuracy while its performance tends to stabilize as c grows larger.

Table 1: Sensitivity to c (baseline $c = 1$).

c	Relative error	Δ vs. baseline
0.6	0.21 (5.4e-3)	+38.1%
0.8	0.17 (5.2e-3)	+13.4%
1.0	**0.15 (4.9e-3)**	**0.0**
1.2	0.14 (4.8e-3)	−6.4%
1.4	0.14 (4.9e-3)	−7.8%

In practice, since the true value of V_{c} is unknown, we suggest the following approach to select a c that

achieves a relatively higher estimation accuracy: (i) given a dataset with size N, generate $B = 50$ bootstrap samples, each with size N, (ii) compute the average $\hat{V}_{n_0,n_1}$ value of the bootstrap samples for each candidate of c, and plot a curve on the average values, (iii) pick the "elbow" point of the curve (the point with largest distance to the line connecting the start and end point of the curve), and choose the corresponding c value. An illustration is shown in Figure 1.

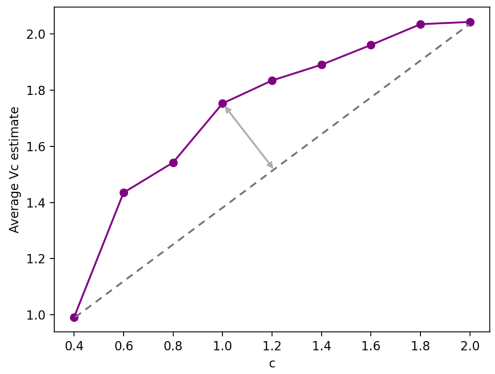


Figure 1: Selection of the elbow point.

5.2 Simulation Results

Synthetic data. To evaluate the effectiveness of the adapted COT value estimator in simulation[3], we introduce the following synthetic data-generating mechanism. For $w = 0, 1$, $Y(w) = F_m(f_w(Z), \varepsilon_w)$, where $Z, \varepsilon_w, Y(w)$ are independent with each other. The model F_m includes: (i) location model $F_1(u, v) = u+v$, and (ii) scale model $F_2(u, v) = (u_j v_j, j \in [d_Y])$. More concretely, we consider $d_Y = d_Z = 1$, $Z, \varepsilon_w \sim N(0, 1)$, and three models:

(a) Linear location model ($m = 1$) with $f_0(z) = -0.6z, f_1(z) = 1.6z$.

(b) Quadratic location model ($m = 1$) with $f_0(z) = -0.2z^2, f_1(z) = 0.6z^2$.

(c) Scale model ($m = 2$) with $f_0(z) = 0.5z - 0.35, f_1(z) = 1.1z + 0.35$.

For these Gaussian models, the V_c's population value has a closed form without referring to an optimization problem (see appendix for a proof), and is computable through a simple Monte Carlo algorithm with parametric convergence rate. Therefore, these models enable the evaluation of the estimation accuracy.

[3]The code and additional experimental results are available at `github.com/siruilin1998/causalPIviaCOT`.

We compare the estimation accuracy of our adapted COT value estimator with the method in (Ji et al., 2023), implemented in their Python package `DualBounds`[4]. Their approach relies on estimating nuisance functions associated with the covariate–outcome model. In our comparison, we evaluate both the ridge regression-based and k-nearest neighbor (KNN)-based `DualBounds` methods. Our implementation is based on the Python Optimal Transport (`POT`[5]) library (Flamary et al., 2021).

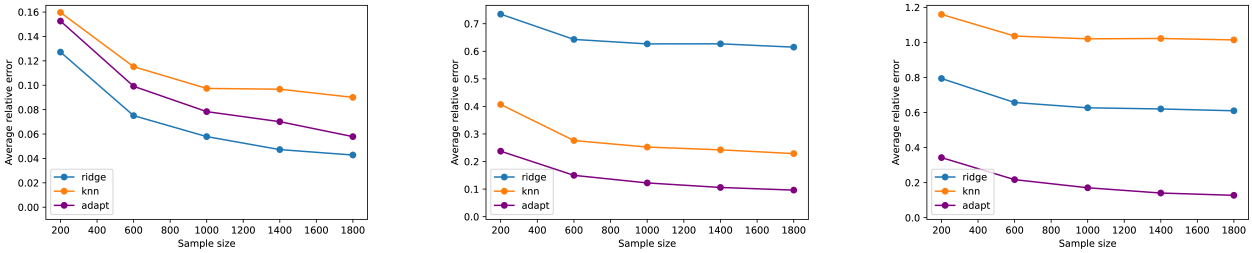
In Figure 2, we present the comparison in the setting of: (i) Bernoulli design ((a)(b)(c)) and (ii) Covariate-dependent treatment ((d)(e)(f)) with $e(z) = (1 + e^{-\frac{3z}{2}})^{-1}$ $\forall z \in \mathbb{R}$, which we assume to be fully known for all methods under evaluation. The figure shows that: (i) For the linear location model, the ridge-based `DualBounds` method performs best due to its use of a linear nuisance estimator that aligns with the true model structure. Nonetheless, the adapted COT value estimator achieves comparable accuracy despite not relying on model-specific assumptions (see (a)). (ii) For the quadratic location model, the adapted COT value estimator outperforms both versions of `DualBounds` and the KNN-based outperforms the ridge-based, as the latter cannot capture the nonlinearity in the true model. (iii) For the scale model, the adapted COT estimator also has the best performance.

Real data. We also consider the real-world dataset from the Student Achievement and Retention (STAR) Demonstration Project (Angrist et al., 2009), where the academic performance (outcome) is measured by the GPA recorded in the first academic year and the baseline GPA serves as the covariate. To make the COT value accessible at the population level, we evaluate the methods using hypothetical data generated from a model fitted to the real data. Table 2 presents the accuracy on estimating the PI bounds of the correlation between two potential outcomes, and our adapted COT value estimator consistently outperforms the others.

6 CONCLUSION AND DISCUSSION

Conclusion. In this paper, we study the optimal covariate-assisted partial identification sets for causal estimands by solving COT problems. Our finite-sample COT estimator avoids nuisance function estimation, does not require well-specified models, and enjoys a provable statistical convergence rate. Furthermore, our adapted COT value estimator answers the question proposed in the discussion of (Lin et al.,

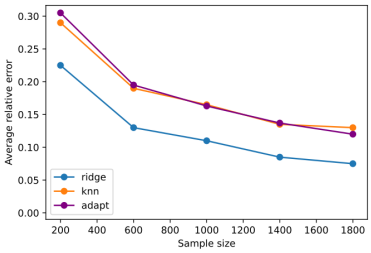
[4]`https://dualbounds.readthedocs.io`
[5]`https://pythonot.github.io/`



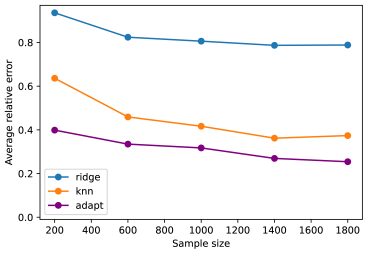
(a) Linear location model (Bernoulli design)

(b) Quadratic location model (Bernoulli design)

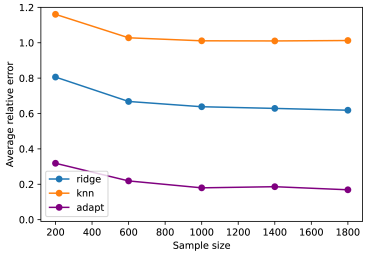
(c) Scale model (Bernoulli design)



(d) Linear location model (Covariate-dependent)



(e) Quadratic location model (Covariate-dependent)



(f) Scale model (Covariate-dependent)

Figure 2: Estimation accuracy comparison between the adapted COT estimator and `DualBounds`. In the legend, "adapt" refers to the adapted COT estimator, "ridge" refers to the ridge regression–based `DualBounds`, and "knn" refers to the KNN–based `DualBounds`. The average relative error is computed over 500 Monte Carlo repetitions. For the uncertainty, we compute the standard error of the mean (SEM), i.e. $\mathrm{SD}/\sqrt{500}$, where SD is the standard deviation. The numerical values of SEM satisfy: (a): < 0.005; (b)(c): < 0.02; (d)(e)(f): < 0.018.

Table 2: Average relative error (SEM) over 500 Monte Carlo repetitions on the STAR-based semi-synthetic data.

Method ($N = n_0 + n_1 = 2n_0$)	$N = 400$	$N = 800$	$N = 1200$	$N = 1600$
ridge regression–based `DualBounds`	4.0e−2 (1.3e-3)	3.2e−2 (1.0e-3)	2.9e−2 (0.9e-3)	2.7e−2 (0.8e-3)
KNN–based `DualBounds`	5.1e−2 (1.6e-3)	4.5e−2 (1.3e-3)	4.5e−2 (1.1e-3)	4.4e−2 (1.0e-3)
Our method	**3.6e-2 (1.2e-3)**	**2.9e-2 (1.0e-3)**	**2.6e-2 (0.8e-3)**	**2.5e-2 (0.8e-3)**

2025) on how to combine the adapted Wasserstein distance into the COT value estimation. Specifically, our adapted estimator is their $V_{\mathrm{causal}}(\eta)$ when η approached infinity.

Triangular transport maps. We next discuss the connection between our COT framework and triangular transport maps. When the transport cost is quadratic and the conditional outcome distributions admit densities, the optimal coupling can be represented by a triangular transport map (Carlier et al., 2010, 2016; Hosseini et al., 2025). Such maps arise from a sequential alignment of conditional distributions and coincide, in the one-dimensional outcome case, with the classical Knothe–Rosenblatt rearrangement (Rosenblatt, 1952; Knothe, 1957).

Triangular transport has been used in several causal modeling and inference frameworks (Charpentier et al., 2023; De Lara et al., 2024; Balakrishnan et al., 2025a), where it is often interpreted as a structural mechanism for generating counterfactual outcomes. Our perspective is complementary. Rather than modeling unit-level counterfactuals via an estimated transport map, we use the induced causal coupling structure to define and estimate a distributional functional. In particular, our estimator targets the COT value directly, without requiring explicit construction of the underlying triangular map.

Nevertheless, we conjecture that our discretization approach (Section 4.1.1) could be adapted to recover a triangular transport map in a manner analogous to plug-in constructions in the unconditional setting with a similar order of the estimation error; see, for example, the one-nearest-neighbor estimator in (Manole et al., 2024, Section 4.2).

Acknowledgements

The material in this paper is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0397. Additional support is gratefully acknowledged from NSF 2118199, 2229012, 2312204, 2403007, and ONR 13983111.

References

Joshua Angrist, Daniel Lang, and Philip Oreopoulos. Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics*, 1(1):136–163, 2009.

Peter M Aronow, Donald P Green, and Donald KK Lee. Sharp bounds on the variance in randomized experiments. *Annals of Statistics*, 42(3):850–871, 2014.

Julio Backhoff, Mathias Beiglbock, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. *SIAM Journal on Optimization*, 27(4):2528–2562, 2017.

Julio Backhoff, Daniel Bartl, Mathias Beiglböck, and Johannes Wiesel. Estimating processes in adapted Wasserstein distance. *Annals of Applied Probability*, 32(1):529–550, 2022.

Julio Backhoff-Veraguas, Daniel Bartl, Mathias Beiglböck, and Manu Eder. Adapted Wasserstein distances and stability in mathematical finance. *Finance and Stochastics*, 24(3):601–632, 2020.

Sivaraman Balakrishnan, Edward Kennedy, and Larry Wasserman. Conservative inference for counterfactuals. *Journal of Causal Inference*, 13(1):20230071, 2025a.

Sivaraman Balakrishnan, Edward Kennedy, and Larry Wasserman. Conservative inference for counterfactuals. *Journal of Causal Inference*, 13(1):20230071, 2025b.

Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. On the representation and learning of monotone triangular transport maps. *Foundations of Computational Mathematics*, 24(6):2063–2108, 2024.

D. P. Bertsekas and S. E. Shreve. *Stochastic Optimal Control: The Discrete Time Case*, volume 139 of *Mathematics in Science and Engineering*. Academic Press, New York, 1978.

Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

Jose Blanchet, Martin Larsson, Jonghwa Park, and Johannes Wiesel. Bounding adapted Wasserstein metrics. *arXiv preprint arXiv:2407.21492*, 2024.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional Monge maps. In *Advances in Neural Information Processing Systems*, volume 35, pages 6859–6872, 2022.

Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41 (6):2554–2576, 2010.

Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: an optimal transport approach. *Annals of Statistics*, pages 1165–1192, 2016.

Arthur Charpentier, Emmanuel Flachaire, and Ewen Gallic. Optimal transport for counterfactual estimation: A method for causal inference. In *Optimal Transport Statistics for Economics and Related Topics*, pages 45–89. Springer, 2023.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional Wasserstein distances with applications in Bayesian OT flow matching. *Journal of Machine Learning Research*, 26(141): 1–47, 2025.

Hugh Dance and Benjamin Bloem-Reddy. Counterfactual cocycles: a framework for robust and coherent counterfactual transports. *arXiv preprint arXiv:2405.13844*, 2025.

Lucas De Lara and Luca Ganassali. What is a good matching of probability measures? a counterfactual lens on transport maps. *arXiv preprint arXiv:2509.16027*, 2025.

Lucas De Lara, Alberto González-Sanz, Nicholas Asher, Laurent Risser, and Jean-Michel Loubes. Transport-based counterfactual models. *Journal of Machine Learning Research*, 25(136):1–59, 2024.

Yanqin Fan and Sang Soo Park. Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951, 2010.

Yanqin Fan, Brendan Pass, and Xuetao Shi. Partial identification in moment models with incomplete data via optimal transport. *arXiv preprint arXiv:2503.16098*, 2025.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162 (3):707–738, 2015.

Zijun Gao, Shu Ge, and Jian Qian. Bridging multiple worlds: multi-marginal optimal transport for causal partial-identification problem. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, 2025.

James J Heckman, Jeffrey Smith, and Nancy Clements. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.

Bamdad Hosseini, Alexander W Hsu, and Amirhossein Taghvaei. Conditional optimal transport on function spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):304–338, 2025.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge university press, 2015.

Wenlong Ji, Lihua Lei, and Asher Spector. Model-agnostic covariate-assisted inference on partially identified causal effects. *arXiv preprint*, 2023.

Reid A. Johnson. quantile-forest: A python package for quantile regression forests. *Journal of Open Source Software*, 9(93):5976, 2024. doi: 10.21105/joss.05976. URL `https://doi.org/10.21105/joss.05976`.

Olav Kallenberg. *Foundations of Modern Probability*. Probability and Its Applications. Springer, Berlin, 2nd edition, 2002.

Yuta Kawakami and Jin Tian. Moments of causal effects. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, pages 2011–2043, 2025.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Dynamic conditional optimal transport through simulation-free flows. *Advances in Neural Information Processing Systems*, 37:93602–93642, 2024.

Herbert Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4(1):39–52, 1957.

Rémi Lassalle. Causal transport plans and their Monge–Kantorovich problems. *Stochastic Analysis and Applications*, 36(3):452–484, 2018.

Sirui Lin, Zijun Gao, Jose Blanchet, and Peter Glynn. Tightening causal bounds via covariate-aware optimal transport. In *International Conference on Machine Learning*, pages 37815–37845. PMLR, 2025.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *Annals of Statistics*, 52(3):966–998, 2024.

Charles F Manski. The mixing problem in programme evaluation. *The Review of Economic Studies*, 64(4): 537–553, 1997.

Piyushi Manupriya, Rachit K. Das, Sayantan Biswas, and SakethaNath N. Jagarlapudi. Consistent optimal transport with empirical conditional measures. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, pages 3646–3654, 2024.

Roger B Nelsen. *An Introduction to Copulas*. Springer, 2006.

Georg Ch Pflug and Alois Pichler. A distance for multistage stochastic optimization models. *SIAM Journal on Optimization*, 22(1):1–23, 2012.

Georg Ch Pflug and Alois Pichler. From empirical observations to tree models for stochastic optimization: convergence properties. *SIAM Journal on Optimization*, 26(3):1715–1740, 2016.

Alain Rakotomamonjy, Rémi Flamary, Gilles Gasso, M El Alaya, Maxime Berar, and Nicolas Courty. Optimal transport for conditional domain matching and label shift. *Machine Learning*, pages 1–20, 2022.

Murray Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3): 470–472, 1952.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Jerzy Splawa-Neyman. On the application of probability theory to agricultural experiments. Essay on principles. section 9. *Statistical Science*, pages 465–472, 1923.

Esteban G Tabak, Giulio Trigila, and Wenjun Zhao. Data driven conditional optimal transport. *Machine Learning*, 110:3135–3155, 2021.

André H Tchen. Inequalities for distributions with given marginals. *Annals of Probability*, pages 814–827, 1980.

William Torous, Florian Gunsilius, and Philippe Rigollet. An optimal transport approach to estimating causal effects via nonlinear difference-in-differences. *Journal of Causal Inference*, 12(1): 20230004, 2024.

Cédric Villani et al. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Zheyu Oliver Wang, Ricardo Baptista, Youssef Marzouk, Lars Ruthotto, and Deepanshu Verma. Effi-

cient neural network approaches for conditional optimal transport with applications in Bayesian inference. *SIAM Journal on Scientific Computing*, 47(4): C979–C1005, 2025.

Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, 2006.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) 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. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes]
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. [Yes]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes]

Supplementary Materials

Organization. The appendix is organized as follows. Section A presents the details of the real data experiment discussed in Section 5.2. Section C presents supplementary experiments to show the robustness and effectiveness of our method. Section D reviews technical background knowledge. Section E provides proofs for the results in Section 2. Section F provides proofs for the results in Section 3. Section G provides proofs for the results in Section 4. Section I collects proofs for supporting lemmas.

A More Details on the Real Data Experiment

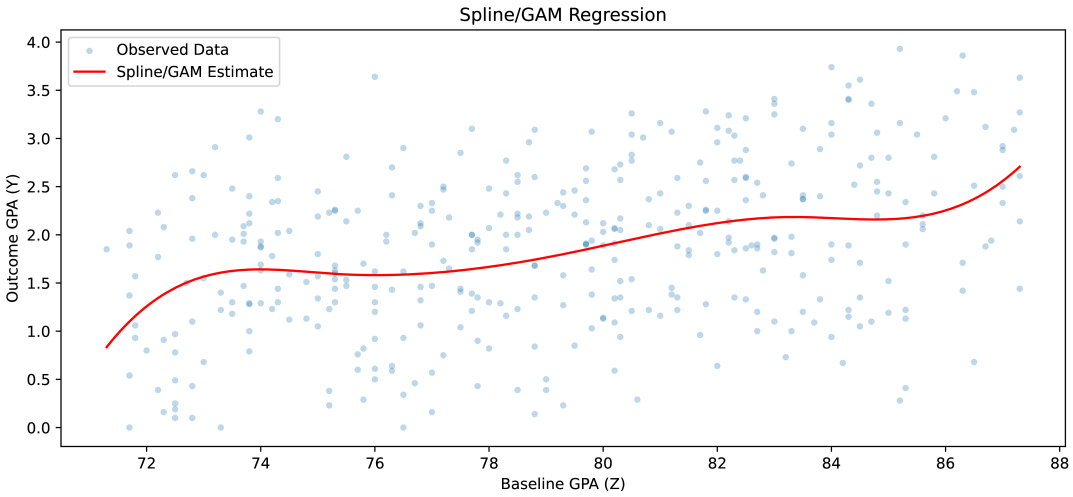


Figure 3: Cubic spline regression of the outcome variable Y (GPA) versus the covariate Z (baseline GPA) for the treatment group. The distribution of Z is modeled using Gaussian kernel density estimation (KDE), while the relationship between the outcome Y and Z is modeled using cubic spline regression. The Wasserstein distance between the empirical and KDE-generated distributions of Z is less than 0.2. The Wasserstein distance between the observed and model-generated distributions of Y is less than 0.07 for the treatment group and less than 0.09 for the control group.

In this section, we present more details of the real data experiment in Section 5.2, which is based on the Student Achievement and Retention (STAR) Demonstration Project Angrist et al. (2009), an initiative designed to assess the impact of scholarship incentive programs on academic performance. Experiments were run on a MacBook Air (Apple M3, 8-core CPU, 16 GB RAM), without GPU acceleration.

In the STAR study, the treatment is access to a scholarship incentive and was randomly assigned. Academic performance is measured by the GPA recorded at the end of the first academic year. In addition to the treatment indicator and outcome variable, the dataset includes baseline GPA (measured prior to treatment assignment), which serves as a key covariate due to its strong predictive influence on academic outcomes.

We are interested in the correlation between two potential outcomes of each unit i, which is defined as

$$\rho \triangleq \frac{\mathbb{E}[Y_i(1)Y_i(0)] - \mathbb{E}[Y_i(1)]\mathbb{E}[Y_i(0)]}{\sqrt{\mathrm{Var}(Y_i(0))\mathrm{Var}(Y_i(1))}}.$$

The unidentifiable part of ρ arises from the estimand $\mathbb{E}[Y_i(1)Y_i(0)]$. Then, to obtain the corresponding V_c value for ρ, we could utilize the cost function $h(y_0, y_1) = (y_0 + y_1)^2$. To ensure the PI bounds are accessible at the

population level, we generate hypothetical data from a model fitted to the real data. Particularly, we fit the real data using the model $Y(w) = f_w(Z(w)) + \mathcal{N}(0, \sigma_w^2)$ $w = 0, 1$, and subsequently generate data based on the estimated functions f_w and noise level σ_w.

To ensure the true (population-level) PI bounds are accessible, we generate hypothetical data from a model fitted to the real data. Particularly, we fit the real data using the model

$$Y(k) = f_k(Z(k)) + \mathcal{N}(0, \sigma_k^2) \quad k = 0, 1.$$

and subsequently generate data based on the estimated functions f_k and noise level σ_k for $k = 0, 1$ (see Figure 3). In addition, we model the distribution of baseline GPA (Z) using Gaussian kernel density estimation. Specifically, we consider

$$f_k(Z) = \sum_{j=1}^{6} \hat{\beta}_{kj} \phi_j(Z) \quad k = 0, 1,$$

where $(\phi_j, j = 1, ..., 6)$ are cubic spline basis functions. The regression coefficients $(\hat{\beta}_{kj}, j = 1, ..., 6, k = 0, 1)$ are estimated via ridge regression applied to the real data, with the regularization parameter selected through cross-validation.

B More Discussion

Multiple treatment levels. We discuss here the possibilities of an extension to settings with multiple treatment levels. Suppose the treatment variable takes more than two levels. A natural extension of our framework is to consider multi-marginal optimal transport, which constructs a joint coupling across all conditional outcome distributions given the covariates. This provides a geometric framework for jointly aligning multiple potential outcome distributions. The unconditional multi-treatment setting has been studied in (Gao et al., 2025).

At the same time, the use of transport maps in counterfactual modeling has important limitations. Recent work (Dance and Bloem-Reddy, 2025; De Lara and Ganassali, 2025) shows that optimal transport maps generally cannot serve as structural models for counterfactual assignments when more than two treatment levels are present. We emphasize that our methodology does not interpret OT maps as structural counterfactual mechanisms. Instead, OT and COT are employed as distributional comparison tools: they measure discrepancies between conditional outcome laws and construct PI intervals via geometric constraints on admissible couplings. Consequently, conditional OT remains meaningful in multi-treatment settings for constructing PI intervals and bounding functionals of potential outcomes.

C Supplementary Experiments

C.1 Robustness to Covariate Distributional Shift

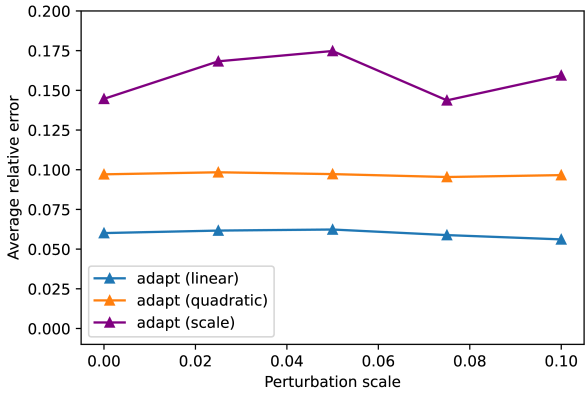


Figure 4: Robustness of our method to covariate mismatch. The average relative error is defined as $\mathbb{E}\left[|\text{estimator} - V_c|\right]/V_c$, where the expectation $\mathbb{E}$ is approximated by averaging over 500 Monte Carlo repetitions. To quantify the uncertainty, we compute the standard error of the mean (SEM) as $\text{SD}/\sqrt{500}$, where SD denotes the standard deviation of the relative errors across repetitions. The maximum SEM values are less than 0.006 in this setting.

In this section, we consider a setting in which the covariates of the treatment group are shifted. Let $\mu_{Z(0)}, \mu_{Z(1)}$ denote the (marginal) generating distribution of $Z(0), Z(1)$. Specifically, we set $\mu_Z = \mu_{Z(0)} = \mathcal{N}(0,1)$ and perturb the treatment covariate distribution as $\mu_{Z(1)} = \mathcal{N}(0,1) + \eta\varepsilon$, where ε is an independent $\mathcal{N}(0,1)$ noise and η controls the magnitude of the perturbation. Using the same three models described in Section 5.2, we evaluate the estimation accuracy of the adapted COT value estimator under varying levels of η. The results are reported in Figure 4. We observe that the adapted COT value estimator remains stable across small perturbation levels ($\eta \in [0, 0.1]$), illustrating its robustness to covariate distribution shift (see also Theorem 5). Intuitively, this robustness stems from the projection of covariates onto a finite set of representative values: small perturbations in the covariates are often absorbed during coarsening, as they rarely alter cell assignments. Moreover, the use of optimal coupling to match the discretized marginals helps mitigate the impact of systematic perturbations such as location shifts.

C.2 Comparison with a Fréchet–Hoeffding-Type Approach

When the outcomes are one-dimensional, the copula models Nelsen (2006), and the Fréchet–Hoeffding bounds Heckman et al. (1997); Manski (1997); Fan and Park (2010); Kawakami and Tian (2025) can provide partial identification bounds in closed forms without solving an OT-type optimization problem. Particularly, for the case $d_Y = 1$ and $h(y_0, y_1) = (y_0 - y_1)^2$, Hoeffding-type results yield a closed-form optimizer for (4). (Note that Hoeffding-type bounds do not provide consistent partial identification sets for $d_Y > 1$, whereas our procedure does.)

Proposition 3 (Fréchet–Hoeffding bounds for $h(y_0, y_1) = (y_0 - y_1)^2$)**.** *Let $F_w(\cdot|z)$ $w = 0, 1$ be the marginal CDFs of Y_w conditioned on $Z = z$ on $\mathbb{R}$ with finite second moments, and denote their quantile functions by $F_w^{-1}(u|z) = \inf\{x : F_w(x|z) \geq u\}$ for $w = 0, 1$. Then, for any coupling π of (Y_0, Y_1, Z) with these marginals,*

$$\int_{\mathcal{Z}} \int_0^1 \left(F_0^{-1}(u|z) - F_1^{-1}(u|z)\right)^2 \mathrm{d}u\mathrm{d}z \;\leq\; \mathbb{E}_\pi\big[(Y_0 - Y_1)^2\big] \;\leq\; \int_{\mathcal{Z}} \int_0^1 \left(F_0^{-1}(u|z) - F_1^{-1}(1-u|z)\right)^2 \mathrm{d}u\mathrm{d}z.$$

Conditioned on $Z = z$, the lower bound is attained by the comonotone coupling $Y_0 = F_0^{-1}(U|z)$, $Y_1 = F_1^{-1}(U|z)$, and the upper bound by the countermonotone coupling $Y_0 = F_0^{-1}(U|z)$, $Y_1 = F_1^{-1}(1-U|z)$, where $U \sim$ Uniform$(0, 1)$.

Based on the above result, a Fréchet-Hoeffding-Type approach Aronow et al. (2014); Balakrishnan et al. (2025a) for estimating V_c proceeds as follows:

(i) **Quantile regression:** Within the control and treatment groups, estimate the conditional quantile functions by running quantile regressions. For example, in the following, we use a non-parametric quantile regression method, random forest quantile regression (python package: quantile_forest Johnson (2024)[6]).

(ii) **Fréchet–Hoeffding bound:** For each covariate value (in either the control or treatment group), use the estimated conditional quantiles to compute the squared loss.

(iii) **Aggregation:** Approximate the overall value by taking the sample average of the values of the conditional squared loss derived in (ii) over the distinct covariate values.

To compare our adapted COT value estimator and the above Fréchet–Hoeffding–based approach, we consider the scale model (Section 5.2 (c)) with the same model parameter as there. We run a Monte Carlo simulation with 200 repetitions. The mean absolute error (standard error) results are collected in Table 3 (In this setting we could compute the oracle true value (=1.92) of the PI bounds with high accuracy). The results show that our method is comparable with the Fréchet–Hoeffding–based method in this special one-dimensional outcome setting.

Table 3: Comparison of our method and Fréchet–Hoeffding–based method using the scale model.

	N = 100	N = 500	N = 1000	N = 1500
Our Method	0.356 (0.025)	0.176 (0.011)	0.138 (0.007)	0.122 (0.007)
F–H Based	0.276 (0.019)	0.180 (0.009)	0.156 (0.007)	0.148 (0.006)

D Technical Background

D.1 Regular Kernel

In this paper, we always assume that joint distribution like $\mu_{Y,Z}$ that can be written as $\mu_{Y,Z} = \mu_Z \otimes \mu_Y^Z$, i.e. for any measurable function g, $\int g(y,z)\mathrm{d}\mu = \int_{\mathcal{Z}} \int_{\mathcal{Y}} g(y,z)\mathrm{d}\mu_Y^z(y)\mathrm{d}\mu_Z(z)$, with a regular kernel μ_Y^z. The definition of a regular kernel is stated as follows.

Definition 5 (Regular kernel)**.** A kernel μ_Y^z is regular if (i) for any $z \in \mathcal{Z}$, $\int \mathbf{1}(y \in \cdot)\mathrm{d}\mu_Y^z(y)$ is a probability measure, and (ii) for any measurable set $B \subseteq \mathcal{Y}$, $z \mapsto \int \mathbf{1}(y \in B)\mathrm{d}\mu_Y^z(y)$ is a measurable function.

We can always assume that μ has a regular kernel due to the following result.

Proposition 4 (Disintegration (Kallenberg, 2002, Theorem 5.3, 5.4))**.** *If $\mathcal{Y} \subseteq \mathbb{R}^{d_Y}$ and is equipped with the Borel σ-algebra, then for a measure $\mu_{Y,Z} \in \mathcal{P}(\mathcal{Y} \times \mathcal{Z})$, there is a regular kernel μ_Y^z such that $\int g(y,z)\mathrm{d}\mu = \int_{\mathcal{Z}} \int_{\mathcal{Y}} g(y,z)\mathrm{d}\mu_Y^z(y)\mathrm{d}\mu_Z(z)$ for any measurable function g. Further, μ_Y^z is unique μ_Z-a.e.*

D.2 Wasserstein vs. Adapted Wasserstein: A Distance Metric Comparison

This section presents the two metrics employed to study the continuity of COT.

Definition 6 (Wasserstein distance)**.** The 1-Wasserstein distance between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is defined as

$$\mathbb{W}_1(\mu,\nu) \triangleq \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_\pi[\|X(0) - X(1)\|_2].$$

The 1-Wasserstein distances possess key properties that will be instrumental in our proof.

Lemma 1 (Triangle inequality)**.** $\mathbb{W}_1(\mu,\mu') \le \mathbb{W}_1(\mu,\nu) + \mathbb{W}_1(\nu,\mu')$.

Lemma 2 (Convexity)**.** *The map $\mu \mapsto \mathbb{W}_1(\mu,\mu')$ is convex.*

[6] https://pypi.org/project/quantile-forest/

For $\mathcal{X} = \mathcal{Z} \times \mathcal{Y}$, the so-called adapted Wasserstein distance induces a topology that is typically stronger than the weak topology, which can be metrized by the Wasserstein distances.

Definition 7 (Adapted Wasserstein distance)**.** The adapted Wasserstein distances between $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is defined as

$$\mathbb{W}_{\mathrm{a}}(\mu, \nu) \triangleq \min_{\pi \in \Pi_{\mathrm{a}}(\mu,\nu)} \mathbb{E}_{\pi}[\|X(0) - X(1)\|_2],$$

where $X(0) = (Z(0), Y(0)), X(1) = (Z(1), Y(1))$, and

$$\Pi_{\mathrm{a}}(\mu, \nu) \triangleq \{\pi \in \Pi(\mu, \nu) : \text{under } \pi,\ Z(1) \perp\!\!\!\perp Y(0) \text{ conditioned on } Z(0),\\ Z(0) \perp\!\!\!\perp Y(1) \text{ conditioned on } Z(1)\}.$$

The following proposition is stated to maintain consistency with Definition 1 introduced earlier.

Proposition 5 (Reformulation of adapted Wasserstein distance; see, e.g. (Backhoff et al., 2017, Proposition 5.2))**.** *Let $\mu = \mu_Z \otimes \mu_Y^Z$ and $\nu = \nu_Z \otimes \nu_Y^Z$, we have*

$$\mathbb{W}_{\mathrm{a}}(\mu, \nu) = \min_{\pi \in \Pi(\mu_Z, \nu_Z)} \int \|z - z'\|_2 + \mathbb{W}_1(\mu_Y^z, \nu_Y^{z'})\ \mathrm{d}\pi(z, z').$$

Since $\Pi_{\mathrm{a}}(\mu, \nu) \subseteq \Pi(\mu, \nu)$, clearly $\mathbb{W}_1 \leq \mathbb{W}_{\mathrm{a}}$, but the reverse may not hold. Consider an i.i.d. sample $(X_i = (Z_i, Y_i), i \in [n])$ drawn from a distribution μ, and let $\hat{\mu}_n \triangleq \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. It is well known that $\mathbb{W}_1(\hat{\mu}_n, \mu)$ converges to zero almost surely as n goes to infinity. However, as shown in (Pflug and Pichler, 2016, Proposition 1), $\mathbb{W}_{\mathrm{a}}(\hat{\mu}_n, \mu)$ may fail to converge to zero, and a remedy to ensure convergence is to replace $\hat{\mu}_n$ with its convolution with a suitable kernel function. The intuition behind this approach is to ensure that the kernel of the convoluted empirical distribution becomes smooth in the variable it is conditioned on. This idea aligns with the findings of Blanchet et al. (2024), which show that for distributions with Lipschitz kernels, the weak topology and the topology induced by the adapted Wasserstein distance are equivalent.

However, since the convolved empirical distribution is not discrete, its practical implementation requires an additional sampling step for computation. To address this, Backhoff et al. (2022) propose a discrete alternative—the adapted empirical distribution—which mitigates the convergence issue by assigning the sample $(X_i)_{i\in[n]}$ to a finite number of cells, which is the approach we adopt in this paper.

D.3 Fournier and Guillin's Result

This section provides the convergence rate of the empirical distribution under the Wasserstein distance, as established by Fournier and Guillin (2015).

Suppose that the sample $(X_i, i \in [n])$ i.i.d. follows μ and denote $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Theorem 1 in Fournier and Guillin (2015) characterizes a nearly sharp convergence rate for empirical Wasserstein distances.

Theorem 7 (Empirical Wasserstein convergence rate, Theorem 1 Fournier and Guillin (2015))**.** *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ and let $q > 0$. Assume that $M_k(\mu) := \int \|x\|_2^k \,\mathrm{d}\mu < \infty$ for some $k > q$. There exists a constant C depending only on q, d, k such that, for all $n \geq 1$,*

$$\mathbb{E}\big(\mathbb{W}_q(\hat{\mu}_n, \mu)^q\big) \leq C M_k^{q/k}(\mu) \begin{cases} n^{-1/2} + n^{-(k-q)/k} & \text{if } q > d/2, k \neq 2q, \\ n^{-1/2}\log(1+n) + n^{-(k-q)/k} & \text{if } q = d/2, k \neq 2q, \\ n^{-q/d} + n^{-(k-q)/k} & \text{if } q \in (0, d/2), k \neq d/(d-q), \end{cases}$$

where $\mathbb{W}_q(\mu, \nu) \triangleq \min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi[\|X(0) - X(1)\|_2^q]^{\frac{1}{q}}$.

In this work, we will utilize the following corollary.

Corollary 1 ($k > 2$)**.** *Under the same conditions of Theorem 7, if $k > 2$, then*

$$\mathbb{E}[\mathbb{W}_1(\hat{\mu}_n, \mu)] \leq C R_d(n),$$

where C is a constant that depends on $k, q, d, \mathbb{E}_\mu[\|x\|_2^k]$, and

$$R_d(n) = \begin{cases} n^{-\frac{1}{2\vee d}} & \text{if } d \neq 2, \\ n^{-\frac{1}{2}} \log(3+n) & \text{if } d = 2. \end{cases} \tag{8}$$

Note that here we take $\log(3+n)$ rather than $\log(1+n)$ to make $x \mapsto xR_d(x)$ a concave function.

E Proofs of Section 2

E.1 Proof of Proposition 1

Proof of Proposition 1. We prove the equality for the treatment potential outcome ($k = 1$); the result for the control potential outcome ($k = 0$) follows analogously. In fact,

$$\begin{aligned} \mathbb{P}(Y_i(1) = y, Z_i = z) &= \frac{\mathbb{P}(Y_i(1) = y, Z_i = z, W_i = 1)}{\mathbb{P}(W_i = 1 \mid Y_i(1) = y, Z_i = z)} \\ &= \frac{\mathbb{P}(Y_i(1) = y, Z_i = z \mid W_i = 1) \cdot \mathbb{P}(W_i = 1)}{\mathbb{P}(W_i = 1 \mid Z_i = z)} \\ &= \mathbb{P}(Y_i(1) = y, Z_i = z \mid W_i = 1) \cdot \frac{\mathbb{P}(W_i = 1)}{e(z)}, \end{aligned}$$

where we use Assumption 2 in the second last equality. □

E.2 Proof of Proposition 2

Remark 1. In this proof, we will use the notions of Borel measurability, lower semianalyticity, and universal measurability, which are discussed in detail in (Bertsekas and Shreve, 1978, Section 7.7).

Proof of Proposition 2. We first show that the right-hand side (RHS) $\int \mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)})\mathrm{d}\mu_Z(z)$ is well-defined. Since h is a measurable function bounded from below, then the map

$$\mathcal{L}_h \triangleq (z, \gamma) \mapsto \int h(y_0, y_1)\mathrm{d}\gamma(y_0, y_1)$$

is a Borel (measurable) and thus lower semianalytic (l.s.a.).

Note that the maps $z \mapsto \mu^z_{Y(0)}$ and $z \mapsto \mu^z_{Y(1)}$ are both Borel since $\mu^z_{Y(0)}, \mu^z_{Y(1)}$ are regular kernels (see Section D.1). Further, $\{(p, q, \gamma) : \gamma \in \Pi(p, q)\}$ is weakly closed. Then, the set

$$D \triangleq \{(z, \gamma) : \gamma \in \Pi(\mu^z_{Y(0)}, \mu^z_{Y(1)})\}$$

is Borel.

As a result, $\mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)})$ is the infimum of the function $\mathcal{L}_h$ over the fiber of D at z. By (Bertsekas and Shreve, 1978, Proposition 7.47), the map $z \mapsto \mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)})$ is l.s.a. and in particular universally measurable. Therefore, the RHS $\int \mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)})\mathrm{d}\mu_Z(z) + \epsilon.$ is well-defined.

Next, we prove the equation from two directions as follows.

LHS $\leq$ RHS: For any $\epsilon > 0$ By (Bertsekas and Shreve, 1978, Proposition 7.50(b)), there is a universally measurable ϵ-optimizer for

$$\inf_{\gamma \in \Pi(\mu^z_{Y(0)}, \mu^z_{Y(1)})} \int h(y_0, y_1)\mathrm{d}\gamma(y_0, y_1).$$

That is, there exists a universally measurable map $z \mapsto \gamma^z_\epsilon \in \Pi(\mu^z_{Y(0)}, \mu^z_{Y(1)})$ such that

$$\int h(y_0, y_1)\mathrm{d}\gamma^z_\varepsilon(y_0, y_1) \leq \inf_{\gamma \in \Pi(\mu^z_{Y(0)}, \mu^z_{Y(1)})} \int h(y_0, y_1)\mathrm{d}\gamma(y_0, y_1) + \varepsilon. \tag{9}$$

Then, apply (Bertsekas and Shreve, 1978, Proposition 7.45), we can construct a measure $\pi_\epsilon \in \mathcal{P}(\mathcal{Y}^2 \times \mathcal{Z})$ such that

$$\int g(y_0, y_1, z) \mathrm{d}\pi_\epsilon(y_0, y_1, z) = \int \int_{\mathcal{Y}^2} g(y_0, y_1, z) \mathrm{d}\gamma_\varepsilon^z(y_0, y_1) \mathrm{d}\mu_Z(z)$$

for any measurable function g.

As a result, $\pi \in \Pi_\mathrm{c}$, and thus

$$V_\mathrm{c} = \min_{\pi \in \Pi_\mathrm{c}} \int h(y_0, y_1) \mathrm{d}\pi \leq \int h(y_0, y_1) \mathrm{d}\pi_\epsilon \leq \int \mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)}) \mathrm{d}\mu_Z(z) + \epsilon,$$

where the second inequality is due to (9).

Send $\epsilon \to 0^+$, we get $V_\mathrm{c} \leq \int \mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)}) \mathrm{d}\mu_Z(z)$.

LHS $\geq$ RHS: For any $\pi \in \Pi_\mathrm{c}$, it can be written as $\pi = \mu_Z \otimes \pi^Z_{Y(0),Y(1)}$, and $\pi^z_{Y(0),Y(1)} \in \Pi(\mu^z_{Y(0)}, \mu^z_{Y(1)})$ μ_Z-almost everywhere. This is because for any measurable function g,

$$\int g(y, z) \mathrm{d}\mu_{Y(0),Z}(y, z) = \int g(y, z) \mathrm{d}\pi_{Y(0),Z}(y, z) = \int_{\mathcal{Z}} \int_{\mathcal{Y}} g(y_0, z) \mathrm{d}\pi^z_{Y(0)}(y) \mathrm{d}\mu_Z(z),$$

and the argument on $(Y(1), Z)$ is similar.

As a result,

$$\begin{aligned}\int h(y, y') \mathrm{d}\pi(y, y', z) &= \int_{\mathcal{Z}} \mathrm{d}\mu_Z(z) \int_{\mathcal{Y} \times \mathcal{Y}} h(y, y') \mathrm{d}\pi^z_{Y(0),Y(1)}(y, y') \\ &\geq \int_{\mathcal{Z}} \mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)}) \mathrm{d}\mu_Z(z).\end{aligned}$$

Taking infimum on π completes this part of proof.

□

F Proofs of Section 3

F.1 Proof of Theorem 1

Lemma 3 (Π_c is compact)**.** Π_c *is a compact subset of* $\mathcal{P}(\mathcal{Y}^2 \times \mathcal{Z})$ *under the weak topology.*

Proof of Theorem 1. Given the function h, the estimand V_c only depends on $\mu_{Y(0),Z}, \mu_{Y(1),Z}$, which are identifiable from the observed samples, thus (i) holds true.

For (ii), by definition of V_c, for any $\pi \in \Pi_\mathrm{c}$, $\mathbb{E}_\pi[h(Y(0), Y(1))] \geq V_\mathrm{c}$, indicating that V_c is a valid lower bound for V^*. Further, since h is bounded and lower semicontinuous, $\mathbb{E}_\pi[h(Y(0), Y(1))]$ is a lower semicontinuous function in π with respect to the weak topology ((Villani et al., 2009, Lemma 4.3)). By Lemma 3, Π_c is compact, thus the minimum V_c can be attained.

For (iii), for any $v = (1-\theta)V_\mathrm{c} + \theta\tilde{V}_\mathrm{c}$, where $\theta \in [0, 1]$, there is a coupling $\pi_v = (1-\theta)\pi_\mathrm{c} + \theta\tilde{\pi}_\mathrm{c}$ such that $\pi_v \in \Pi_\mathrm{c}$ (since Π_c is a convex set) and $v = \mathbb{E}_{\pi_v}[h(Y(0), Y(1))]$. Then, the PI set is convex and thus is an interval.

Since now h is continuous, then by (ii), $\tilde{V}_\mathrm{c}$ is the tightest upper bound for the PI set. As a result, the PI set is equal to $[V_\mathrm{c}, \tilde{V}_\mathrm{c}]$. □

Remark 2 (Unbounded h)**.** The results of Theorem 1 can be extended to possibly unbounded function h such that, there are functions $a(y) \in L^1(\mu_{Y(0)}), b(y) \in L^1(\mu_{Y(1)})$ ($f \in L^1(\mu)$ denotes that $\int |f| \mathrm{d}\mu < \infty$) and $|h(y_0, y_1)| \leq a(y_0) + b(y_1)$. For the proof, we can consider $\tilde{h} = h(y_0, y_1) - (a(y_0) + b(y_1))$ for the lower bound and $\tilde{h}' = a(y_0) + b(y_1) - h(y_0, y_1)$ for the upper bound.

F.2 Proof of Theorem 2

Lemma 4 (Uniformly continuous kernel)**.** *Suppose that $\mathcal{Y}, \mathcal{Z}$ are both compact and $g_w(z) = \mu^z_{Y(w)} : \mathcal{Z} \to \mathcal{P}(\mathbb{R}^{d_Y})$ $w = 0, 1$ are continuous under the weak topology. Then, for any $\delta > 0$, there is a constant $C(\delta)$ such that $\mathbb{W}_1\left(\mu^{z_0}_{Y(k)}, \mu^{z_1}_{Y(k)}\right) \le \delta + C(\delta)\|z_0 - z_1\|_2$.*

Lemma 5 (Lipschitz continuity of $\mathbb{W}_h$)**.** *For any $\epsilon \ge 0$, suppose that $h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ satisfies that for any $y, y' \in \mathcal{Y}$,*

$$|h(y) - h(y')| \le L_h \|y - y'\|_2 + \epsilon. \tag{10}$$

Then, for any $(\mu, \nu, \mu', \nu') \in \mathcal{P}(\mathcal{Y})$,

$$|\mathbb{W}_h(\mu, \nu) - \mathbb{W}_h(\mu', \nu')| \le L_h \left(\mathbb{W}_1(\mu, \mu') + \mathbb{W}_1(\nu, \nu')\right) + 2\epsilon.$$

Proof of Theorem 2. For $\nu = \nu_{Y(0),Y(1),Z}$, we bound the following gap:

$$\begin{aligned}
&\left|V_{\mathrm{c}} - \int \mathbb{W}_h\left(\nu^z_{Y(0)}, \nu^z_{Y(1)}\right) \mathrm{d}\nu_Z(z)\right| \\
=&\left|\int \mathbb{W}_h\left(\mu^z_{Y(0)}, \mu^z_{Y(1)}\right) \mathrm{d}\mu_Z(z) - \int \mathbb{W}_h\left(\nu^z_{Y(0)}, \nu^z_{Y(1)}\right) \mathrm{d}\nu_Z(z)\right| \\
\le&\underbrace{\left|\int \mathbb{W}_h\left(\mu^z_{Y(0)}, \mu^z_{Y(1)}\right) \mathrm{d}\mu_Z(z) - \int \mathbb{W}_h\left(\mu^z_{Y(0)}, \nu^z_{Y(1)}\right) \mathrm{d}\nu_Z(z)\right|}_{\text{(Term A)}} \\
&+\underbrace{\left|\int \mathbb{W}_h\left(\mu^z_{Y(0)}, \nu^z_{Y(1)}\right) \mathrm{d}\nu_Z(z) - \int \mathbb{W}_h\left(\nu^z_{Y(0)}, \nu^z_{Y(1)}\right) \mathrm{d}\nu_Z(z)\right|}_{\text{(Term B)}}.
\end{aligned}$$

Here, the first equation is due to Proposition 2.

By Assp. 4-5, h is continuous and supported on a compact domain, then h is uniformly continuous. Therefore, for any $\epsilon > 0$, there is a constant $\widetilde{C}(\epsilon)$ such that for any $y, y' \in \mathcal{Y}$,

$$|h(y) - h(y')| \le \widetilde{C}(\epsilon) \|y - y'\|_2 + \epsilon.$$

Thus, Lemma 5 is applicable and for any $(\mu, \nu, \mu', \nu') \in \mathcal{P}(\mathcal{Y})$,

$$|\mathbb{W}_h(\mu, \nu) - \mathbb{W}_h(\mu', \nu')| \le \widetilde{C}(\epsilon) \left(\mathbb{W}_1(\mu, \mu') + \mathbb{W}_1(\nu, \nu')\right) + 2\epsilon. \tag{11}$$

Similarly, by Lemma 4, we have

$$\mathbb{W}_1\left(\mu^{z_0}_{Y(k)}, \mu^{z_1}_{Y(k)}\right) \le C(\epsilon)\|z_0 - z_1\|_2 + \epsilon. \tag{12}$$

For **(Term A)**: we highlight that we consider the optimal coupling $\pi \in \mathcal{P}(\mathcal{Y}^2 \times \mathcal{Z}^2)$ that is attained in the adapted Wasserstein distance $\mathbb{W}_{\mathrm{a}}(\mu_{Y(1),Z}, \nu_{Y(1),Z})$ to connect μ_Z and ν_Z, then for $\epsilon, \epsilon' > 0$,

$$\begin{aligned}
\text{(Term A)} =&\left|\int \mathbb{W}_h\left(\mu^z_{Y(0)}, \mu^z_{Y(1)}\right) \mathrm{d}\mu_Z(z) - \int \mathbb{W}_h\left(\mu^z_{Y(0)}, \nu^z_{Y(1)}\right) \mathrm{d}\nu_Z(z)\right| \\
=&\left|\int \mathbb{W}_h\left(\mu^z_{Y(0)}, \mu^z_{Y(1)}\right) - \mathbb{W}_h\left(\mu^{z'}_{Y(0)}, \nu^{z'}_{Y(1)}\right) \mathrm{d}\pi_{Z,Z'}(z, z')\right| \\
\le&\int \left|\mathbb{W}_h\left(\mu^z_{Y(0)}, \mu^z_{Y(1)}\right) - \mathbb{W}_h\left(\mu^{z'}_{Y(0)}, \nu^{z'}_{Y(1)}\right)\right| \mathrm{d}\pi_{Z,Z'}(z, z') \\
\le&\widetilde{C}(\epsilon) \int \mathbb{W}_1\left(\mu^z_{Y(0)}, \mu^{z'}_{Y(0)}\right) + \mathbb{W}_1\left(\mu^z_{Y(1)}, \nu^{z'}_{Y(1)}\right) \mathrm{d}\pi_{Z,Z'}(z, z') + 2\epsilon \\
\le&\widetilde{C}(\epsilon) \int (C(\epsilon')\|z - z'\|_2 + \epsilon') + \mathbb{W}_1\left(\mu^z_{Y(1)}, \nu^{z'}_{Y(1)}\right) \mathrm{d}\pi_{Z,Z'}(z, z') + 2\epsilon \\
\le&\widetilde{C}(\epsilon)(C(\epsilon') \vee 1)\mathbb{W}_{\mathrm{a}}(\mu_{Y(1),Z}, \nu_{Y(1),Z}) + 2\epsilon + \widetilde{C}(\epsilon)\epsilon'.
\end{aligned}$$

Here, the second inequality is due to (11); the third inequality is due to (12); the last inequality is due to the definition of π, which achieves the *optimality* in $\mathbb{W}_{\mathrm{a}}(\mu_{Y(1),Z}, \nu_{Y(1),Z})$.

For **(Term B)**: we have

$$\begin{aligned}
(\text{Term B}) =& \left| \int \mathbb{W}_h \left(\mu^z_{Y(0)}, \nu^z_{Y(1)} \right) \mathrm{d}\nu_Z(z) - \int \mathbb{W}_h \left(\nu^z_{Y(0)}, \nu^z_{Y(1)} \right) \mathrm{d}\nu_Z(z) \right| \\
\leq & \int \left| \mathbb{W}_h \left(\mu^z_{Y(0)}, \nu^z_{Y(1)} \right) - \mathbb{W}_h \left(\nu^z_{Y(0)}, \nu^z_{Y(1)} \right) \right| \mathrm{d}\nu_Z(z) \\
\leq & \widetilde{C}(\epsilon) \int \mathbb{W}_1(\mu^z_{Y(0)}, \nu^z_{Y(0)}) \mathrm{d}\nu_Z(z) + 2\epsilon,
\end{aligned}$$

where the last inequality is due to (11).

Now, we consider the optimal coupling $\tilde{\pi} \in \mathcal{P}(\mathcal{Y}^2 \times \mathcal{Z}^2)$ that is attained in $\mathbb{W}_{\mathrm{a}}(\mu_{Y(0),Z}, \nu_{Y(0),Z})$, then

$$\begin{aligned}
& \int \mathbb{W}_1(\mu^z_{Y(0)}, \nu^z_{Y(0)}) \mathrm{d}\nu_Z(z) \\
= & \int \mathbb{W}_1(\mu^{z'}_{Y(0)}, \nu^{z'}_{Y(0)}) \mathrm{d}\tilde{\pi}(z, z') \\
= & \int \mathbb{W}_1(\mu^{z'}_{Y(0)}, \nu^{z'}_{Y(0)}) - \mathbb{W}_1(\mu^{z'}_{Y(0)}, \mu^{z}_{Y(0)}) \mathrm{d}\tilde{\pi}(z, z') + \int \mathbb{W}_1(\mu^{z'}_{Y(0)}, \mu^{z}_{Y(0)}) \mathrm{d}\tilde{\pi}(z, z') \\
\leq & \int \mathbb{W}_1(\nu^{z'}_{Y(0)}, \mu^{z}_{Y(0)}) \mathrm{d}\tilde{\pi}(z, z') + \int \mathbb{W}_1(\mu^{z'}_{Y(0)}, \mu^{z}_{Y(0)}) \mathrm{d}\tilde{\pi}(z, z') \\
\leq & \int \mathbb{W}_1(\nu^{z'}_{Y(0)}, \mu^{z}_{Y(0)}) \mathrm{d}\tilde{\pi}(z, z') + C(\epsilon') \int \|z - z'\|_2 \, \mathrm{d}\tilde{\pi}(z, z') + \epsilon' \\
\leq & (C(\epsilon') \vee 1) \mathbb{W}_{\mathrm{a}}(\mu_{Y(0),Z}, \nu_{Y(0),Z}) + \epsilon'.
\end{aligned} \tag{13}$$

Here, the first inequality is due to the triangle inequality of the metric $\mathbb{W}_1$; the second inequality is due to (12); the last inequality is due to the definition of $\tilde{\pi}$.

Therefore, we get

$$(\text{Term B}) \leq \widetilde{C}(\epsilon)(C(\epsilon') \vee 1) \mathbb{W}_{\mathrm{a}}(\mu_{Y(1),Z}, \nu_{Y(1),Z}) + 2\epsilon + \widetilde{C}(\epsilon)\epsilon'.$$

Finally, combine (Term A) and (Term B), and we get

$$\begin{aligned}
& \left| V_{\mathrm{c}} - \int \mathbb{W}_h \left(\nu^z_{Y(0)}, \nu^z_{Y(1)} \right) \mathrm{d}\nu_Z(z) \right| \\
\leq & \widetilde{C}(\epsilon)(C(\epsilon') \vee 1)(\mathbb{W}_{\mathrm{a}}(\mu_{Y(0),Z}, \nu_{Y(0),Z}) + \mathbb{W}_{\mathrm{a}}(\mu_{Y(1),Z}, \nu_{Y(1),Z})) + 2(2\epsilon + \widetilde{C}(\epsilon)\epsilon').
\end{aligned}$$

Replace ν by the measures from the sequence of $(\nu_n, n \geq 1)$ and send $\epsilon', \epsilon \to 0$, we get the desired result. □

G Proofs of Section 4

Notation. In this section and the following, to simplify the notation, we will let $n_0 = n, n_1 = m$, and $\xi_0 = w, \xi_1 = \xi$.

G.1 Proof of Theorem 4

The proof of Theorem 4 requires several useful results from Backhoff et al. (2022). Since our definition of adapted empirical distribution is slightly different from that of Backhoff et al. (2022), to make the paper self-contained, we include proofs of the lemmas in Section I.

Remark 3 (Notation)**.** We introduce the following notation, which is aligned with Backhoff et al. (2022). For notational convenience, in this proof and the proofs of the associated lemmas and corollaries, we write $\hat{\boldsymbol{\mu}}^{z_0}_{Y(0)}$ to represent $\hat{\boldsymbol{\mu}}^{z_0}_{Y(0),n}$ and $\hat{\boldsymbol{\mu}}^{z_1}_{Y(1)}$ to represent $\hat{\boldsymbol{\mu}}^{z_1}_{Y(1),m}$. We also write $\hat{\pi}$ to represent $\hat{\pi}_{n,m}$. The constants, if not specified explicitly, are independent of n, m but may depend on d_Z, d_Y, L_Z, L_h.

Let Φ_r^n be the small cubes that partition $[0,1]^{d_Z}$ as defined in Definition 2, i.e.,

$$\Phi_r^n \triangleq \left\{(\varphi_r^n)^{-1}(\{x\}) : x \in \varphi_r^n([0,1]^{d_Z})\right\}.$$

For the sets $G \subseteq [0,1]^{d_Z}$, we define the conditional distribution of μ given G as

$$\mu_{Y(0)}^G = \frac{\int_G \mu_{Y(0)}^\zeta \mathrm{d}\mu_Z(\zeta)}{\mu_Z(G)}.$$

We also define the conditional distributions of $\hat{\mu}, \hat{\boldsymbol{\mu}}$ given G as

$$\hat{\mu}_{Y(0)}^G \triangleq \frac{1}{|I_G|}\sum_{i\in I_G}\delta_{Y_i(0)}, \quad \hat{\boldsymbol{\mu}}_{Y(0)}^G \triangleq \frac{1}{|\tilde{I}_G|}\sum_{i\in \tilde{I}_G}\delta_{Y_i(0)},$$

where $I_G \triangleq \{i \in [n] : Z_i(0) \in G\}, \tilde{I}_G \triangleq \{i \in [n] : \varphi_r^n(Z_i(0)) \in G\}$.

To prove the finite-sample complexity result, we introduce the following lemmas.

Lemma 6 (Variance of $\mathbb{W}_h$)**.** *Under Assp. 7, the following holds.*

$$\int \mathbb{W}_h(\mu_{Y(0)}^z, \mu_{Y(1)}^z)^2 \mathrm{d}\mu_Z(z) \le 16(L_h L_Z)^2 \int \|z\|_2^2 \mathrm{d}\mu_Z(z) + C_{\mu,h},$$

where $C_{\mu,h}$ is a constant that depends on μ, h.

Lemma 7 (Gap between $\hat{\boldsymbol{\mu}}$ and $\hat{\mu}$)**.** *Under Assp. 4, we have*

$$\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0)}, \hat{\mu}_{Z(0)}) \le \sqrt{d_Z} n^{-r}.$$

Lemma 8 (Gap between $\hat{\boldsymbol{\mu}}$ and μ)**.** *Under Assp. 4 and 8, we have*

$$\begin{aligned}\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Y(0)}^z, \mu_{Y(0)}^z)\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z) &\le \sum_{G\in\Phi_r^n} \hat{\boldsymbol{\mu}}_{Z(0)}(G)\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Y(0)}^G, \mu_{Y(0)}^G) + L_Z\sqrt{d_Z}n^{-r} \\ &= \sum_{G\in\Phi_r^n} \hat{\mu}_{Z(0)}(G)\mathbb{W}_1(\hat{\mu}_{Y(0)}^G, \mu_{Y(0)}^G) + L_Z\sqrt{d_Z}n^{-r}.\end{aligned}$$

Lemma 9 ((Backhoff et al., 2022, Lemma 3.4))**.** *The following inequalities hold.*

(i) Under Assp. 4 and 6, $\mathbb{E}[\mathbb{W}_1(\hat{\mu}_{Z(0)}, \mu_Z)] \le C_1 R_{d_Z}(n)$.

(ii) Under Assp. 4 and 6,

$$\mathbb{E}\left[\sum_{G\in\Phi_r^n} \hat{\mu}_{Z(0)}(G)\mathbb{W}_1(\hat{\mu}_{Y(0)}^G, \mu_{Y(0)}^G)\right] \le C_2 R_{d_Y}(n^{1-rd_Z}).$$

Here, $R_d(\cdot)$ is defined in (8), C_1 is a constant that depends on d_Z, and C_2 is a constant that depends on d_Y.

Proof of Theorem 4. We start with

$$|\hat{V}_{n,m} - V_{\mathrm{c}}| = \left|\int \mathbb{W}_h(\hat{\boldsymbol{\mu}}_{Y(0)}^{z_0}, \hat{\boldsymbol{\mu}}_{Y(1)}^{z_1})\mathrm{d}\hat{\pi}(z_0,z_1) - V_{\mathrm{c}}\right|. \qquad \text{(by Definition 4)}$$

For the integrand, we apply Lemma 5 and get

$$\left|\mathbb{W}_h(\hat{\boldsymbol{\mu}}_{Y(0)}^{z_0}, \hat{\boldsymbol{\mu}}_{Y(1)}^{z_1}) - \mathbb{W}_h(\mu_{Y(0)}^{z_0}, \mu_{Y(1)}^{z_1})\right| \le L_h(\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Y(0)}^{z_0}, \mu_{Y(0)}^{z_0}) + \mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Y(1)}^{z_1}, \mu_{Y(1)}^{z_1})). \tag{14}$$

As a result, we have

$$
\begin{aligned}
&|\hat{V}_{n,m} - V_{\mathrm{c}}| \\
=&\left| \int \mathbb{W}_h(\hat{\boldsymbol{\mu}}^{z_0}_{Y(0)}, \hat{\boldsymbol{\mu}}^{z_1}_{Y(1)}) - \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) \mathrm{d}\hat{\pi}(z_0, z_1) + \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) \mathrm{d}\hat{\pi}(z_0, z_1) - V_{\mathrm{c}} \right| \\
\leq& \int \left| \mathbb{W}_h(\hat{\boldsymbol{\mu}}^{z_0}_{Y(0)}, \hat{\boldsymbol{\mu}}^{z_1}_{Y(1)}) - \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) \right| \mathrm{d}\hat{\pi}(z_0, z_1) + \left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) \mathrm{d}\hat{\pi}(z_0, z_1) - V_{\mathrm{c}} \right| \\
\leq& L_h \underbrace{\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^{z_0}_{Y(0)}, \mu^{z_0}_{Y(0)}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0) + \int \mathbb{W}_1(\mu^{z_1}_{Y(1)}, \hat{\boldsymbol{\mu}}^{z_1}_{Y(1)}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(1)}(z_1)}_{\text{(Term A)}} \qquad \text{(by Equation (14))} \\
&+ \underbrace{\left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) \mathrm{d}\hat{\pi}(z_0, z_1) - V_{\mathrm{c}} \right|}_{\text{(Term B)}}. \qquad (15)
\end{aligned}
$$

For **(Term B)**, again, we apply Lemma 5 to the integrand and get

$$
\left| \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) - \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) \right| \leq L_h \mathbb{W}_1(\mu^{z_0}_{Y(1)}, \mu^{z_1}_{Y(1)}) \leq L_h L_Z \left\| z_0 - z_1 \right\|_2,
$$

where the second inequality is due to Assp. 8. Therefore, we have

$$
\begin{aligned}
&\text{(Term B)} \\
\leq& \left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) - \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) \mathrm{d}\hat{\pi}(z_0, z_1) \right| + \left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) \mathrm{d}\hat{\pi}(z_0, z_1) - V_{\mathrm{c}} \right| \\
\leq& \underbrace{L_h L_Z \int \left\| z_0 - z_1 \right\|_2 \mathrm{d}\hat{\pi}(z_0, z_1)}_{\text{(Term C)}} + \underbrace{\left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0) - V_{\mathrm{c}} \right|}_{\text{(Term D)}}. \qquad (16)
\end{aligned}
$$

For **(Term D)**, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0) - V_{\mathrm{c}} \right| \right] \\
\leq& \mathbb{E}\left[\left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) (\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0) - \mathrm{d}\hat{\mu}_{Z(0)}(z_0)) \right| \right] + \mathbb{E}\left[\left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) \mathrm{d}\hat{\mu}_{Z(0)}(z_0) - V_{\mathrm{c}} \right| \right].
\end{aligned}
$$

By Lemma 5 and Assp. 8, the mapping $z \mapsto \mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)})$ is $(2L_hL_Z)$-Lipschitz continuous. Then, by Lemma 7,

$$
\mathbb{E}\left[\left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) (\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0) - \mathrm{d}\hat{\mu}_{Z(0)}(z_0)) \right| \right] \leq 2L_hL_Z \mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0)}, \hat{\mu}_{Z(0)}) \leq 2L_hL_Z \sqrt{d_Z} n^{-r}.
$$

By Assp. 6, Lemma 6 and Markov's inequality,

$$
\begin{aligned}
\mathbb{E}\left[\left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) \mathrm{d}\hat{\mu}_{Z(0)}(z_0) - V_{\mathrm{c}} \right| \right] \leq& \left(16(L_hL_Z)^2 \int \|z\|_2^2 \mathrm{d}\mu_Z(z) + C_{\mu,h} \right)^{\frac{1}{2}} n^{-\frac{1}{2}} \\
\leq& \hat{C} n^{-r},
\end{aligned}
$$

where $r = \frac{1}{d_Z + 2 \vee d_Y} \leq \frac{1}{3}$.

Note that, the $n^{-\frac{1}{2}}$ above can be replaced by $(m \vee n)^{-\frac{1}{2}}$ since $Z(0), Z(1)$ are symmetric thus can be swapped in the derivation above.

For **(Term C)**, we apply Assp. 9. Finally, combine (Term C) and (Term D), and we arrive at

$$
\mathbb{E}[\text{(Term B)}] \leq C' (m \wedge n)^{-r}.
$$

For **(Term A)**, we present the upper bound for $\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^{z_0}_{Y(0)}, \mu^{z_0}_{Y(0)})\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0)$ and the remaining part can be derived similarly.

By Lemma 8,

$$\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^{z}_{Y(0)}, \mu^{z}_{Y(0)})\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z) \le \sum_{G\in\Phi^n_r} \hat{\mu}_{Z(0)}(G)\mathbb{W}_1(\hat{\mu}^{G}_{Y(0)}, \mu^{G}_{Y(0)}) + L_Z\sqrt{d_Z}n^{-r}.$$

Then, apply Lemma 9, we get

$$\begin{aligned}\mathbb{E}\left[\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^{z}_{Y(0)}, \mu^{z}_{Y(0)})\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z)\right] \le& \mathbb{E}\left[\sum_{G\in\Phi^n_r} \hat{\mu}_{Z(0)}(G)\mathbb{W}_1(\hat{\mu}^{G}_{Y(0)}, \mu^{G}_{Y(0)})\right] + L_Z\sqrt{d_Z}n^{-r}\\ \le& C_2 R_{d_Y}(n^{1-rd_Z}) + L_Z\sqrt{d_Z}n^{-r}.\end{aligned}$$

$R_{d_Y}(\cdot)$ is defined in (8), and C_2 is a constant that depends on d_Y and $\max_{y\in\mathcal{Y}}\|y\|_2$.

Therefore, we get

$$\mathbb{E}[(\text{Term A})] \le \tilde{C}(R_{d_Y}((m\wedge n)^{1-rd_Z}) + (m\wedge n)^{-r}).$$

Finally, combine (Term A) and (Term B), we get

$$\mathbb{E}\left[|\hat{V}_{n,m} - V_{\mathrm{c}}|\right] \le \bar{C}(R_{d_Y}((m\wedge n)^{1-rd_Z}) + (m\wedge n)^{-r}).$$

Let $r = \frac{1}{d_Z+2\vee d_Y}$, which minimizes the right-hand side on the order of $m\wedge n$, we get

$$\mathbb{E}\left[|\hat{V}_{n,m} - V_{\mathrm{c}}|\right] \le \begin{cases}\bar{C}(m\wedge n)^{-\frac{1}{d_Z+2\vee d_Y}}\log(m\wedge n) & \text{if } d_Y \ne 2,\\ \bar{C}(m\wedge n)^{-\frac{1}{d_Z+2\vee d_Y}} & \text{if } d_Y = 2.\end{cases}$$

□

Remark 4 (Choice of $\hat{\pi}_{n,m}$ in (6))**.** When $\hat{\pi}_{n,m}$ is the optimal coupling of the marginals, by Lemma 7 and the triangle inequality (Lemma 1),

$$\begin{aligned}&\int \|z_0 - z_1\|_2 \,\mathrm{d}\hat{\pi}(z_0, z_1)\\ =&\mathbb{E}\left[\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0)}, \hat{\boldsymbol{\mu}}_{Z(1)})\right]\\ \le&\mathbb{E}\left[\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0)}, \hat{\mu}_{Z(0)})\right] + \mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(0)}, \mu_Z)\right] + \mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(1)}, \mu_Z)\right] + \mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(1)}, \hat{\boldsymbol{\mu}}_{Z(1)})\right]\\ \le&2\sqrt{d_Z}(n^{-r} + m^{-r}) + \mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(0)}, \mu_Z)\right] + \mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(1)}, \mu_Z)\right].\end{aligned}$$

Then, by Lemma 9 (i), we get

$$\int \|z_0 - z_1\|_2 \,\mathrm{d}\hat{\pi}(z_0, z_1) \le 2\sqrt{d_Z}(n^{-r} + m^{-r}) + 2C_1 R_{d_Z}(n\wedge m).$$

Then, for $r = \frac{1}{d_Z+2\vee d_Y}$, we have

$$R_{d_Z}(n\wedge m) \le C'(n\wedge m)^{-\frac{1}{2\vee d_Z+1}} \le C'(n\wedge m)^{-r}.$$

Therefore, $\hat{\pi}_{n,m}$ in (6) satisfies Assp. 9 when $r = \frac{1}{d_Z+2\vee d_Y}$.

G.2 Proof of Theorem 5

Proof of Theorem 5. If the generating distribution of $Z(0), Z(1)$ are not equal to μ_Z, i.e., $\mu_{Z(0)} \neq \mu_Z$ and $\mu_{Z(1)} \neq \mu_Z$, then for $r = \frac{1}{d_Z + 2 \vee d_Y}$, according to the derivation in Remark 4, we have

$$\begin{aligned}
&\int \|z_0 - z_1\|_2 \, \mathrm{d}\hat{\pi}(z_0, z_1) \\
\leq & 2\sqrt{d_Z}(n^{-r} + m^{-r}) + \mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(0)}, \mu_Z)\right] + \mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(1)}, \mu_Z)\right] \\
\leq & 2\sqrt{d_Z}(n^{-r} + m^{-r}) + \mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(0)}, \mu_{Z(0)})\right] + \mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(1)}, \mu_{Z(1)})\right] + \mathbb{W}_1(\mu_{Z(0)}, \mu_Z) + \mathbb{W}_1(\mu_{Z(1)}, \mu_Z) \\
\leq & C'(n \wedge m)^{-r} + 2\epsilon.
\end{aligned}$$

Plug this condition back to the derivation in the proof of Theorem 4, specifically (16), we get the desired result. □

G.3 Proof of Theorem 3

Lemma 10 (Convergence of adapted empirical distribution)**.** *Under Assp. 4 - 8, then almost surely,*

$$\lim_{n\to\infty} \mathbb{W}_{\mathrm{a}}(\hat{\boldsymbol{\mu}}_{Y(0),Z(0),n}, \mu_{Y(0),Z}) = \lim_{m\to\infty} \mathbb{W}_{\mathrm{a}}(\hat{\boldsymbol{\mu}}_{Y(1),Z(1),m}, \mu_{Y(1),Z}) = 0.$$

Proof of Theorem 3. The proof strategy of Theorem 3 is similar to that of Theorem 4. By (15), we have

$$\begin{aligned}
&|\hat{V}_{n,m} - V_{\mathrm{c}}| \\
\leq & \underbrace{\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^{z_0}_{Y(0),n}, \mu^{z_0}_{Y(0)}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0),n}(z_0) + \int \mathbb{W}_1(\mu^{z_1}_{Y(1)}, \hat{\boldsymbol{\mu}}^{z_1}_{Y(1),m}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(1),m}(z_1)}_{\text{(Term A)}} \\
& + \underbrace{\left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) \mathrm{d}\hat{\pi}_{n,m}(z_0, z_1) - V_{\mathrm{c}} \right|}_{\text{(Term B)}}.
\end{aligned}$$

For **(Term A)**, using a similar argument for Eq. (13), we get

$$\text{(Term A)} \leq (L_Z \vee 1)(\mathbb{W}_{\mathrm{a}}(\hat{\boldsymbol{\mu}}_{Y(0),Z(0),n}, \mu_{Y(0),Z}) + \mathbb{W}_{\mathrm{a}}(\hat{\boldsymbol{\mu}}_{Y(1),Z(1),m}, \mu_{Y(1),Z})).$$

By Lemma 10, we get (Term A) converges to zero almost surely.

For **(Term B)**, by Assp. 4, 8, $(z_0, z_1) \mapsto \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)})$ is a bounded continuous function. Indeed, by Assp. 8 and Lemma 5,

$$\begin{aligned}
\mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) - \mathbb{W}_h(\mu^{z_0'}_{Y(0)}, \mu^{z_1'}_{Y(1)}) \leq & L_h(\mathbb{W}_1(\mu^{z_0}_{Y(0)}, \mu^{z_0'}_{Y(0)}) + \mathbb{W}_1(\mu^{z_1}_{Y(1)}, \mu^{z_1'}_{Y(1)})) \\
\leq & L_h(\|z_0 - z_0'\|_2 + \|z_1 - z_1'\|_2).
\end{aligned}$$

Further, by Assp. 4, we get that $\mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)})$ is bounded on $\mathcal{Z}$.

Then, since $\hat{\pi}_{n,m}$ weakly converges to $(\mathrm{id}, \mathrm{id})_{\#}\mu_Z$ a.s., we get

$$\begin{aligned}
&\lim_{n,m\to\infty} \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) \mathrm{d}\hat{\pi}_{n,m}(z_0, z_1) - V_{\mathrm{c}} \\
= & \lim_{n,m\to\infty} \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) \mathrm{d}\hat{\pi}_{n,m}(z_0, z_1) - \int \mathbb{W}_h(\mu^{z}_{Y(0)}, \mu^{z}_{Y(1)}) \mathrm{d}\mu_Z(z) = 0.
\end{aligned}$$

Thus, (Term B) converges to zero almost surely. The desired result is proved. □

G.4 Proof of Theorem 6

Lemma 11 (TV bound). *Suppose $Z, Z' \in \mathcal{Z}$, and $r = \sup_{z,z' \in \mathcal{Z}} \|z - z'\|_2 < \infty$. Then,*

$$\mathbb{W}_1(Z, Z') \le r \times d_{TV}(Z, Z') = \frac{r}{2} \int |\mathrm{d}\mu_Z - \mathrm{d}\mu_{Z'}|.$$

Lemma 12 (Linear functional gap). *Suppose that $(Z_i, 1 \le i \le n)$ are i.i.d. drawn from $\mu \in \mathcal{P}([0,1]^{d_Z})$, and let $\hat{\mu}_{Z(0),n} = \sum_{i=1}^n \hat{w}_i \delta_{Z_i}$, where $\hat{w}_i$ is defined as*

$$\hat{w}_i = \frac{(1 - \hat{e}(Z_i(0)))^{-1}}{\sum_{l=1}^n (1 - \hat{e}(Z_l(0)))^{-1}}, \ i \in [n].$$

Additionally, let $\tilde{\mu}_{Z(0),n} = \sum_{i=1}^n w_i \delta_{Z_i(0)}$, where w_i is defined as

$$w_i = \frac{\frac{\mathrm{d}\mu}{\mathrm{d}\mu_{|W=0}}(Z_i(0))}{\sum_{i=1}^n \frac{\mathrm{d}\mu}{\mathrm{d}\mu_{|W=0}}(Z_i(0))} \quad i \in [n].$$

Under Assp. 3 and 13, we have

$$\mathbb{E}[\mathbb{W}_1(\hat{\mu}_{Z(0),n}, \tilde{\mu}_{Z(0),n})] \le \frac{\sqrt{d_Z}}{2\delta^3} n^{-\frac{1}{2}} + \frac{\sqrt{d_Z} C_{\mathrm{w}}}{2\delta} n^{-r}.$$

Lemma 13 (Empirical Wasserstein convergence rate for reweighted empirical distribution). *Under the same condition of Lemma 12, we have*

$$\mathbb{E}[\mathbb{W}_1(\hat{\mu}_{Z(0),n}, \mu_{Z(0),n})] \le \delta^{-2} C R_{d_Z}(n) + \frac{\sqrt{d_Z}}{2\delta^3} n^{-\frac{1}{2}} + \frac{\sqrt{d_Z} C_{\mathrm{w}}}{2\delta} n^{-r},$$

where C is a constant depending on d_Z.

Lemma 14 (Discretization gap). *Under Assp. 11 and 12 with*

$$\hat{\boldsymbol{\mu}}_{Z(0),n} = \sum_{G \in \Phi_r^n} \frac{(1 - \hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}|}{\sum_G (1 - \hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}|} \delta_{\varphi_r^n(Z_i(0))},$$

and

$$\hat{\mu}_{Z(0),n} = \sum_{i=1}^n \hat{w}_i \delta_{Z_i(0)},$$

where $\hat{w}_i$ is defined as in Lemma 12, we have

$$\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0),n}, \hat{\mu}_{Z(0),n}) \le \sqrt{d_Z} \left(1 + \frac{L_e}{\eta^3}\right) n^{-r}.$$

Lemma 15 (Gap between $\hat{\boldsymbol{\mu}}$ and μ). *Under Assp. 4 and 8 with $\hat{\boldsymbol{\mu}}_{Z(0),n}$ defined as in Lemma 14, we have*

$$\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^z_{Y(0),n}, \mu^z_{Y(0)}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0),n}(z) \le \sum_{G \in \Phi_r^n} \hat{\boldsymbol{\mu}}_{Z(0),n}(G) \mathbb{W}_1(\hat{\boldsymbol{\mu}}^G_{Y(0),n}, \mu^G_{Y(0)}) + L_Z \sqrt{d_Z} n^{-r}.$$

Proof of Lemma 15. The proof follows exactly the same as the proof of Lemma 8, which holds as long as $\hat{\boldsymbol{\mu}}_{Z(0),n}$ is a probability measure supported on the cell centers $\{\varphi_r^n(G) : G \in \Phi_r^n\}$. □

We also introduce the following theorem adapted from (Backhoff et al., 2022, Lemma 3.3).

Lemma 16 (Conditional distribution of Y). *Under Assp. 10, conditionally on $\{|\{i : Z_i(0) \in G\}|, G \in \Phi_r^n\}$, for each G, the sample $(Y_i(0), i : Z_i(0) \in G)$ i.i.d. follows $\mu(\cdot | Z(0) \in G, W = 0)$.*

Proof of Theorem 6. Without loss of generality, due to the cross-fitting, we can assume that the random estimation of propensity score function $\hat{e}$ is independent of $((Y_i(0), Z_i(0)), i \in [n])$ and $((Y_j(1), Z_j(1)), j \in [m])$.

The proof strategy of Theorem 3 is similar to that of Theorem 4.

Recall that

$$\hat{\boldsymbol{\mu}}_{Y(0),Z(0),n} = \sum_{G \in \Phi_r^n} \frac{(1-\hat{e}(\varphi_r^n(G)))^{-1}}{\sum_G (1-\hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}|} \sum_{i: Z_i(0) \in G} \delta_{Y_i(0), \varphi_r^n(Z_i(0))},$$
$$\hat{\boldsymbol{\mu}}_{Y(1),Z(1),m} = \sum_{G \in \Phi_r^m} \frac{\hat{e}(\varphi_r^m(G))^{-1}}{\sum_G \hat{e}(\varphi_r^m(G))^{-1} |\{j : Z_j(1) \in G\}|} \sum_{j: Z_j(1) \in G} \delta_{Y_j(1), \varphi_r^m(Z_j(0))},$$

First, note that

$$\hat{\boldsymbol{\mu}}_{Z(0),n} = \sum_{G \in \Phi_r^n} \frac{(1-\hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}|}{\sum_G (1-\hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}|} \delta_{\varphi_r^n(Z_i(0))},$$
$$\hat{\boldsymbol{\mu}}_{Z(1),m} = \sum_{G \in \Phi_r^m} \frac{\hat{e}(\varphi_r^m(G))^{-1} |\{j : Z_j(1) \in G\}|}{\sum_G \hat{e}(\varphi_r^m(G))^{-1} |\{j : Z_j(1) \in G\}|} \delta_{\varphi_r^m(Z_j(0))}.$$

and $\hat{\pi}_{n,m}$ is an optimal coupling of $\hat{\boldsymbol{\mu}}_{Z(0),n}$ and $\hat{\boldsymbol{\mu}}_{Z(1),m}$ associated with $\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0),n}, \hat{\boldsymbol{\mu}}_{Z(1),m})$. For the conditional distributions, we have, for $G \in \Phi_r^n, G' \in \Phi_r^m$

$$\hat{\boldsymbol{\mu}}^G_{Y(0),n} = \frac{1}{|\{i : Z_i(0) \in G\}|} \sum_{i: Z_i(0) \in G} \delta_{Y_i(0)},$$
$$\hat{\boldsymbol{\mu}}^{G'}_{Y(1),m} = \frac{1}{|\{j : Z_j(1) \in G'\}|} \sum_{j: Z_j(1) \in G'} \delta_{Y_j(1)}.$$

Additionally, we let

$$\hat{\mu}_{Z(0),n} = \sum_{i=1}^n \hat{w}_i \delta_{Z_i(0)}, \quad \hat{\mu}_{Z(1),m} = \sum_{j=1}^m \hat{\xi}_j \delta_{Z_j(1)}.$$

and

$$\tilde{\mu}_{Z(0),n} = \sum_{i=1}^n w_i \delta_{Z_i(0)}, \quad \tilde{\mu}_{Z(1),m} = \sum_{j=1}^m \xi_j \delta_{Z_j(1)},$$

where for $i \in [n], j \in [m]$,

$$w_i = \frac{\frac{\mathrm{d}\mu}{\mathrm{d}\mu_{|W=0}}(Z_i(0))}{\sum_{i=1}^n \frac{\mathrm{d}\mu}{\mathrm{d}\mu_{|W=0}}(Z_i(0))}, \quad \xi_j = \frac{\frac{\mathrm{d}\mu}{\mathrm{d}\mu_{|W=1}}(Z_j(1))}{\sum_{j=1}^m \frac{\mathrm{d}\mu}{\mathrm{d}\mu_{|W=1}}(Z_j(1))}.$$

In the following, we adopt the same notation as in the proof of Theorem 4, omitting the dependence of n, m in subscripts and superscripts for notational simplicity.

By (15), we have

$$|\hat{V}_{n,m} - V_{\mathrm{c}}| \leq \underbrace{L_h \left(\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^{z_0}_{Y(0)}, \mu^{z_0}_{Y(0)}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0) + \int \mathbb{W}_1(\mu^{z_1}_{Y(1)}, \hat{\boldsymbol{\mu}}^{z_1}_{Y(1)}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(1)}(z_1) \right)}_{\text{(Term A)}}$$
$$+ \underbrace{\left| \int \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_1}_{Y(1)}) \mathrm{d}\hat{\pi}(z_0, z_1) - V_{\mathrm{c}} \right|}_{\text{(Term B)}}.$$

For **(Term B)**, by Eq.(16), we have

$$\begin{aligned}
&\text{(Term B)}\\
\leq&\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_1}_{Y(1)})-\mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})\mathrm{d}\hat{\pi}(z_0,z_1)\right|+\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})\mathrm{d}\hat{\pi}(z_0,z_1)-V_{\mathrm{c}}\right|\\
\leq&\underbrace{L_hL_Z\int\|z_0-z_1\|_2\,\mathrm{d}\hat{\pi}(z_0,z_1)}_{\text{(Term C)}}+\underbrace{\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0)-V_{\mathrm{c}}\right|}_{\text{(Term D)}}.
\end{aligned}$$

For **(Term C)**, similar to the derivation in Remark 4, we have

$$\begin{aligned}
&\int\|z_0-z_1\|_2\,\mathrm{d}\hat{\pi}(z_0,z_1)\\
=&\mathbb{E}\left[\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0)},\hat{\boldsymbol{\mu}}_{Z(1)})\right]\\
\leq&\mathbb{E}\left[\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0)},\hat{\mu}_{Z(0)})\right]+\mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(0)},\mu_Z)\right]+\mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(1)},\mu_Z)\right]+\mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(1)},\hat{\boldsymbol{\mu}}_{Z(1)})\right]\\
\leq&2\sqrt{d_Z}(1+\frac{L_e}{\eta^3})(n^{-r}+m^{-r})+\mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(0)},\mu_Z)\right]+\mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(1)},\mu_Z)\right],
\end{aligned}$$

where the first inequality is the triangle inequality (Lemma 1) and the second is due to Lemma 14.

Under Assp. 3 and 13, by Lemma 13, we get

$$\mathbb{E}\left[\mathbb{W}_1(\hat{\mu}_{Z(0)},\mu_Z)\right]\leq\delta^{-2}CR_{d_Z}(n)+\sqrt{d_Z}C_{\mathrm{w}}n^{-r}.$$

This is due to $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_{|W=0}}(z)=(1-e(z))^{-1}\mathbb{P}(W=0)\leq\delta^{-1}$ (Assp. 13). Combine these, we get

$$\text{(Term C)}\leq C_1'(R_{d_Z}(n\wedge m)+(n\wedge m)^{-r}).$$

For **(Term D)**, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0)-V_{\mathrm{c}}\right|\right]\\
\leq&\mathbb{E}\left[\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})(\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0)-\mathrm{d}\hat{\mu}_{Z(0)}(z_0))\right|\right]+\mathbb{E}\left[\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})\mathrm{d}\hat{\mu}_{Z(0)}(z_0)-V_{\mathrm{c}}\right|\right].
\end{aligned}$$

By Lemma 5 and Assp. 8, the mapping $z\mapsto\mathbb{W}_h(\mu^z_{Y(0)},\mu^z_{Y(1)})$ is $(2L_hL_Z)$-Lipschitz continuous. Then, by Lemma 14,

$$\begin{aligned}
\mathbb{E}\left[\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})(\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0)-\mathrm{d}\hat{\mu}_{Z(0)}(z_0))\right|\right]\leq&2L_hL_Z\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0)},\hat{\mu}_{Z(0)})\\
\leq&2L_hL_Z\sqrt{d_Z}(1+\frac{L_e}{\eta^3})n^{-r}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})\mathrm{d}\hat{\mu}_{Z(0)}(z_0)-V_{\mathrm{c}}\right|\right]\\
\leq&\underbrace{\mathbb{E}\left[\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})\mathrm{d}\hat{\mu}_{Z(0)}(z_0)-\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})\mathrm{d}\tilde{\mu}_{Z(0)}(z_0)\right|\right]}_{\text{(Term E)}}\\
&+\underbrace{\mathbb{E}\left[\left|\int \mathbb{W}_h(\mu^{z_0}_{Y(0)},\mu^{z_0}_{Y(1)})\mathrm{d}\tilde{\mu}_{Z(0)}(z_0)-V_{\mathrm{c}}\right|\right]}_{\text{(Term F)}}
\end{aligned}$$

By Lemma 12, and the fact that the mapping $z \mapsto \mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)})$ is $(2L_hL_Z)$-Lipschitz continuous, we get

$$(\text{Term E}) \leq 2L_hL_Z(\frac{\sqrt{d_Z}}{2\delta^3}n^{-\frac{1}{2}} + \frac{\sqrt{d_Z}C_{\mathrm{w}}}{2\delta}n^{-r}).$$

And by Assp. 2, Lemma 6 and Markov's inequality, we get

$$(\text{Term F}) \leq \delta^{-1}\left(16(L_hL_Z)^2\int\|z\|_2^2\mathrm{d}\mu_Z(z) + C_{\mu,h}\right)^{\frac{1}{2}}n^{-\frac{1}{2}} \leq C_2'n^{-r},$$

where $r = \frac{1}{d_Z+2\vee d_Y} \leq \frac{1}{3}$. Combine these, we get

$$(\text{Term D}) \leq C_3'(n\wedge m)^{-r}.$$

For **(Term A)**, we present the upper bound for $\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^{z_0}_{Y(0)}, \mu^{z_0}_{Y(0)})\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z_0)$ and the remaining part can be derived similarly.

Let $\nu = \mu_{Z|W=0}\otimes\mu^Z_{Y(0)}$, thus by Assp. 2, $\nu^z_{Y(0)} = \mu^z_{Y(0)}$. Then, by Lemma 8,

$$\begin{aligned}
&\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^z_{Y(0)}, \mu^z_{Y(0)})\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z)\\
=&\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^z_{Y(0)}, \nu^z_{Y(0)})\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z)\\
\leq&\sum_{G\in\Phi^n_r}\hat{\boldsymbol{\mu}}_{Z(0)}(G)\mathbb{W}_1(\hat{\boldsymbol{\mu}}^G_{Y(0)}, \nu^G_{Y(0)}) + L_Z\sqrt{d_Z}n^{-r}\\
=&\sum_{G\in\Phi^n_r}\hat{\boldsymbol{\mu}}_{Z(0)}(G)\mathbb{W}_1\left(\frac{1}{|\{i:Z_i(0)\in G\}|}\sum_{i:Z_i(0)\in G}\delta_{Y_i(0)}, \nu^G_{Y(0)}\right) + L_Z\sqrt{d_Z}n^{-r}.
\end{aligned}$$

Note that $(\hat{\boldsymbol{\mu}}_{Z(0)}(G), G\in\Phi^n_r)$ only depends on $\hat{e}$ and $\mathcal{G} = \{|\{i:Z_i(0)\in G\}|, G\in\Phi^n_r\}$, then apply Lemma 16, we get,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{G\in\Phi^n_r}\hat{\boldsymbol{\mu}}_{Z(0)}(G)\mathbb{W}_1\left(\frac{1}{|\{i:Z_i(0)\in G\}|}\sum_{i:Z_i(0)\in G}\delta_{Y_i(0)}, \nu^G_{Y(0)}\right)\right]\\
=&\mathbb{E}\left[\sum_{G\in\Phi^n_r}\hat{\boldsymbol{\mu}}_{Z(0)}(G)\mathbb{E}\left[\mathbb{W}_1\left(\frac{1}{|\{i:Z_i(0)\in G\}|}\sum_{i:Z_i(0)\in G}\delta_{Y_i(0)}, \nu^G_{Y(0)}\right)|\mathcal{G}, \hat{e}\right]\right]\\
\leq&C_4'\mathbb{E}\left[\sum_{G\in\Phi^n_r}\frac{(1-\hat{e}(\varphi^n_r(G)))^{-1}|\{i:Z_i(0)\in G\}|}{\sum_G(1-\hat{e}(\varphi^n_r(G)))^{-1}|\{i:Z_i(0)\in G\}|}R_{d_Y}(|\{i:Z_i(0)\in G\}|)\right]\\
\leq&\frac{C_4'\eta^{-1}}{n}\mathbb{E}\left[\sum_{G\in\Phi^n_r}|\{i:Z_i(0)\in G\}|R_{d_Y}(|\{i:Z_i(0)\in G\}|)\right]\\
=&\frac{C_4'\eta^{-1}|\Phi^n_r|}{n}\mathbb{E}\left[\frac{1}{|\Phi^n_r|}\sum_{G\in\Phi^n_r}|\{i:Z_i(0)\in G\}|R_{d_Y}(|\{i:Z_i(0)\in G\}|)\right]\\
\leq&C_4'\eta^{-1}R_{d_Y}(n^{1-rd_Z}),
\end{aligned}$$

where the first inequality is also due to Lemma 16 with independence between $\hat{e}$ and the sample Z, Y, the second is due to Assp. 12, and the last inequality is due to $x\mapsto xR_{d_Y}(x)$ is concave.

As a result,

$$\mathbb{E}\left[\int\mathbb{W}_1(\hat{\boldsymbol{\mu}}^z_{Y(0)}, \mu^z_{Y(0)})\mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z)\right] \leq C_4'\eta^{-1}R_{d_Y}(n^{1-rd_Z}) + L_Z\sqrt{d_Z}n^{-r}.$$

$R_{d_Y}(\cdot)$ is defined in (8), and C_4' is a constant that depends on d_Y and $\max_{y\in\mathcal{Y}}\|y\|_2$. Therefore, we get

$$\mathbb{E}[(\text{Term A})] \le C_5'(R_{d_Y}((m\wedge n)^{1-rd_Z}) + (m\wedge n)^{-r}).$$

Let $r = \frac{1}{d_Z+2\vee d_Y}$, the remaining steps follow the same as the proof of Theorem 4. □

H Proof of Section 5

Lemma 17 (V_c for Gaussian noise model, (Lin et al., 2025, Lemma 6.1))**.** *Assume that $h(y,\bar{y}) = \|y-\bar{y}\|_2^2$, and the noise variables follow the distribution $\varepsilon_k \sim N(0,\Sigma_k), k=0,1$. Then,*

(i) For the location model ($m=1$), $V_c = \mathbb{E}_Z\left[\|f_0(Z)-f_1(Z)\|_2^2\right] + S(\Sigma_0,\Sigma_1)$.

(ii) For the scale model ($m=2$), $V_c = \mathbb{E}_Z\left[S(\mathcal{D}(f_0(Z))\Sigma_0\mathcal{D}(f_0(Z)), \mathcal{D}(f_1(Z))\Sigma_1\mathcal{D}(f_1(Z)))\right]$.

Here, $S(\Sigma_0,\Sigma_1) = Tr\left(\Sigma_0+\Sigma_1-2(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}})^{\frac{1}{2}}\right)$, and $\mathcal{D}(v)$ is the matrix with v at the diagonal.

I Proofs of Supporting Lemmas

I.1 Proof of Lemma 1

Proof of Lemma 1. Let π ($\tilde{\pi}$, resp.) be an optimal coupling associated with $\mathbb{W}_1(\mu,\nu)$ ($\mathbb{W}_1(\nu,\mu')$, resp.). By the gluing lemma in Chapter 1 of Villani et al. (2009), there is a coupling $\gamma = \gamma(x,y,x')$ such that

$$\gamma_{xy} = \pi, \quad \gamma_{yx'} = \tilde{\pi}.$$

As a result, we have

$$\begin{aligned}
\mathbb{W}_1(\mu,\mu') \le& \int \|x-x'\|_2 \,\mathrm{d}\gamma(x,y,x') \\
\le& \int (\|x-y\|_2 + \|y-x'\|_2)\mathrm{d}\gamma(x,y,x') \\
=& \int \|x-y\|_2 \,\mathrm{d}\pi(x,y) + \int \|y-x'\|_2 \,\mathrm{d}\tilde{\pi}(y,x') \\
=& \mathbb{W}_1(\mu,\nu) + \mathbb{W}_1(\nu,\mu').
\end{aligned}$$

□

I.2 Proof of Lemma 2

Proof of Lemma 2. Let π ($\tilde{\pi}$, resp.) be an optimal coupling associated with $\mathbb{W}_1(\mu,\mu')$ ($\mathbb{W}_1(\nu,\mu')$, resp.). Then, for $\lambda\in[0,1]$, $\lambda\pi + (1-\lambda)\tilde{\pi}$ is a coupling for $\lambda\mu+(1-\lambda)\nu$ and μ'.

As a result, we have

$$\begin{aligned}
\mathbb{W}_1(\lambda\mu+(1-\lambda)\nu,\mu') \le& \int \|x-x'\|_2 \,\mathrm{d}(\lambda\pi + (1-\lambda)\tilde{\pi}) \\
=& \lambda\int \|x-x'\|_2 \,\mathrm{d}\pi + (1-\lambda)\int \|x-x'\|_2 \,\mathrm{d}\tilde{\pi} \\
=& \lambda\mathbb{W}_1(\mu,\mu') + (1-\lambda)\mathbb{W}_1(\nu,\mu').
\end{aligned}$$

□

I.3 Proof of Lemma 3

Proof of Lemma 3. Since $\Pi_c \subseteq \Pi(\mu_{Y(0)},\mu_{Y(1)})$ and $\Pi(\mu_{Y(0)},\mu_{Y(1)})$ is tight ((Villani et al., 2009, Lemma 4.4)), then Π_c is also tight. It remains to show that Π_c is closed with respect to the weak topology. Suppose that $\pi_k\in\Pi_c$ for $k\ge 1$ and $\lim_k \pi_k = \pi_\infty$, then for any bounded continuous function $\ell: \mathcal{Y}\times\mathcal{Z}\to\mathbb{R}$, $\mathbb{E}_{\pi_\infty}[\ell(Y(0),Z)] = \lim_k \mathbb{E}_{\pi_k}[\ell(Y(0),Z)] = \int \ell \mathrm{d}\mu_{Y(0),Z}$. Similarly, $\mathbb{E}_{\pi_\infty}[\ell(Y(1),Z)] = \int \ell\mathrm{d}\mu_{Y(1),Z}$. Therefore, $\pi_\infty\in\Pi_c$, thus Π_c is closed. □

I.4 Proof of Lemma 4

Proof of Lemma 4. On the compact space $\mathcal{Y}$, the weak topology can be induced by the metric $\mathbb{W}_1$, thus $g_w(z) = \mu^z_{Y(w)}$ $w = 0, 1$ are continuous with respect to $\mathbb{W}_1$. Since $\mathcal{Z}$ is compact, then $g_w(z)$ $w = 0, 1$ are uniformly continuous with respect to $\mathbb{W}_1$. As a result, we get that, for any $\delta > 0$, there is a constant $C(\delta)$ such that $\mathbb{W}_1\left(\mu^{z_0}_{Y(k)}, \mu^{z_1}_{Y(k)}\right) \leq \delta + C(\delta)\|z_0 - z_1\|_2$. □

I.5 Proof of Lemma 5

Proof of Lemma 5. We only need to prove that

$$|\mathbb{W}_h(\mu, \nu) - \mathbb{W}_h(\mu, \nu')| \leq L_h \mathbb{W}_1(\nu, \nu') + \epsilon.$$

If this holds, then the desired result is followed by applying the triangle inequality:

$$|\mathbb{W}_h(\mu, \nu) - \mathbb{W}_h(\mu', \nu')| \leq |\mathbb{W}_h(\mu, \nu) - \mathbb{W}_h(\mu, \nu')| + |\mathbb{W}_h(\mu, \nu') - \mathbb{W}_h(\mu', \nu')| + 2\epsilon. \tag{17}$$

In the following, we use the similar technique for proving the triangle inequality of Wasserstein distances.

Let $\pi_1 \in \mathcal{P}(\mathcal{Y}^2)$ be the optimal coupling associated with $\mathbb{W}_h(\mu, \nu)$, and let $\pi_2 \in \mathcal{P}(\mathcal{Y}^2)$ be the optimal coupling associated with $\mathbb{W}_1(\nu, \nu')$. By the gluing lemma in Chapter 1 of Villani et al. (2009), there is a coupling $\pi_3 \in \mathcal{P}(\mathcal{Y}^3)$, where we denote $\mathcal{Y}^3 = \{(y^{(1)}, y^{(2)}, y^{(3)}), y^{(k)} \in \mathcal{Y}\}$, such that

$$\begin{aligned}\pi_3(y^{(1)}, y^{(2)}) &= \pi_1,\\ \pi_3(y^{(2)}, y^{(3)}) &= \pi_2.\end{aligned}$$

Therefore, by Assp. 7, we have

$$\begin{aligned}\mathbb{W}_h(\mu, \nu') \leq& \mathbb{E}_{\pi_3}\left[h(y^{(1)}, y^{(3)})\right]\\ \leq& \mathbb{E}_{\pi_3}\left[h(y^{(1)}, y^{(2)}) + L_h\|y^{(2)} - y^{(3)}\|_2 + \epsilon\right]\\ =& \mathbb{W}_h(\mu, \nu) + L_h\mathbb{W}_1(\nu, \nu') + \epsilon.\end{aligned}$$

where the first inequality is due to $\pi_3(y^{(1)}, y^{(3)}) \in \Pi(z, z_1)$; the second inequality is due to the smoothness assumption of h (10); the equation is due to the construction of π_3 by gluing π_1, π_2.

Swapping ν and ν', we can the other direction of the inequality. Jointly, we get

$$|\mathbb{W}_h(\mu, \nu) - \mathbb{W}_h(\mu, \nu')| \leq L_h\mathbb{W}_1(\nu, \nu') + \epsilon.$$

Plug it back to (17), we get the desired result. □

I.6 Proof of Lemma 6

Proof of Lemma 6. Let π_z be the optimal coupling associated with $\mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)})$, then we have, for a point $z_0 \in \mathcal{Z}$,

$$\begin{aligned}&\int \mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)})^2 \mathrm{d}\mu_Z(z)\\ =& \int (\mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)}) - \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}) + \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}))^2 \mathrm{d}\mu_Z(z)\\ \leq& 2\int (\mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)}) - \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}))^2 + (\mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}))^2 \mathrm{d}\mu_Z(z)\\ =& \underbrace{2\int (\mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)}) - \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}))^2 \mathrm{d}\mu_Z(z)}_{\text{(Term A)}} + C.\end{aligned}$$

For **(Term A)**, by Lemma 5 and Assp. 8, we have

$$(\mathbb{W}_h(\mu^z_{Y(0)}, \mu^z_{Y(1)}) - \mathbb{W}_h(\mu^{z_0}_{Y(0)}, \mu^{z_0}_{Y(1)}))^2 \le 4(L_h L_Z)^2 \|z - z_0\|_2^2 \le 8(L_h L_Z)^2(\|z\|_2^2 + \|z_0\|_2^2).$$

Take it back, we get

$$\text{(Term A)} \le 16(L_h L_Z)^2 \int \|z\|_2^2 \mathrm{d}\mu_Z(z) + \tilde{C},$$

which proves the desired result. □

I.7 Proof of Lemma 7

Proof of Lemma 7. Note that $\hat{\boldsymbol{\mu}}_{Z(0)}$ is the push-forward measure of $\hat{\mu}_{Z(0)}$ under φ_r^n. Then,

$$\begin{aligned}
\mathbb{W}_1(\hat{\mu}_{Z(0)}, \hat{\boldsymbol{\mu}}_{Z(0)}) \le& \left(\int \|\zeta - \varphi_r^n(\zeta)\|_2 \, \mathrm{d}\hat{\mu}_{Z(0)}(\zeta) \right) \\
\le& \sup_{\zeta \in [0,1]^{d_Z}} \|\zeta - \varphi_r^n(\zeta)\|_2 && \text{(due to Assp. 4)} \\
\le& \sqrt{d_Z} n^{-r}.
\end{aligned}$$

□

I.8 Proof of Lemma 8

Lemma 18. *For any $G \in \Phi_r^n$, we have: (i) $\hat{\mu}^G_{Y(0)} = \hat{\boldsymbol{\mu}}^G_{Y(0)}$, and (ii) $\hat{\boldsymbol{\mu}}_{Z(0)}(G) = \hat{\mu}_{Z(0)}(G)$.*

Proof of Lemma 18. Note that, due to the construction of φ_r^n and $G \in \Phi_r^n$, we have $I_G = \tilde{I}_G$. As a result, we get $\hat{\mu}^G_{Y(0)} = \hat{\boldsymbol{\mu}}^G_{Y(0)}$, which proves (i). For (ii), note that $n\hat{\mu}_{Z(0)}(G) = |I_G| = |\tilde{I}_G| = n\hat{\boldsymbol{\mu}}_{Z(0)}(G)$. □

Proof of Lemma 8. We have

$$\int \mathbb{W}_1(\hat{\boldsymbol{\mu}}^z_{Y(0)}, \mu^z_{Y(0)}) \mathrm{d}\hat{\boldsymbol{\mu}}_{Z(0)}(z) = \sum_{\zeta \in \varphi_r^n([0,1]^{d_Z})} \mathbb{W}_1(\hat{\boldsymbol{\mu}}^\zeta_{Y(0)}, \mu^\zeta_{Y(0)}) \hat{\boldsymbol{\mu}}_{Z(0)}(\zeta). \tag{18}$$

Note that, in (18), we have $\hat{\boldsymbol{\mu}}^\zeta_{Y(0)} = \hat{\boldsymbol{\mu}}^G_{Y(0)}$, and $\hat{\boldsymbol{\mu}}_{Z(0)}(\zeta) = \hat{\boldsymbol{\mu}}_{Z(0)}(G)$ for $G = (\varphi_r^n)^{-1}(\zeta) \in \Phi_r^n$.

Then, by Lemma 18, we have

$$\begin{aligned}
\mathbb{W}_1(\hat{\boldsymbol{\mu}}^G_{Y(0)}, \mu^\zeta_{Y(0)}) \le& \mathbb{W}_1(\hat{\boldsymbol{\mu}}^G_{Y(0)}, \mu^G_{Y(0)}) + \mathbb{W}_1(\mu^G_{Y(0)}, \mu^\zeta_{Y(0)}) && \text{(by the triangle ineq.)} \\
\le& \mathbb{W}_1(\hat{\mu}^G_{Y(0)}, \mu^G_{Y(0)}) + \mathbb{W}_1(\mu^G_{Y(0)}, \mu^\zeta_{Y(0)}). && \text{(by Lemma 18)}
\end{aligned}$$

By the convexity of $\mu \mapsto \mathbb{W}_1(\mu, \mu^\zeta_{Y(0)})$ (Lemma 2) and Assp. 8, for $\zeta \in \varphi_r^n([0,1]^{d_Z})$, we have

$$\begin{aligned}
\mathbb{W}_1(\mu^G_{Y(0)}, \mu^\zeta_{Y(0)}) =& \mathbb{W}_1\left(\int_G \mu^{\zeta'}_{Y(0)} \mathrm{d}\mu_Z(\zeta'), \mu^\zeta_{Y(0)} \right) \\
\le& \int_G \mathbb{W}_1(\mu^{\zeta'}_{Y(0)}, \mu^\zeta_{Y(0)}) \mathrm{d}\mu_Z(\zeta') \\
\le& \int_G L_Z \|\zeta' - \zeta\|_2 \, \mathrm{d}\mu_Z(\zeta') && \text{(due to Assp. 8)} \\
\le& L_Z \sup_{\zeta' \in G} \|\zeta - \zeta'\|_2 . && \text{(due to Assp. 4)} \\
\le& L_Z \sqrt{d_Z} n^{-r}
\end{aligned}$$

Back to (18), by Lemma 18 that $\hat{\mu}_{Z(0)}(G) = \hat{\boldsymbol{\mu}}_{Z(0)}(G)$, we obtain

$$\begin{aligned}
&\sum_{\zeta \in \varphi_r^n([0,1]^{d_Z})} \mathbb{W}_1(\hat{\boldsymbol{\mu}}^{\zeta}_{Y(0)}, \mu^{\zeta}_{Y(0)}) \hat{\boldsymbol{\mu}}_{Z(0)}(\zeta) \\
&\leq \sum_{G \in \Phi_r^n} \hat{\boldsymbol{\mu}}_{Z(0)}(G) \left(\mathbb{W}_1(\hat{\mu}^G_{Y(0)}, \mu^G_{Y(0)}) + L_Z \sqrt{d_Z} n^{-r} \right) \\
&= \sum_{G \in \Phi_r^n} \hat{\boldsymbol{\mu}}_{Z(0)}(G) \mathbb{W}_1(\hat{\mu}^G_{Y(0)}, \mu^G_{Y(0)}) + L_Z \sqrt{d_Z} n^{-r} \\
&= \sum_{G \in \Phi_r^n} \hat{\mu}_{Z(0)}(G) \mathbb{W}_1(\hat{\mu}^G_{Y(0)}, \mu^G_{Y(0)}) + L_Z \sqrt{d_Z} n^{-r}. \qquad \text{(by Lemma 18)}
\end{aligned}$$

Plugging back to (18) proves the desired result. □

I.9 Proof of Lemma 11

Proof of Lemma 11. Note that $d_{\mathrm{TV}}(Z, Z') = \inf_{\lambda \in \Pi(\mathbb{Q}, \mathbb{P})} \mathbb{E}_\lambda [\mathbf{1}(Z \neq Z')]$. Then, since

$$\|Z - Z'\|_2 \leq r \times \mathbf{1}(Z \neq Z'),$$

then we get

$$\mathbb{W}_1(Z, Z') \leq r \times d_{\mathrm{TV}}(Z, Z').$$

□

I.10 Proof of Lemma 12

Proof of Lemma 12. By Lemma 11,

$$\mathbb{W}_1(\hat{\mu}_{Z(0),n}, \tilde{\mu}_{Z(0),n}) \leq \sqrt{d_Z} d_{\mathrm{TV}}(\hat{\mu}_{Z(0),n}, \tilde{\mu}_{Z(0),n}).$$

Then, we have

$$\begin{aligned}
&\mathbb{W}_1(\hat{\mu}_{Z(0),n}, \tilde{\mu}_{Z(0),n}) \\
\leq &\frac{\sqrt{d_Z}}{2} \sum_{i=1}^n \left| \hat{w}_i - \frac{\frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0))}{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0))} \right| \\
\leq &\frac{\sqrt{d_Z}}{2} \sum_{i=1}^n \left| \frac{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0)) \hat{w}_i - \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0))}{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0))} \right| \\
\leq &\frac{\sqrt{d_Z}}{2} \sum_{i=1}^n \left| \frac{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0)) - n}{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0))} \right| \hat{w}_i + \frac{\sqrt{d_Z}}{2} \sum_{i=1}^n \left| \frac{n\hat{w}_i - \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0))}{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0))} \right| \\
\leq &\underbrace{\frac{\sqrt{d_Z}}{2} \left| \frac{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0)) - n}{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0))} \right|}_{\text{(Term A)}} + \underbrace{\frac{\sqrt{d_Z}}{2\delta n} \sum_{i=1}^n \left| n\hat{w}_i - \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0)) \right|}_{\text{(Term B)}}.
\end{aligned}$$

Here, for the last inequality, we apply $\sum_{i=1}^n w_i = 1$ and $\frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(z) \geq \mathbb{P}(W=0)\ \forall z$, where $\mathbb{P}(W=0) = \int \mathbb{P}(W = 0|Z=z)\mathrm{d}\mu_Z(z) \geq \delta$ by Assp. 3.

For **(Term A)**, since for $x, y \geq \delta$, $x^{-1} - y^{-1} \leq \delta^{-2}|x-y|$, then we have

$$\begin{aligned}
\mathbb{E}[(\text{Term A})] &= \frac{\sqrt{d_Z}}{2}\mathbb{E}\left|1 - \frac{1}{\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0))}\right| \\
&\leq \frac{\sqrt{d_Z}}{2\delta^2}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i(0)) - 1\right| \\
&\leq \frac{\sqrt{d_Z}}{2\delta^2\sqrt{n}}\sqrt{\mathrm{Var}\left(\frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z)\right)} \\
&\leq \frac{\sqrt{d_Z}}{2\delta^3\sqrt{n}}.
\end{aligned}$$

Here, for the last inequality, we apply $\frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(z) = \mathbb{P}(W=0)/(1-e(z)) \leq \delta^{-1}$.

For **(Term B)**, we apply Assp. 13 and get

$$\mathbb{E}[(\text{Term B})] \leq \frac{\sqrt{d_Z}C_{\mathrm{w}}}{2\delta}n^{-r}.$$

Combine (Term A) and (Term B), we get the desired result. □

I.11 Proof of Lemma 13

Proof of Lemma 13. We write $Z_i(0)$ as Z_i in this proof. By the triangle inequality (Lemma 1),

$$\mathbb{E}[\mathbb{W}_1(\hat{\mu}_{Z(0),n}, \mu_{Z(0)})] \leq \underbrace{\mathbb{E}[\mathbb{W}_1(\hat{\mu}_{Z(0),n}, \tilde{\mu}_{Z(0),n})]}_{\text{(Term A)}} + \underbrace{\mathbb{E}[\mathbb{W}_1(\tilde{\mu}_{Z(0),n}, \mu_{Z(0)})]}_{\text{(Term B)}}.$$

For **(Term A)**, apply Lemma 12, we get

$$(\text{Term A}) \leq \frac{\sqrt{d_Z}}{2\delta^3}n^{-\frac{1}{2}} + \frac{\sqrt{d_Z}C_{\mathrm{w}}}{2\delta}n^{-r}.$$

For **(Term B)**, we follow the similar argument of the proof of (Fournier and Guillin, 2015, Theorem 1), and see that it depends on the bound of $\mathbb{E}[|\tilde{\mu}_{Z(0),n}(A) - \mu_{Z(0)}(A)|]$ for any subset A of $\mathbb{R}^{d_Z}$.

Indeed, we have,

$$\begin{aligned}
&\mathbb{E}[|\tilde{\mu}_{Z(0),n}(A) - \mu_{Z(0)}(A)|] \\
=&\mathbb{E}\left[\left|\sum_{i=1}^n w_i \delta_{Z_i}(A) - \mu_{Z(0)}(A)\right|\right] \\
=&\mathbb{E}\left[\left|\frac{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i)(\delta_{Z_i}(A) - \mu_{Z(0)}(A))}{\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i)}\right|\right] \\
\leq&\frac{1}{\mathbb{P}(W=0)}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i)(\delta_{Z_i}(A) - \mu_{Z(0)}(A))\right|\right] \\
\leq&\frac{1}{\sqrt{n}\mathbb{P}(W=0)}\left(\mathbb{E}_{Z\sim\mu}\left[\left(\frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z)\delta_Z(A) - \mu_{Z(0)}(A)\right)^2\right]\right)^{\frac{1}{2}} \qquad \text{(Jensen's inequality)} \\
\leq&\sqrt{\frac{\mu_{Z(0)}(A)}{n}} \cdot \frac{\delta^{-1}}{\mathbb{P}(W=0)}.
\end{aligned}$$

In addition, we have

$$\mathbb{E}\left[|\tilde{\mu}_{Z(0)}(A) - \mu_{Z(0)}(A)|\right] \leq \mathbb{E}\left[\tilde{\mu}_{Z(0)}(A)\right] + \mu_{Z(0)}(A) \leq \frac{2}{\mathbb{P}(W=0)} \cdot \mu_{Z(0)}(A),$$

where we apply the inequality $\mathbb{P}(W=0) \leq \frac{\mathrm{d}\mu_Z}{\mathrm{d}\mu_{Z|W=0}}(Z_i) \leq \delta^{-1}$.

Further, $\mathbb{P}(W=0) = \int \mathbb{P}(W=0|Z=z)\mathrm{d}\mu_Z(z) \geq \delta$ by Assp. 3.

Jointly, we have

$$\mathbb{E}[|\tilde{\mu}_{Z(0)}(A) - \mu_{Z(0)}(A)|] \leq \delta^{-2} \min\left(2\mu_{Z(0)}(A), \sqrt{\frac{\mu_{Z(0)}(A)}{n}}\right).$$

This inequality is similar to the first inequality in (Fournier and Guillin, 2015, Section 3), except that instead of δ^{-2}, they have 1 as the constant.

Then, following the same argument as in the proof of (Fournier and Guillin, 2015, Theorem 1), we get

$$\text{(Term B)} \leq \mathbb{E}[\mathbb{W}_1(\tilde{\nu}, \nu)] \leq C\delta^{-2} R_{d_Z}(n),$$

where C is a constant depending on d_Z.

Combining (Term A) and (Term B), we get the desired result. □

I.12 Proof of Lemma 14

Proof of Lemma 14. Let

$$W_G = \frac{(1-\hat{e}(\varphi_r^n(G)))^{-1}|\{i : Z_i(0) \in G\}|}{\sum_G (1-\hat{e}(\varphi_r^n(G)))^{-1}|\{i : Z_i(0) \in G\}|},$$

then by the triangle inequality (Lemma 1), we have

$$\begin{aligned}
&\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0),n}, \hat{\mu}_{Z(0),n}) \\
=&\mathbb{W}_1\Big(\sum_{G\in\Phi_r^n} W_G \delta_{\varphi_r^n(Z_i(0))}, \sum_{i=1}^n \hat{w}_i \delta_{Z_i(0)}\Big) \\
\leq&\mathbb{W}_1\Big(\sum_{G\in\Phi_r^n} W_G \delta_{\varphi_r^n(Z_i(0))}, \sum_{G\in\Phi_r^n} \frac{W_G}{|\{i : Z_i(0)\in G\}|}\sum_{i\in G}\delta_{Z_i(0)}\Big) \\
&+ \mathbb{W}_1\Big(\sum_{G\in\Phi_r^n} \frac{W_G}{|\{i : Z_i(0)\in G\}|}\sum_{i\in G}\delta_{Z_i(0)}, \sum_{i=1}^n \hat{w}_i\delta_{Z_i(0)}\Big) \\
\leq&\sqrt{d_Z}n^{-r} + \underbrace{\mathbb{W}_1\Big(\sum_{G\in\Phi_r^n} \frac{W_G}{|\{i : Z_i(0)\in G\}|}\sum_{i\in G}\delta_{Z_i(0)}, \sum_{i=1}^n \hat{w}_i\delta_{Z_i(0)}\Big)}_{\text{(Term A)}},
\end{aligned}$$

where the last inequality is due to Lemma 7.

For **(Term A)**, by Lemma 11, we have

$$\begin{aligned}
\text{(Term A)} \leq & \frac{\sqrt{d_Z}}{2} \sum_{G\in\Phi_r^n} \sum_{i\in G} \left| \frac{W_G}{|\{i : Z_i(0) \in G\}|} - \hat{w}_i \right| \\
\leq & \frac{\sqrt{d_Z}}{2} \sum_{G\in\Phi_r^n} \sum_{i\in G} \Bigg(\underbrace{\frac{(1-\hat{e}(\varphi_r^n(G)))^{-1} - (1-\hat{e}(Z_i(0)))^{-1}}{\sum_G (1-\hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}|}}_{\text{(Term B)}_i} \\
& + \underbrace{\frac{(1-\hat{e}(Z_i(0)))^{-1}}{\sum_G (1-\hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}|} - \frac{(1-\hat{e}(Z_i(0)))^{-1}}{\sum_{i=1}^n (1-\hat{e}(Z_i(0)))^{-1}}}_{\text{(Term C)}_i} \Bigg).
\end{aligned}$$

For $(\textbf{Term B})_i$, by Assp. 11 and 12, we have

$$|(1-\hat{e}(\varphi_r^n(G)))^{-1} - (1-\hat{e}(Z_i(0)))^{-1}| \leq \frac{L_e}{\eta^2} n^{-r}, \quad \sum_G (1-\hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}| \geq n.$$

Thus,

$$\sum_{G\in\Phi_r^n} \sum_{i\in G} |(\text{Term B})_i| \leq \frac{L_e}{\eta^2} n^{-r}.$$

For $(\textbf{Term C})_i$, by the same reasoning above, we have

$$\begin{aligned}
|(\text{Term C})_i| &= \frac{|\sum_{i=1}^n (1-\hat{e}(Z_i(0)))^{-1} - \sum_G (1-\hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}||}{(\sum_G (1-\hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}|)(\sum_{i=1}^n (1-\hat{e}(Z_i(0)))^{-1})} \\
&\leq \frac{n \frac{L_e}{\eta^2} n^{-r}}{(\sum_G (1-\hat{e}(\varphi_r^n(G)))^{-1} |\{i : Z_i(0) \in G\}|)(\sum_{i=1}^n (1-\hat{e}(Z_i(0)))^{-1})} \leq \frac{1}{n} \frac{L_e}{\eta^3} n^{-r}.
\end{aligned}$$

Thus, we get

$$\sum_{G\in\Phi_r^n} \sum_{i\in G} |(\text{Term C})_i| \leq \frac{L_e}{\eta^3} n^{-r}.$$

Combine these, we get

$$\mathbb{W}_1(\hat{\boldsymbol{\mu}}_{Z(0),n}, \hat{\mu}_{Z(0),n}) \leq \sqrt{d_Z}(1 + \frac{L_e}{\eta^3}) n^{-r}.$$

□