Statistical Inference in Causal Partial Identification under Smooth Densities

Sirui Lin[1*] **Zijun Gao**[2,*] **Jose 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

Many causal quantities are only partially identifiable due to the inherent missingness of potential outcomes, and the associated partial identification (PI) sets can be obtained by solving an optimal transport (OT) problem. Covariates often provide additional information about the potential outcomes and thus yield tighter PI sets, which can be obtained via conditional optimal transport (COT). However, COT-based PI set estimators are susceptible to the curse of dimensionality in the covariates, which precludes the asymptotic normality and hinders statistical inference. In this paper, we exploit the smoothness in the marginal densities of covariates and potential outcomes, and develop a wavelet-based primal approach for COT that attains a faster convergence rate. Moreover, for quadratic cost functions, we establish a stability result for COT and prove asymptotic normality of the proposed estimator, enabling valid statistical inference for the PI set. Empirically, we validate the estimation and inference performance of our approach through numerical experiments in comparison with existing benchmarks.

1 Introduction

We study statistical inference for causal estimands that are only partially identifiable under the potential outcome framework. The object of interest is a partial identification (PI) set, whose endpoints can be characterized through an optimal transport (OT) formulation [Gao et al., 2025]. When covariates are available, conditional optimal transport (COT) can further tighten these PI sets by conditioning on the covariate distribution, making it a natural framework for inference in modern observational and experimental datasets [Ji et al., 2023, Lin et al., 2026].

Partial identification sets are central to causal inference because they quantify the uncertainty induced by the unobservability of counterfactual outcomes. Equivalently, they describe the range of causal estimands compatible with the identified marginal distributions of potential outcomes. Their widths directly inform decision-making in policy evaluation, fairness assessment, and treatment-effect analysis. In particular, the expectation of the squared difference between potential outcomes is of special interest, as it connects to several important quantities in causal inference: the variance lower bound of the Neymanian estimator [Aronow et al., 2014], causal ANOVA [Gao and Zhao, 2025], and the cross-world correlation between potential outcomes [Bodik et al., 2025].

The study of partial identification for counterfactual outcomes was pioneered by Manski [1990, 1997], Manski and Pepper [2000], who developed instrumental-variable-based bounds under various assumptions; a large subsequent literature, surveyed by Tamer [2010], formulates PI problems through moment inequalities. More recently, Tan et al. [2026] studied partial identification of policy-relevant treatment effects in generalized Roy models [Heckman and Vytlacil, 2005], where limited instrument

*Equal contribution. Correspondence: Zijun Gao (zijungao@marshall.usc.edu).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

support makes many policy counterfactuals only partially identifiable. Their work sharpens this earlier PI literature by imposing distributional constraints directly on a joint conditional law, rather than relying on moment-inequality relaxations. Although their target differs from the joint distribution of potential outcomes considered here, their results further highlight COT as a language for sharp PI analysis. We complement this line of work by developing statistically efficient estimation and inference tools for COT under smooth density assumptions.

Beyond causal inference, COT is also of independent interest in OT theory as a conditional analogue of Wasserstein distances, closely related to adapted Wasserstein metrics studied in stochastic and dynamic settings [Backhoff et al., 2022, Eckstein and Pammer, 2024]. This perspective connects causal PI inference with statistical OT [Villani, 2008, Panaretos and Zemel, 2019, Niles-Weed and Berthet, 2022, Hütter and Rigollet, 2021, Manole et al., 2024, Hundrieser et al., 2024] and with distributional robustness and model-risk formulations, where sharp guarantees arise by optimizing over ambiguity sets defined by distributional constraints [Mohajerin Esfahani and Kuhn, 2018, Kuhn et al., 2019, Blanchet and Murthy, 2019, Xu et al., 2020, Nguyen et al., 2024]. Here, however, the ambiguity set is induced by the unobserved joint law of potential outcomes, making inference for causal PI sets the central statistical problem. Although motivated by causal PI, our estimation and inference tools apply more generally to COT problems with smooth conditional laws and covariate-level heterogeneity.

Current approaches to COT-based PI sets face substantial limitations. Dual formulations can yield valid confidence sets but may be conservative and require strong, hard-to-verify assumptions on potential functions [Ji et al., 2023, Al-Jarrah et al., 2025]. Penalization and discretization approaches reduce COT to OT at the cost of conservative or slow convergence rates and an acute curse of dimensionality in the covariates [Manupriya et al., 2024, Lin et al., 2025, 2026]. Quantile regression can provide asymptotically exact confidence sets but is limited to scalar outcomes [Fan and Park, 2010]. Sampling-based methods learn transport maps or samplers rather than the optimal objective value that defines the PI set [Hosseini et al., 2025]. Overall, existing methods do not provide statistically efficient estimators of the COT value, together with valid inference, for multivariate outcomes without suffering from the curse of dimensionality in high-dimensional covariates.

We address these challenges with a smoothness-based, primal COT estimator built on wavelet expansions of the observable marginal densities, inspired by recent advances in statistical OT [Hütter and Rigollet, 2021, Niles-Weed and Berthet, 2022, Manole et al., 2024]. The estimator exploits density smoothness, a more directly verifiable condition in causal settings than regularity of unobserved Kantorovich potentials required by dual approaches [Ji et al., 2023, Al-Jarrah et al., 2025]. Under smoothness assumptions, we establish faster convergence rates for general Lipschitz costs, and for quadratic costs we prove a COT stability result that leads to asymptotic normality and valid inference.

Contributions. Our contributions span causal PI inference and COT theory. On the causal side, we provide an inferential procedure with improved rates under Lipschitz costs and asymptotically normal confidence sets under quadratic costs; to our knowledge, this is the first PI inference approach via COT that delivers asymptotic normality for multivariate outcomes under smoothness. On the OT side, we establish a stability-based central limit theorem for conditional transport under quadratic costs, extending smooth OT results to a setting not covered by existing OT CLTs. A comparison of rates and assumptions with prior COT estimators appears in Table 1; the main contributions are:

1. We develop a primal, smoothness-based COT estimator (Definition 3.1) using wavelet expansions of observable marginal densities, complementing dual approaches that rely on potential smoothness [Ji et al., 2023, Al-Jarrah et al., 2025]. Under smoothness conditions, we establish convergence rates for general Lipschitz costs on average (Theorem 3.6) and under high probability (Theorem 3.7). We also extend the convergence result to the setting with covariate-dependent treatment (Theorem 3.14), which typically arises in observational studies.

2. For quadratic costs, we generalize OT stability to the COT setting (Proposition 3.8) and leverage it to prove asymptotic normality (Theorem 3.9) and conduct valid statistical inference for causal PI sets.

3. We provide empirical evidence that the proposed estimator improves the estimation accuracy over existing COT benchmarks and asymptotically controls the type I error for inference.

Notations. We denote $n \wedge m = \min(n, m), n \vee m = \max(n, m)$. We denote by $\mathcal{P}(\Omega)$ the set of probability measures on Ω. We use $\mathcal{P}_{\mathrm{ac}}(\Omega)$, for $\Omega \subseteq \mathbb{R}^d$, to denote the set of probability measures that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. d_X denotes the dimension of X. We typically use X to denote $X := (Y, Z)$. 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 $Z = z$. $\|g\|_{L^2(\mu)} := (\mathbb{E}_{X\sim\mu}[g(X)^2])^{1/2}$. By $\mu = \mu_Z \otimes \mu_Y^Z$, we mean that for any measurable function g, $\int g \mathrm{d}\mu = \int_{\mathcal{Z}} \int_{\mathcal{Y}} g(z, y) \mathrm{d}\mu_Y^z(y) \mathrm{d}\mu_Z(z)$. Given a density function $f(y, z)$, we denote its marginal density of z by $f_Z(z)$, and the conditional density of y given z by $f(y \mid z) = f(y, z)/f_Z(z)$. Given a measurable objective function h and probability distributions $P, Q \in \mathcal{P}(\mathbb{R}^{d_X})$, the optimal transport distance is defined as

$$\mathsf{W}_h(P, Q) := \min_{\pi\in\Pi(P,Q)} \mathbb{E}_\pi[h(X, X')],$$

where $\Pi(\mu, \nu) := \left\{\pi \in \mathcal{P}(\mathcal{X}^2) : \pi_X = \mu,\ \pi_{X'} = \nu\right\}$ and π_X denotes the marginal of π on X. When $h(x, x') = \|x - x'\|_2^p$, we denote $\mathsf{W}_p(P, Q) := \mathsf{W}_h(P, Q)^{\frac{1}{p}}$, which is the Wasserstein p-distance.

2 Preliminaries

2.1 PI under Potential Outcome Model and Conditional Optimal Transport

Consider $2n$ units. Each unit has potential outcomes $Y_i(0)$ under control and $Y_i(1)$ under treatment, with $Y_i(0), Y_i(1) \in [0, 1]^{d_Y}$; see, e.g., Rubin [1974]. We allow $d_Y \geq 1$, so outcomes may be vector-valued. Each unit also has covariates $Z_i \in [0, 1]^{d_Z}$, with $d_Z \geq 1$. We assume the super-population model $(Y_i(0), Y_i(1), Z_i) \sim_{\text{i.i.d.}} \mu$, where μ is an unknown joint distribution.

Each unit receives a binary treatment $W_i \in \{0, 1\}$, and we observe only $Y_i = Y_i(W_i)$. For simplicity, we focus on a completely randomized design with exactly n treated and n control units; the analysis can be extended to assignment mechanisms satisfying unconfoundedness and overlap [Imbens and Rubin, 2015]. Under this design, the marginal laws of $(Y_i(0), Z_i)$ and $(Y_i(1), Z_i)$ are identifiable from the observed data. We denote them by P and Q, respectively, and write p and q for their densities when they exist. After relabeling, the control sample $\mathcal{S}_0$ is an i.i.d. sample from P, and the treatment sample $\mathcal{S}_1$ is an i.i.d. sample from Q, both of size n.

Although P and Q are identifiable, the joint law μ is not, because $Y_i(0)$ and $Y_i(1)$ are never observed together. Consequently, causal functionals of the form $\mathbb{E}_\mu[h(Y(0), Y(1))]$ are generally only partially identifiable, and their compatible values form a PI set. We focus on its lower endpoint, which can be written as the conditional optimal transport problem

$$V_{\mathrm{c}} = \mathsf{CW}_h(P, Q) := \min_{\pi\in\Pi_{\mathrm{c}}(P,Q)} \mathbb{E}_\pi\big[h(Y(0), Y(1))\big],$$

where $h : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a pre-specified objective function and

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

Thus, $\Pi_{\mathrm{c}}(P, Q)$ is the set of joint laws of $(Y(0), Y(1), Z)$ compatible with the observed marginals. The upper endpoint is $-\mathsf{CW}_{-h}(P, Q)$.

Example 2.1 (Quadratic cost) *Consider the quadratic cost $h(y_0, y_1) = \|y_0 - y_1\|_2^2$ with $d_Y \geq 1$. In the scalar case, since the average treatment effect is identifiable, sharp bounds on this quantity directly imply sharp bounds on the variance of treatment effects. This variance is a classical unidentified component in the Neymanian variance formula for the difference-in-means estimator. More broadly, quadratic functionals of counterfactual differences appear in recent work on conservative inference for counterfactuals [Balakrishnan et al., 2025], causal ANOVA [Gao and Zhao, 2025], and cross-world dependence between potential outcomes [Bodik et al., 2025]. For vector-valued outcomes, the same quadratic cost provides a sharp bound on $\mathrm{tr}\{\mathrm{Var}(Y(1) - Y(0))\}$, an aggregate measure of treatment-effect heterogeneity across multiple outcomes.*

2.2 Addressing Curse of Dimensionality with Wavelet Density Estimator

Although the sharp PI bound V_{c} is defined by a COT problem over the identifiable marginals, directly plugging in their empirical counterparts leads to poor statistical behavior. Empirical plug-in estimation

of OT values suffers from the curse of dimensionality, even without covariates [Fournier and Guillin, 2015]; e.g., replacing P and Q by empirical distributions typically yields error of order n^{-1/d_Y}, which deteriorates quickly with the outcome dimension. For COT, existing plug-in approaches also depend on the covariate dimension with rates of order $n^{-1/(d_Y+d_Z)}$ [Lin et al., 2026]. In this paper, we mitigate this curse of dimensionality by exploiting smoothness of the densities. Our estimator is based on wavelet expansions, so we first introduce the relevant Besov spaces.

Definition 2.2 (Besov space) *Let $s > 0$ and $p, q \in [1, \infty]$. Let $\Psi = (\Phi, \Psi_1, \Psi_2, \ldots)$ be a wavelet system on $[0,1]^d$, where Φ denotes the collection of scaling functions and Ψ_j the collection of wavelet functions at resolution level j. Let $j_0 \geq 0$ denote a fixed coarse resolution level.*

For $f \in L^p([0,1]^d)$ admitting the wavelet expansion

$$f(x) = \sum_{\zeta \in \Phi} \theta_\zeta \, \zeta(x) + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j} \theta_\xi \, \xi(x), \text{ where } \theta_g := \int_{[0,1]^d} g(x) f(x) \, dx, \; g \in \Phi \cup \bigcup_j \Psi_j,$$

the Besov norm is defined as

$$\|f\|_{\mathcal{B}^s_{p,q}([0,1]^d)} := \|(\theta_\zeta)_{\zeta \in \Phi}\|_{\ell_p} + \left\| \left(2^{j(s+d/2-d/p)} \|(\theta_\xi)_{\xi \in \Psi_j}\|_{\ell_p} \right)_{j \geq j_0} \right\|_{\ell_q}.$$

The Besov space is defined as $\mathcal{B}^s_{p,q}([0,1]^d) := \left\{ f \in L^p([0,1]^d) : \|f\|_{\mathcal{B}^s_{p,q}([0,1]^d)} < \infty \right\}$.

The Besov norm can be extended to negative smoothness indices $s < 0$ via duality:

$$\mathcal{B}^s_{p',q'}([0,1]^d) := (\mathcal{B}^{-s}_{p,q}([0,1]^d))^*, \text{ where } p'^{-1} + p^{-1} = 1, \; q'^{-1} + q^{-1} = 1.$$

In this paper, we choose the boundary-corrected wavelet system on $[0,1]^{d_Y+d_Z}$ and defer the details of this wavelet system to Section C. The usefulness of wavelet expansions in our analysis is driven by the following stability result, which motivates the wavelet-based density estimator introduced next.

Proposition 2.3 (Niles-Weed and Berthet [2022, Theorem 4]) *Suppose that $\widehat{\mu}, \mu \in \mathcal{P}_{\text{ac}}([0,1]^d)$ with densities $\hat{f}, f \in L^2([0,1]^d)$. If $f(x) \geq \gamma^{-1} \, \forall x \in [0,1]^d$ for some $\gamma > 0$, then there is a universal constant $C_0 > 0$ such that $\mathsf{W}_2(\widehat{\mu}, \mu) \leq C_0 \gamma^{\frac{1}{2}} \|\hat{f} - f\|_{\mathcal{B}^{-1}_{2,1}([0,1]^d)}$.*

Definition 2.4 (Wavelet density estimator) *Let $J_n \geq j_0$ be a resolution level. Given i.i.d. observations $(X_i)_{i=1}^n$ with density f, the boundary-corrected wavelet density estimator is defined as*

$$\widehat{f}_{J_n}(x) = \frac{\widetilde{f}_{J_n}(x) \, \mathbf{1}\{\widetilde{f}_{J_n}(x) \geq 0\}}{\int_{\widetilde{f}_{J_n}(x) \geq 0} \widetilde{f}_{J_n}(x) \, dx}, \text{ where } \widetilde{f}_{J_n}(x) := \sum_{\zeta \in \Phi} \widehat{\theta}_\zeta \, \zeta(x) + \sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi_j} \widehat{\theta}_\xi \, \xi(x), \tag{1}$$

and $\widehat{\theta}_\zeta := \frac{1}{n} \sum_{i=1}^n \zeta(X_i), \widehat{\theta}_\xi := \frac{1}{n} \sum_{i=1}^n \xi(X_i)$.

Specifically, in this work, we take $J_n = \lfloor \frac{\log_2(n)}{2s+d_Y+d_Z} \rfloor$ and $j_0 \geq \lfloor \log_2(s/0.18 + 1) \rfloor + 1$ (which is standard as discussed in Manole et al. [2024, Section A.2]).

2.3 Assumptions

Assumption 2.5 (Bounded density) *Assume $P, Q \in \mathcal{P}_{\text{ac}}([0,1]^{d_Y+d_Z})$, $\gamma^{-1} \leq p(y,z), q(y,z) \leq \gamma$ for all $(y,z) \in [0,1]^{d_Y+d_Z}$ and some $\gamma \geq 1$.*

Assumption 2.6 (Smoothness) *Assume that, for a fixed smoothness parameter $s \geq 1$, the densities $p, q \in \mathcal{C}^{s+\epsilon}([0,1]^{d_Y+d_Z})$ for some $\epsilon > 0$. $\mathcal{C}^s$ denotes the s-order Hölder smoothness; see Section C.*

Assumption 2.7 (Lipschitz objective) *The objective function h is L_h-Lipschitz.*

Assumption 2.5 postulates that the density is bounded away from zero, a standard condition in nonparametric OT. When this assumption is violated, the convergence rate of smoothness-based OT estimators changes qualitatively [Niles-Weed and Berthet, 2022, Section 5]. Assumption 2.6 enables a wavelet-based expansion of the density and is crucial for constructing estimators that are statistically efficient. For technical reasons, we require the density to be slightly smoother than order s. Assumption 2.7 guarantees stability of the optimal transport value with respect to perturbations of the input measures; see [Lin et al., 2026, Assumption 7].

Table 1: Comparison of statistical convergence rates for COT value estimation.

Reference	Estimation Rate	Key Assumptions
[Lin et al., 2026]	$n^{-\frac{1}{2\vee d_Y+d_Z}}$	Lipschitz kernel
[Ji et al., 2023]	$n^{-1/2}$	Density, optimal dual are estimated at rate $O(n^{-1/4})$
This work	$n^{-\frac{s}{2s+d_Y+d_Z}}$	Densities in $\mathcal{C}^{s+\epsilon}$, bounded away from 0
This work	$n^{-(\frac{2s}{2s+d_Y+d_Z}\wedge\frac{1}{2})}$	Densities in $\mathcal{C}^{s+\epsilon}$, bounded away from 0, h quadratic

3 Main Results

3.1 Wavelet-Based COT Estimator

Definition 3.1 (Wavelet-based COT estimator) *Given i.i.d. samples $\mathcal{S}_0, \mathcal{S}_1$ drawn from P, Q respectively, let the resolution level be $J_n = \lfloor \frac{\log_2(n)}{2s+d_Y+d_Z} \rfloor$.*

(i) Let $\widehat{P}_n \in \mathcal{P}([0,1]^{d_Y+d_Z})$ be the distribution associated with the wavelet density estimator based on $\mathcal{S}_0$ (see Definition 2.4), and $\widehat{Q}_n$ is defined similarly based on $\mathcal{S}_1$.

(ii) Further, let $\widehat{R}_n = (\widehat{P}_{n,Z} + \widehat{Q}_{n,Z})/2 \in \mathcal{P}([0,1]^{d_Z})$.

(iii) Let $\widehat{P}_n^\dagger = \widehat{R}_n \otimes \widehat{P}^Z_{n,Y}$, $\widehat{Q}_n^\dagger = \widehat{R}_n \otimes \widehat{Q}^Z_{n,Y}$. The wavelet-based COT estimator is defined by

$$\boxed{\widehat{V}_{\mathrm{c},n} := \mathsf{CW}_h(\widehat{P}_n^\dagger, \widehat{Q}_n^\dagger).}$$

Since conditional optimal transport requires alignment of the Z-marginal, we construct a shared Z-marginal using the pooled Z-samples from both data sources. The paired (Y, Z) observations from each source are then used to estimate the corresponding conditional distributions.

Remark 3.2 (Computation) *Note that $\widehat{P}_n^\dagger$ and $\widehat{Q}_n^\dagger$ are continuous distributions. Consequently, $\widehat{V}_{\mathrm{c},n}$ is approximated via a resampling procedure, which is standard in the literature (see, e.g., [Weed and Berthet, 2019, Deb et al., 2021]). The detailed algorithm is deferred to Section E.1. We discuss the computational complexity in Section E.3 and show that it is typically no larger than that of the estimator with oracle samplers of P and Q.*

Remark 3.3 (Adaptiveness to smoothness) *In practice, if the smoothness order s of the underlying density is unknown and cannot be estimated, the truncation level J_n cannot be chosen directly from our theoretical prescription. This can be addressed using a standard adaptive truncation procedure, such as the one described in Section 4.2 of Niles-Weed and Berthet [2022], at the cost of an additional $\log n$ factor in the convergence rate. Nevertheless, even when the smoothness is known, the theoretical choice of J_n still involves unspecified constants that must be selected in practice. For this reason, our empirical implementation chooses J_n directly using heuristic tuning rules; see Appendix F.2.3.*

3.2 Fast Convergence Rate with General Objective

To bound the statistical convergence rate of the wavelet-based COT estimator, we first present the following stability result of COT under Lipschitz kernels and objectives.

Proposition 3.4 (Stability) *Suppose that $\mathsf{W}_1(P_Y^z, P_Y^{z'}) \le L_p\|z - z'\|_2, \mathsf{W}_1(Q_Y^z, Q_Y^{z'}) \le L_p\|z - z'\|_2$ for any z, z', and h is L_h-Lipschitz, then*

$$|\mathsf{CW}_h(\widehat{P}, \widehat{Q}) - \mathsf{CW}_h(P, Q)| \le 2L_hL_p\mathsf{W}_1(P_Z, \widehat{P}_Z) + L_h \int \mathsf{W}_1(P_Y^z, \widehat{P}_Y^z) + \mathsf{W}_1(Q_Y^z, \widehat{Q}_Y^z)\,\mathrm{d}\widehat{P}_Z(z).$$

The first term on the right-hand side above can be bounded using standard results from wavelet density estimation [Niles-Weed and Berthet, 2022]. In contrast, we derive the convergence rate of the second term for the density estimator introduced in Section 3.1.

Proposition 3.5 (Key estimation bound) *Under Assumption 2.5-2.6,*

$$\mathbb{E}\left[\int \mathsf{W}_2^2(\widehat{Q}_Y^z, Q_Y^z)\,\widehat{R}(\mathrm{d}z)\right] \le Cn^{-\frac{2s}{2s+d_Y+d_Z}},$$

where C is a constant that depends on $P, Q, \gamma, s, d_Y, d_Z$.

We are now ready to state our main result.

Theorem 3.6 (Fast convergence of COT under smooth densities) *Under Assumption 2.5 (γ), 2.6 (s) and 2.7 (L_h), we have*

$$\mathbb{E}[|V_{\mathrm{c}} - \widehat{V}_{\mathrm{c},n}|] \le CL_h n^{-\frac{s}{2s+d_Y+d_Z}},$$

where C is a constant that depends on $P, Q, \gamma, s, d_Y, d_Z$.

Further, for inferential purpose, we have

Theorem 3.7 (High probability bound under smooth densities) *Under Assumption 2.5 (γ), 2.6 (s) and 2.7 (L_h), for every fixed $\delta \in (0,1)$ and all sufficiently large n (depending on δ), we have*

$$\mathbb{P}\left(|V_{\mathrm{c}} - \widehat{V}_{\mathrm{c},n}| \le CL_h n^{-\frac{s}{2s+d_Y+d_Z}} \sqrt{\log(4/\delta)}\right) \ge 1-\delta,$$

where C depends only on $P, Q, \gamma, s, d_Y, d_Z$.

When s is sufficiently large, the convergence rate of the COT estimator is close to the parametric rate of order $n^{-1/2}$, relieving the dependence of dimension $d_Y + d_Z$. We compare the rates and assumptions of our method with those of existing approaches having finite-sample convergence guarantees; see Table 1. On the other hand, we conjecture that the rate can be further improved to be $n^{-(s+1)/(2s+d_Y+d_Z)}$ as in the unconditional setting. We are not able to achieve this due to our proving technique that uses L_2 norm to upper bound the $\mathcal{B}_{2,1}^{-1}([0,1]^d)$ norm for densities (Lemma D.2). We leave the study of the minimax optimal rate to future work.

3.3 Statistical Inference with Quadratic Objective

In this section, we focus on the quadratic objective $h(y_0, y_1) = \|y_0 - y_1\|_2^2$, which allows a stronger stability bound and thus an analysis of the asymptotic distribution of the wavelet-based COT estimator, enabling standard statistical inference.

In our conditional setting, we let $\varphi(y, z)$ be the Brenier potential [Knott and Smith, 1984, Brenier, 1991] between P_Y^z and Q_Y^z for each $z \in [0,1]^{d_Z}$. To obtain a sharper convergence-rate analysis, we will impose stronger smoothness assumptions on the densities p, q and the functions φ. In particular, under a quadratic objective h, the smoothness of φ can often be deduced from that of p and q. Roughly speaking, for each fixed z, if the conditional densities $p_Y^z(y)$ and $q_Y^z(y) \in \mathcal{C}^s([0,1]^{d_Y})$, then typically the corresponding potentials $\varphi(\cdot, z) \in \mathcal{C}^{s+2}([0,1]^{d_Y})$. See, e.g., Villani [2008, Theorem 12.50].

Proposition 3.8 (Stability with quadratic objective) *Under Assumption 2.5 (γ), and further assume that: (i) $\varphi(\cdot, z) \in \mathcal{C}^2([0,1]^{d_Y})\,\forall z \in [0,1]^{d_Z}$; (ii) $C_2 := \sup_{z\in[0,1]^{d_Z}} \|\varphi(\cdot, z)\|_{\mathcal{C}^2([0,1]^{d_Y})} < \infty$. Then, there exists a constant $\lambda > 0$ that depends on γ, C_2 such that*

$$\begin{aligned} 0 \le& \mathsf{CW}_2(\widehat{P}_n^\dagger, \widehat{Q}_n^\dagger)^2 - \mathsf{CW}_2(P,Q)^2 - \int \phi \mathrm{d}(\widehat{P}_n^\dagger - P) - \int \psi \mathrm{d}(\widehat{Q}_n^\dagger - Q) \\ \le& 2\lambda \int \mathsf{W}_2^2(\widehat{P}_Y^z, P_Y^z) + \mathsf{W}_2^2(\widehat{Q}_Y^z, Q_Y^z)\widehat{R}_n(\mathrm{d}z), \end{aligned}$$

where $\phi(y,z) = \|y\|_2^2 - 2\varphi(y,z)$ and $\psi(y,z) = \|y\|_2^2 - 2\varphi^*(y,z)$, with $\varphi^*(y,z) := \sup_{y'\in[0,1]^{d_Y}} \{y^\top y' - \varphi(y',z)\}$.

When $2s > d_Y + d_Z$, Proposition 3.5 implies that the right-hand side of the above inequality is $o_p(n^{-1/2})$. Thus, the centered estimator and the linear term $\int \phi \mathrm{d}(\widehat{P}_n^\dagger - P) + \int \psi \mathrm{d}(\widehat{Q}_n^\dagger - Q)$ have the same distributional limit after multiplication by $\sqrt{n}$. As a result, we have the following result.

Theorem 3.9 (Central limit theorem of COT) *Under Assumption 2.5 (γ) - 2.6 (s), and that p, q, $\varphi \in \mathcal{C}^{s+1}([0,1]^{d_Y+d_Z})$, $2s > d_Y + d_Z$,*

$$\sqrt{n}\left(\widehat{V}_{\mathrm{c},n} - V_{\mathrm{c}}\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 := Var_P(\eta(Y,Z)) + Var_Q(\kappa(Y,Z))$ with

$$\eta(y,z) := \phi(y,z) - \frac{1}{2}\left(\int\left(\phi(y',z)\frac{p(y',z)}{p_Z(z)} - \psi(y',z)\frac{q(y',z)}{q_Z(z)}\right)\mathrm{d}y'\right),$$
$$\kappa(y,z) := \psi(y,z) - \frac{1}{2}\left(\int\left(\psi(y',z)\frac{q(y',z)}{q_Z(z)} - \phi(y',z)\frac{p(y',z)}{p_Z(z)}\right)\mathrm{d}y'\right).$$

3.4 Covariate-Dependent Treatment

We now consider treatment assignment mechanisms that may depend on covariates, as often arises in observational studies. Let $e(z) := \mathbb{P}(W = 1 \mid Z = z)$ be the propensity score, which captures heterogeneity in treatment assignment across covariate profiles. Because the observed sample may no longer have the same covariate distribution as the target population, we modify our COT estimator using doubly robust estimation. We then introduce the assumptions needed for this extension.

Definition 3.10 (Wavelet-based COT estimator under covariate shift) *Given an $i.i.d.$ sample (Y_i, Z_i, W_i), let $\hat{e}$ be an estimator of e and denote $\hat{\pi}_a(z) := a\hat{e}(z) + (1-a)(1-\hat{e}(z))$ for $a = 0, 1$. Let $\hat{m}_{a,\xi}(z)$ denote an estimator for $m_{a,\xi}(z) := \mathbb{E}[\xi(Y(a), Z) \mid Z = z]$ for $a = 0, 1$ and any wavelet basis ξ. For the wavelet density estimator, define the adjusted wavelet coefficient to be*

$$\widehat{\theta}^{\mathrm{dr}}_{a,\xi} := \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{m}_{a,\xi}(Z_i) + \frac{\mathbf{1}\{W_i = a\}}{\widehat{\pi}_a(Z_i)}\{\xi(Y_i, Z_i) - \widehat{m}_{a,\xi}(Z_i)\}\right].$$

Let $\widehat{P}_n$ (resp. $\widehat{Q}_n$) be the distribution associated with the wavelet density estimator using $(\widehat{\theta}^{\mathrm{dr}}_{0,\xi})_\xi$ (resp. $(\widehat{\theta}^{\mathrm{dr}}_{1,\xi})_\xi$). The remaining steps follow as in Definition 3.1.

Assumption 3.11 (Unconfoundedness and overlap) *Assume that $(Y(0), Y(1)) \perp W \mid Z$. Moreover, there exists $\eta \in (0, 1/2)$ such that $\eta \le e(z) \le 1 - \eta$ $\forall z \in [0,1]^{d_Z}$.*

Assumption 3.12 (Doubly robust error) *Assume that for any basis ξ, $\widehat{e}, \widehat{m}_{a,\xi} \perp (Y_i, Z_i, W_i)_i$, and $\eta/2 \le \widehat{e}(z) \le 1 - \eta/2$ $\forall z$, $\|\hat{m}_{a,\xi}\|_{L^2(P_Z)} \le M$ $\forall \xi$ and $\|\hat{m}_{a,\xi}\|_\infty \le b 2^{j(d_Y+d_Z)}$ $\forall \xi$ at level j a.s. Further, $\sup_{\xi\text{-basis}} \mathbb{E}[\|\widehat{e} - e\|^2_{L^2(P_Z)}\|\widehat{m}_{a,\xi} - m_{a,\xi}\|^2_{L^2(P_Z)}] = o(n^{-1})$.*

The independence assumption holds with independent training and evaluation samples; for cross-fitting, it must be formulated fold by fold. We also introduce the following technical assumptions on $\hat{e}$ and $\hat{m}$. It is possible to remove this assumption by redefining $\widehat{f}_{J_n}$ in (1) through an $\|\|_{L^2}$-projection onto the space of probability densities.

Assumption 3.13 (High-probability error) *There is a constant $\bar{C}$ such that, for any ξ and $u > 0$, with probability at least $1 - \exp(-\bar{C}u)$, we have $\|\widehat{e} - e\|_{L^2(P_Z)}\|\widehat{m}_{a,\xi} - m_{a,\xi}\|_{L^2(P_Z)} \le u$.*

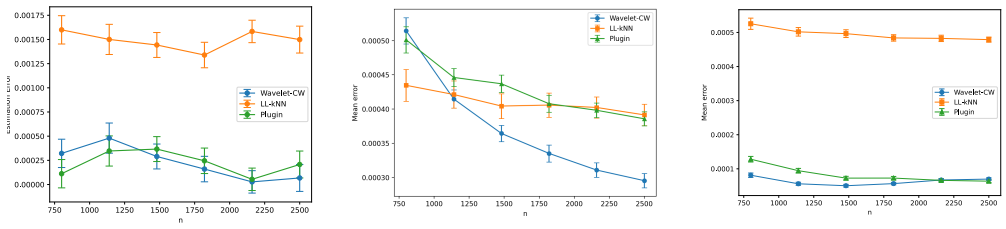
Author check: The proof of Proposition D.22, and hence the stated proof of Theorem 3.14, uses the stronger first-stage tail bound $\exp(-n\bar{C}u)$ rather than $\exp(-\bar{C}u)$. Please verify that stronger assumption for the nuisance estimators, or revise the argument. The stronger assumption has not been imposed in this redline.

We now state the result of fast convergence under smooth densities.

Theorem 3.14 (Fast convergence of COT under covariate shift) *Under Assumption 2.5 (γ), 2.6 (s), 2.7 (L_h), 3.11 (η), 3.12 (M, b) and 3.13 ($\bar{C}$), we have*

$$\mathbb{E}[|V_{\mathrm{c}} - \widehat{V}_{\mathrm{c},n}|] \le C L_h n^{-\frac{s}{2s+d_Y+d_Z}},$$

where C is a constant that depends on $P, Q, \gamma, s, \eta, M, b, \bar{C}, d_Y, d_Z$.

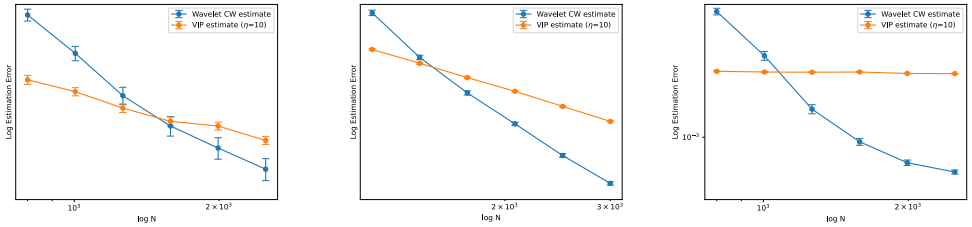


(a) Location model (b) Quadratic model (c) Scale model

Figure 1: Plots of estimation error comparison of our method (*Wavelet CW estimate*) with [Ji et al., 2023] (*LL-kNN*, for which we use "knn"-based nuisance estimator) and [Lin et al., 2026] (*Plugin*) with $d_Y = 1$, $d_Z = 1$. The estimation error is defined by $\mathbb{E}[|V_c - \widehat{V}_c|]$. The mean error curve and the corresponding standard-error confidence bands are computed by aggregating results over 100 Monte Carlo repetitions. The plots show that the convergence rate of our method is at least comparable with the existing approaches in low dimension.

4 Numerical Experiment

4.1 Estimation



(a) Location model (b) Quadratic model (c) Scale model

Figure 2: Log-log plots of estimation error comparison of our method (*Wavelet CW estimate*) and [Lin et al., 2025] (*VIP estimate* with $\eta = 10$) with $d_Y = 2$, $d_Z = 2$. The estimation error is defined by $\mathbb{E}[|V_c - \widehat{V}_c|]$. The mean error curve and the corresponding standard-error confidence bands are computed by aggregating over 300 Monte Carlo repetitions. The plots show that the convergence rate of our method is significantly faster than the rate of VIP estimate, which suffers from the curse of dimensionality based on its theory.

We empirically compare the convergence rate of our estimator (Section 3.1) with existing approaches including [Ji et al., 2023, Lin et al., 2025, 2026].

Data generation mechanism. We outline a default data-generating mechanism. The coordinates of the covariate $Z \in [0,1]^{dz}$ are drawn i.i.d. from $\mathrm{Uniform}(0,1)$. The potential outcomes Y follow

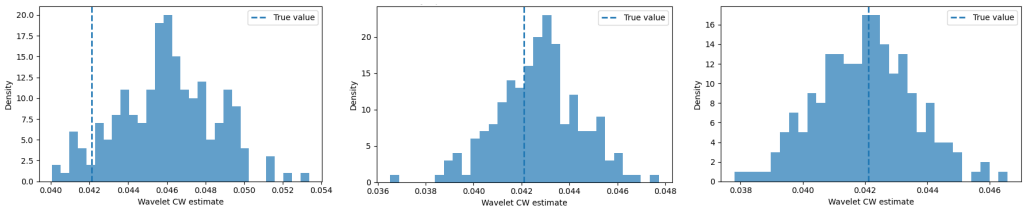
$$Y(w) \mid Z = z \sim \mathcal{N}\big(\mu_w(z), \Sigma_w(z)\big), \quad w \in \{0,1\}.$$

The objective function is quadratic, $h = \|y_0 - y_1\|_2^2$. This underlying distribution and objective enable a closed form evaluation of the COT value thanks to the Gelbrich formula (Proposition B.2). Note that we shall choose the parameter $\mu_w(z), \Sigma_w(z)$ appropriately such that most of its measure is supported on $[0,1]^{d_Y}$. Specifically, we consider three data models.

1. (Location) $\mu_w(z) = \alpha_w z + \beta_w$, $\Sigma_w(z) \equiv \Sigma_w$.
2. (Quadratic) $\mu_w(z) = A_w(z - \alpha_w)^2 + \beta_w, \Sigma_w(z) \equiv \Sigma_w$, $(z - \alpha_w)^2$ is an entry-wise square.
3. (Scale) $\mu_w(z) \equiv \alpha_w, \Sigma_w(z) = (\zeta_w^\top z) \times \Sigma_w$, where $(\zeta_w^\top z)$ is a scalar.

Implementation. Code is publicly available.[2] We follow Algorithms 1 and 2, with wavelet parameters specified in Section F. Since smoothness is unknown, the appendix describes tuning the wavelet

[2] `https://github.com/siruilin1998/causalOT_inference/releases/tag/neurips2026-camera-ready`



(a) $n = 1000$ (b) $n = 2000$ (c) $n = 3000$

Figure 3: Histogram of the proposed estimator with the true value marked by a vertical dashed line. Here the default setting $d_Y = 2$, $d_Z = 2$ is adopted. The results are aggregated over 200 repeats.

Table 2: Coverage of the proposed 95% confidence intervals across different settings. We consider three choices of (d_Y, d_Z), each evaluated at three sample sizes. Confidence intervals are constructed using the bootstrap with 100 resamples. Reported coverage rates are aggregated over 200 repeats.

Dimension	$d_Y = 2, d_Z = 1$			$d_Y = 2, d_Z = 2$			$d_Y = 3, d_Z = 2$		
Sample size (n)	1000	2000	3000	1000	2000	3000	4000	5000	6000
Coverage	92.5%	93%	93.5%	86%	98.5%	97.5%	90%	93.5%	96%

resolution by minimizing either the estimator or its bootstrap standard deviation. We implement Ji et al. [2023] using `DualBounds` (`https://dualbounds.readthedocs.io/en/latest/index.html`) and Lin et al. [2025, 2026] using their public code.

Case i. ($d_Y = 1$) We first compare our method with [Ji et al., 2023] and [Lin et al., 2026] in the setting of $d_Y = d_Z = 1$ across the three models. Specifically, for [Ji et al., 2023], we use the knn-based (nonparametric) nuisance estimator to avoid model misspecification. The plots of estimation error comparison are shown in Figure 1. The plots imply that in low-dimensional cases, our method performs comparably to existing approaches, consistent with the fact that smoothness assumptions do not improve convergence rates beyond the parametric regime.
Case ii. ($d_Y > 1$) Since neither the approach of [Ji et al., 2023] nor [Lin et al., 2026] is implemented for $d_Y > 1$, we focus on comparing with the VIP estimate of [Lin et al., 2025], for which we take the fixed penalty $\eta = 10$. A fixed penalty can retain relaxation bias relative to the COT value. The log-log plots of the estimation error comparison of these two methods are shown in Figure 2, which demonstrate a significant improvement of convergence rate of our method comparing to the VIP estimate across multiple scenarios.

4.2 Inference

We empirically validate the performance of the inference procedure based on our proposed estimator on the case $d_Y > 1$. A real-data analysis is deferred to Section A.

Data generation mechanism. We use the location model to evaluate the inference performance. Specifically, we consider a total of three scenarios with different dimensions: $(d_Y = 2, d_Z = 1)$, $(d_Y = 2, d_Z = 2)$ as the default setting, and $(d_Y = 3, d_Z = 2)$. The details of $\mu_w(z)$, $\Sigma_w(z)$ are in Section F.2. In each scenario, we use three different sample sizes n (specific values listed in Table 2).
Bootstrap. We use the bootstrap with 100 resamples to estimate the variance σ^2 in Theorem 3.9 and plug it in to construct the confidence interval. The alternative approach for $d_Y > 1$ discussed in Section 4.1 either has non-negligible bias or lacks an established asymptotic distribution, making it unsuitable for statistical inference. We therefore focus on our method.
Coverage rate. We report the coverage results for 95% confidence intervals in Table 2. Overall, the coverage rates are acceptable when the sample size n is moderately large. As expected, for each configuration of (d_Y, d_Z), coverage improves as the sample size increases; as the dimensions d_Y and d_Z grow, larger samples are required to achieve proper coverage.
Asymptotic distribution. In Figure 3, we display the distribution of the proposed estimator $\mathsf{CW}_2^2(\widehat{P}_n^\dagger, \widehat{Q}_n^\dagger)$, with the true value $\mathsf{CW}_2^2(P, Q)$ marked by a vertical dashed line. We focus on the default setting $d_Y = 2$, $d_Z = 2$; histograms for the other two settings are provided in Section F. The distribution of the estimator $\mathsf{CW}_2^2(\widehat{P}_n^\dagger, \widehat{Q}_n^\dagger)$ becomes increasingly normal as n grows, consistent with the CLT in Theorem 3.9. The bias also decreases with larger sample sizes: for $n = 1000$, the estimator is upward biased, while by $n = 3000$, the bias becomes negligible.

5 Conclusion

We propose a primal COT solver that leverages the smoothness of the marginal density to enable accurate estimation and statistical inference for causal partial identification sets. Our method accommodates multivariate outcomes, which are not supported by existing quantile-based approaches, and relieves the curse of dimensionality faced by alternative methods.

Acknowledgments and Disclosure of Funding

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.

The authors declare no competing interests.

References

Mohammad Al-Jarrah, Bamdad Hosseini, Niyizhen Jin, Michele Martino, and Amirhossein Taghvaei. Error analysis of triangular optimal transport maps for filtering. *arXiv preprint arXiv:2510.19283*, 2025.

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

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

Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Juraj Bodik, Yaxuan Huang, and Bin Yu. Cross-world assumption and refining prediction intervals for individual treatment effects. *arXiv preprint arXiv:2507.12581*, 2025.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021.

Stephan Eckstein and Gudmund Pammer. Computational methods for adapted optimal transport. *Annals of Applied Probability*, 34(1A):675–713, 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.

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 and Qingyuan Zhao. Causal ANOVA: Estimation, inference, applications. In *Conference on Causal Learning and Reasoning (CLeaR)*, May 2025.

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.

Matthias Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Nicola Gigli. On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proceedings of the Edinburgh Mathematical Society*, 54(2):401–409, 2011.

James J Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, 73(3):669–738, 2005.

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.

Shayan Hundrieser, Marcel Klatt, Axel Munk, and Thomas Staudt. A unifying approach to distributional limits for empirical optimal transport. *Bernoulli*, 30(4):2846–2877, 2024.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *Annals of Statistics*, 49(2):1166–1194, 2021. doi: 10.1214/20-AOS1997.

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.

Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49, 1984.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

Sirui Lin, Zijun Gao, Jose Blanchet, and Peter Glynn. Tightening causal bounds via covariate-aware optimal transport. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Sirui Lin, Zijun Gao, Jose Blanchet, and Peter Glynn. Causal partial identification via conditional optimal transport. In *Proceedings of the Twenty-Ninth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

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. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

Charles F. Manski. Monotone treatment response. *Econometrica*, 65(6):1311–1334, 1997.

Charles F. Manski and John V. Pepper. Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, 68(4):997–1012, 2000.

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.

Yves Meyer. *Wavelets and operators*. Number 37. Cambridge university press, 1992.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Viet Anh Nguyen, Fan Zhang, Wang Shanshan, Jose Blanchet, Erick Delage, and Yinyu Ye. Robustifying conditional portfolio decisions via optimal transport. *Operations Research*, 2024.

Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in wasserstein distance. *Annals of Statistics*, 50(3):1519–1540, 2022.

Victor M Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application*, 6:405–431, 2019.

Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

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

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195, 2010.

Jiyuan Tan, Jose Blanchet, and Vasilis Syrgkanis. Partial identification of policy-relevant treatment effects with instrumental variables via optimal transport. *arXiv preprint arXiv:2604.12263*, 2026.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2008.

Jonathan Weed and Quentin Berthet. Estimation of smooth densities in wasserstein distance. In *conference on Learning Theory*, pages 3118–3119. PMLR, 2019.

Tianlin Xu, Li Kevin Wenliang, Michael Munn, and Beatrice Acciaio. COT-GAN: Generating sequential data via causal optimal transport. *Advances in Neural Information Processing Systems*, 33:8798–8809, 2020.

Appendix for "Statistical Inference in Causal Partial Identification under Smooth Densities"

In Section A, we report a real-data analysis. In Section B, we review the basics of optimal transport. In Section C, we introduce the wavelet-based system. In Section D, we present additional technical details and results. In Section E, we provide the detailed algorithm for our proposal. In Section F, we include additional numerical results.

Notation. 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 $\mathcal{P}_{\mathrm{ac}}(\Omega)$, for $\Omega \subseteq \mathbb{R}^d$, to denote the set of probability measures that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. 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, $\int g \mathrm{d}\mu = \int_{\mathcal{Z}} \int_{\mathcal{Y}} g(z, y) \mathrm{d}\mu_Y^z(y) \mathrm{d}\mu_Z(z)$. Given a density function $f(y, z)$, we denote its marginal density of z by $f_Z(z)$, and the conditional density of y given z by $f(y \mid z) = f(y, z)/f_Z(z)$.

For joint distributions P and Q on $\mathcal{Y} \times \mathcal{Z}$, we write ϕ_Z and ψ_Z for the Kantorovich potentials associated with the conditional laws P_Y^z and Q_Y^z. We use P_Z to denote the marginal distribution of Z induced from P, and similarly for Q_Z. We use $\mathsf{W}_p(P, Q)$ for the Wasserstein-p distance between P and Q (Section B), and $\mathsf{CW}_p(P, Q)$ for the conditional Wasserstein-p distance defined as $(\mathbb{E}[\mathsf{W}_p(P(Y \mid Z), Q(Y \mid Z))^p])^{1/p}$.

A Real Data Analysis

We investigate the randomized experiment from the Student Achievement and Retention Project (STAR) [Angrist et al., 2009], which evaluated the effects of academic support services and financial incentives on college student performance. We use gender and baseline GPA (supported on $[0, 4]$) as covariates, and first-year and second-year GPA (supported on $[0, 4]$) as outcomes. Because the true treatment effect is not directly observable, we construct hypothetical potential outcomes under control and treatment based on the observed outcomes. Specifically, we treat the original outcomes as control potential outcomes and generate gender-dependent treatment effects, motivated by the finding in [Angrist et al., 2009] that female students benefit more from the intervention. We add these treatment effects to the control outcomes to obtain treatment potential outcomes. Based on the constructed hypothetical dataset, in each simulation replicate, we randomly sample 1300 control and 1300 treatment observations and use them as the observed dataset. We generate 200 datasets and aggregate the results.

We apply our method to each generated dataset as described in Section 4.2, with outcome dimension $d_Y = 2$ and covariate dimension $d_Z = 1$ or 2. Particularly, we consider three sets of covariates for conditioning:

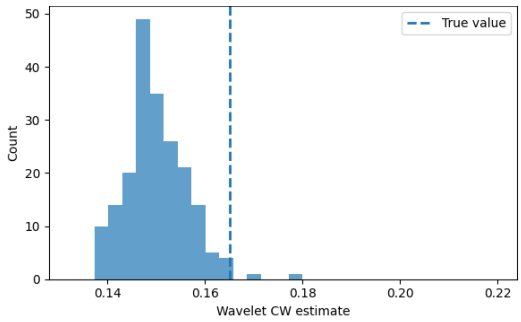
(a) Conditioning on baseline GPA only.

(b) Conditioning on gender only.

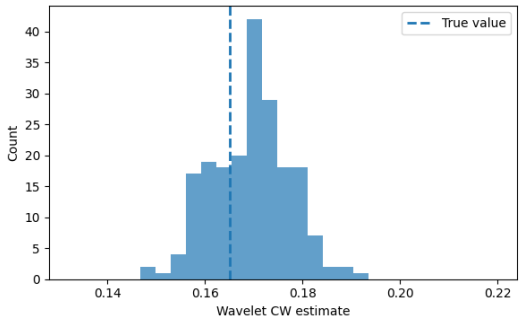
(c) Conditioning on baseline GPA and gender.

By the construction of our hypothetical treatment effects, the conditional distribution under treatment differs from that under control by a location shift depending on the gender. Therefore, conditioning on gender, either alone or together with the baseline GPA, the optimal values of the population-wise COT recover the sharp lower bound given by the Cauchy-Schwarz inequality,

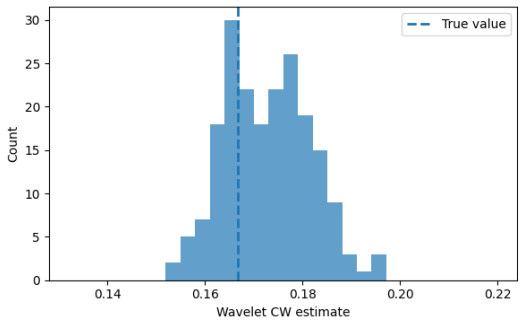
$$\mathbb{E}\left[\|\mathbb{E}[Y(1) \mid \text{Gender}] - \mathbb{E}[Y(0) \mid \text{Gender}]\|_2^2\right].$$



(a) Conditioning on baseline GPA only



(b) Conditioning on gender only



(c) Conditioning on baseline GPA and gender

Figure 4: Histogram of the proposed estimator applied to hypothetical STAR datasets under different conditioning sets of covariates. Results are aggregated over 200 simulated hypothetical datasets.

In contrast, when conditioning only on baseline GPA, the conditioning set does not fully capture the treatment effect heterogeneity, and therefore the optimal value of the COT is strictly smaller, leading to a looser PI set.

As shown in Figure 4, when conditioning on gender (panels (b) and (c)), the estimators' distributions are centered at similar values, both substantially larger than when conditioning only on baseline GPA (panel (a)). This is consistent with the theoretical discussion above and highlights the importance of conditioning on informative covariates to tighten the partial identification sets. When conditioning on both gender and baseline GPA (panel (c), $d_Z = 2$), the estimator exhibits slightly slower convergence, resulting in mildly larger upward bias and a less normal distribution compared to conditioning on gender alone (panel (b), $d_Z = 1$). One possible future direction is to combine the proposed method with automatic selection of influential covariates.

B Optimal Transport

Definition B.1 (Optimal transport) *Given a measurable objective function h and probability distributions $P, Q \in \mathcal{P}(\mathbb{R}^{d_X})$, the OT distance is defined as*

$$\mathsf{W}_h(P,Q) := \min_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi[h(X,X')],$$

where

$$\Pi(P,Q) := \left\{\pi \in \mathcal{P}(\mathcal{X}^2) : \pi_X = P,\ \pi_{X'} = Q\right\}. \tag{2}$$

When $h(x,x') = \|x - x'\|_2^p$, we denote $\mathsf{W}_p(P,Q) = \mathsf{W}_h(P,Q)^{\frac{1}{p}}$, which is the so-called Wasserstein p-distance.

The 2-Wasserstein distance between Gaussian measures admits a closed-form expression.

Proposition B.2 (OT between Gaussian distributions[Gelbrich, 1990])

$$\mathsf{W}_2^2(\mathcal{N}(m_0,\Sigma_0), \mathcal{N}(m_1,\Sigma_1)) = \|m_0 - m_1\|_2^2 + \mathrm{tr}\Big(\Sigma_0 + \Sigma_1 - 2(\Sigma_1^{1/2}\Sigma_0\Sigma_1^{1/2})^{1/2}\Big).$$

C Boundary-corrected Wavelet System

Next, we describe the specific wavelet basis used in our proposal.

Definition C.1 (Boundary-corrected wavelet basis) *Let $j_0 \geq 0$ be the minimal resolution level at which boundary correction is applied. The boundary-corrected wavelet basis on $L^2([0,1])$ is defined by*

$$\Phi_0^{\mathrm{bc}} = \{\phi_{j_0 k}^{\mathrm{bc}}, k \in \mathcal{K}_0(j_0)\},$$
$$\Psi_0^{\mathrm{bc}} = \{\psi_{jk}^{\mathrm{bc}}, k \in \mathcal{K}_0(j), j \geq j_0\},$$

where $\mathcal{K}_0(j) := \{0, 1, \ldots, 2^j - 1\}$. Specifically, the wavelet basis is compactly supported orthonormal on $[0,1]$. The concrete definition of $\phi^{\mathrm{bc}}, \psi^{\mathrm{bc}}$ can be found in Manole et al. [2024, Section A.2.1].

For outcomes $y \in [0,1]^{d_Y}$ and covariates $z \in [0,1]^{d_Z}$, define

$$x = ((y_s)_{1 \leq s \leq d_Y},\ (z_{s'})_{1 \leq s' \leq d_Z}) \in [0,1]^{d_Y + d_Z}.$$

The tensor-product boundary-corrected wavelet functions on $[0,1]^{d_Y+d_Z}$ are defined by

$$\begin{cases} \Phi_{j_0,\mathbf{k}}^{\mathrm{bc}}(x) = \displaystyle\prod_{r=1}^{d_Y+d_Z} \phi_{j_0 k_r}^{\mathrm{bc}}(x_r),\ \mathbf{k} \in \mathcal{K}_0(j_0)^{d_Y+d_Z} \\ \Psi_{j,\mathbf{k},\mathbf{l}}^{\mathrm{bc}}(x) = \displaystyle\prod_{r=1}^{d_Y+d_Z} \begin{cases} \phi_{j,k_r}^{\mathrm{bc}}(x_r),\ l_r = 0, \\ \psi_{j,k_r}^{\mathrm{bc}}(x_r),\ l_r = 1, \end{cases} \\ \mathbf{l} \in \{0,1\}^{d_Y+d_Z} \backslash \{(0,\ldots,0)\}, \quad \mathbf{k} \in \mathcal{K}_0(j)^{d_Y+d_Z}, \quad j \geq j_0 \end{cases}$$

where $\mathbf{k} = (k_1, \ldots, k_{d_Y+d_Z})$ *and* $\mathbf{l} = (l_1, \ldots, l_{d_Y+d_Z})$.

In the following, we denote

$$\Phi := \{\Phi^{\mathrm{bc}}_{j_0,\mathbf{k}}\}_{\mathbf{k}}, \qquad \Psi_j := \{\Psi^{\mathrm{bc}}_{j,\mathbf{k},\mathbf{l}}\}_{\mathbf{k},\mathbf{l}}\ \forall j \geq j_0.$$

Additional to a wavelet expansion, we also focus on density with uniform smoothness described by the Hölder norm.

Definition C.2 (Hölder space) *Let* $s > 0$*, the Hölder space* $C^s([0,1]^d)$ *consists of functions* f *whose derivatives up to order* $\lfloor s \rfloor$ *extend continuously to* $[0,1]^d$ *and satisfy*

$$\|f\|_{C^s([0,1]^d)} := \sum_{|\alpha| \leq \lfloor s \rfloor} \|D^\alpha f\|_\infty + \sum_{|\alpha| = \lfloor s \rfloor} \sup_{\substack{x,y \in [0,1]^d \\ x \neq y}} \frac{|D^\alpha f(x) - D^\alpha f(y)|}{\|x-y\|^{s-\lfloor s \rfloor}} < \infty,$$

where $D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$, $|\alpha| = \sum_{i=1}^d \alpha_i$.

To conclude this section, we present the following result (c.f. Manole et al. [2024, Lemma 25]), showing that any smooth density function embeds into the Besov space, admitting a wavelet expansion.

Proposition C.3 (Besov embedding) $\mathcal{C}^s([0,1]^d) \subseteq \mathcal{B}^s_{\infty,\infty}([0,1]^d)$, $\mathcal{C}^s \subseteq \mathcal{B}^{s-\epsilon}_{2,2}$ *for any* $\epsilon > 0$.

D Additional Theoretical Results and Proofs

Notation. In the following, for simplicity, we will write $\widehat{P}_n$ (resp. $\widehat{Q}_n$, $\widehat{R}_n$) as $\widehat{P}$ (resp. $\widehat{Q}$, $\widehat{R}$). $\lesssim$ means $\leq$ with suppressing a constant independent of n.

D.1 Proof of Proposition 3.5

In this section, we present the technical proof of the core result (Proposition D.4). As a preparation, we introduce the following lemmas.

Lemma D.1 (Niles-Weed and Berthet [2022, Theorem 4]) *Suppose that* $\widehat{\mu}, \mu \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$*, and let* $\hat{f}, f \in L^2([0,1]^d)$ *be the density function of* $\widehat{\mu}, \mu$*, respectively. If there exists a constant* $\gamma > 0$ *such that* $\hat{f}(x) \vee f(x) \geq \gamma^{-1}\ \forall x \in [0,1]^d$*, then*

$$\mathsf{W}_2(\widehat{\mu}, \mu) \leq C_{\mathrm{wb}} \gamma^{\frac{1}{2}} \|\hat{f} - f\|_{\mathcal{B}^{-1}_{2,1}([0,1]^d)},$$

where $C_{\mathrm{wb}} > 0$ *is a universal constant.*

Lemma D.2 (Meyer [1992, Section 6.10, Proposition 7]) *There is a universal constant* $C_{\mathrm{emb}} > 0$*, such that for any* $f \in L^2([0,1]^d)$*,*

$$\|f\|_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} \leq C_{\mathrm{emb}} \|f\|_{L_2([0,1]^d)}.$$

Lemma D.3 (Manole et al. [2024, Lemma 27 and proof of Lemma 29]) *Assume there exist* $\gamma, s > 0$ *such that a density function* $q(x) \geq \gamma^{-1}$*,* $\forall x \in [0,1]^d$*, and such that* $q \in \mathcal{B}^s_{\infty,\infty}([0,1]^d)$*. Let* $\tilde{q}_n$ *be the wavelet projection based on an i.i.d. sample of size* n *drawn from the distribution with density* q*. Then,*

(i) $\int \tilde{q}_n(x)\mathrm{d}x = 1$ *with probability* 1.

(ii) There exists $c_{\mathrm{b}} > 0$*, depending on* $[0,1]^d$ *and* $\|q\|_{\mathcal{B}^s_{\infty,\infty}([0,1]^d)}$*, such that with probability at least* $1 - c_{\mathrm{b}}/n^2$*,*

$$\sup_{x \in [0,1]^d} |q(x) - \tilde{q}_n(x)| \leq \gamma^{-1}/2.$$

Proposition D.4 (Key estimation bound, Proposition 3.5) *Under Assumption 2.5-2.6,*

$$\mathbb{E}\left[\int \mathsf{W}_2^2(\widehat{Q}^z_Y, Q^z_Y)\, \widehat{R}(\mathrm{d}z)\right] \leq C n^{-\frac{2s}{2s+d_Y+d_Z}},$$

where $C = C^2_{\mathrm{wb}} \gamma^2 (16\gamma^5 (C_{\mathrm{emb}}\gamma)^2 (C_{\mathrm{emb}}+1) C^\dagger + c_{\mathrm{b}})$*, and* $C^\dagger, c_{\mathrm{b}}$ *depend on* d_Y, d_Z, q*.*

Proof D.5 (Proof of Proposition D.4) *Denote the (wavelet) density of Q (resp. P) by q (resp. p), the (wavelet) density of $\widehat{Q}$ (resp. $\widehat{P}$) by $\hat{q}$ (resp. $\hat{p}$), and the (wavelet) density of $\widehat{R}$ by $\hat{r}$. As for the wavelet projection, we replace $\hat{p}$ with $\tilde{p}$. Then, by Assumption 2.5,*

$$q(y|z) = \frac{q(y,z)}{q_Z(z)} = \frac{q(y,z)}{\int q(y,z)\mathrm{d}y} \geq \gamma^{-2}.$$

Then, applying Lemma D.1, we get

$$\mathbb{E}\left[\int \mathsf{W}_2^2(\widehat{Q}_Y^z, Q_Y^z)\,\widehat{R}(\mathrm{d}z)\right] \leq C_{\mathrm{wb}}^2\gamma^2\mathbb{E}\left[\int_{\mathcal{Z}} \|\hat{q}_Y^z(y) - q_Y^z(y)\|^2_{\mathcal{B}^{-1}_{2,1}([0,1]^{d_Y})}\mathrm{d}\widehat{R}(z)\right].$$

To further investigate the conditional densities, we apply Lemma D.3 under Assumption 2.6 and get, there exists $c_{\mathrm{b}} > 0$, depending on d_Y, d_Z, q, such that with probability at least $1 - c_{\mathrm{b}}/n^2$,

$$\sup_{(y,z)\in[0,1]^{d_Y+d_Z}} |q(y,z) - \tilde{q}(y,z)| \leq \gamma^{-1}/2,$$
$$\sup_{(y,z)\in[0,1]^{d_Y+d_Z}} |p(y,z) - \tilde{p}(y,z)| \leq \gamma^{-1}/2.$$

As a result,

$$\min_{(y,z)\in[0,1]^{d_Y+d_Z}} \tilde{q}(y,z) \geq \gamma^{-1}/2, \tag{3}$$

$$\tilde{r}(z) = \frac{1}{2}(\tilde{p}_Z(z) + \tilde{q}_Z(z)) \in [\gamma^{-1}/2, 2\gamma] \ \ \forall z \in [0,1]^{d_Z}. \tag{4}$$

Then, on this high probability event, the wavelet projection $\tilde{q}$ is equal to the wavelet density estimator $\hat{q}$ (so does $\hat{r}$), thus Eq. (3)-(4) *also hold for $\hat{q}, \hat{r}$.*

Also, on this high probability event, we apply the following decomposition to bound the right-hand side of the above inequality.

$$\begin{aligned}
&\int_{[0,1]^{d_Z}} \|\hat{q}_Y^z(y) - q_Y^z(y)\|^2_{\mathcal{B}^{-1}_{2,1}([0,1]^{d_Y})}\mathrm{d}\widehat{R}(z)\\
=&\int_{[0,1]^{d_Z}} \left\|\frac{\hat{q}(\cdot,z)}{\hat{q}_Z(z)} - \frac{q(\cdot,z)}{q_Z(z)}\right\|^2_{\mathcal{B}^{-1}_{2,1}([0,1]^{d_Y})}\mathrm{d}\widehat{R}(z) \quad (\textit{by the definition of } \hat{q}_Y^z(y))\\
\leq& 2\int_{[0,1]^{d_Z}} \frac{1}{\hat{q}_Z(z)^2}\,\|\hat{q}(\cdot,z) - q(\cdot,z)\|^2_{\mathcal{B}^{-1}_{2,1}([0,1]^{d_Y})}\,\mathrm{d}\widehat{R}(z)\\
&+ 2\int_{[0,1]^{d_Z}} \|q(\cdot,z)\|^2_{\mathcal{B}^{-1}_{2,1}([0,1]^{d_Y})}\left(\frac{1}{\hat{q}_Z(z)} - \frac{1}{q_Z(z)}\right)^2\mathrm{d}\widehat{R}(z)\\
\leq& 2(2\gamma)^3\underbrace{\int_{[0,1]^{d_Z}} \|\hat{q}(\cdot,z) - q(\cdot,z)\|^2_{\mathcal{B}^{-1}_{2,1}([0,1]^{d_Y})}\,\mathrm{d}z}_{\textit{(Term A)}}\\
&+ 2(2\gamma)^3\gamma^2(C_{\mathrm{emb}}\gamma)^2\underbrace{\int_{[0,1]^{d_Z}} (\hat{q}_Z(z) - q_Z(z))^2\,\mathrm{d}z}_{\textit{(Term B)}},
\end{aligned}$$

where the last inequality is due to Assumption 2.5 and Eq. (3)-(4)*. Specifically, by Lemma D.2,*

$$\|q(\cdot,z)\|^2_{\mathcal{B}^{-1}_{2,1}([0,1]^{d_Y})} \leq C_{\mathrm{emb}}^2\|q(\cdot,z)\|^2_{L^2([0,1]^{d_Y})} \leq C_{\mathrm{emb}}^2\gamma^2 \qquad \forall z \in [0,1]^{d_Z}.$$

We analyze the two terms separately.

(Term A). *By Lemma D.2,*

$$\begin{aligned}
&\int_{[0,1]^{d_Z}} \|\hat{q}(\cdot,z) - q(\cdot,z)\|^2_{\mathcal{B}^{-1}_{2,1}([0,1]^{d_Y})}\,\mathrm{d}z\\
\leq& C_{\mathrm{emb}}\int_{[0,1]^{d_Z}} \|\hat{q}(\cdot,z) - q(\cdot,z)\|^2_{L_2([0,1]^{d_Y})}\mathrm{d}z\\
\leq& C_{\mathrm{emb}}\|\hat{q} - q\|^2_{L_2([0,1]^{d_Y+d_Z})}.
\end{aligned}$$

(Term B). *Let $\delta(y,z) = \hat{q}(y,z) - q(y,z)$. Then, by the Cauchy–Schwarz inequality,*

$$\begin{aligned}\int_{[0,1]^{d_Z}} (\hat{q}_Z(z) - q_Z(z))^2 \,\mathrm{d}z &= \int_{[0,1]^{d_Z}} \Big(\int_{[0,1]^{d_Y}} \delta(y,z)\mathrm{d}y \Big)^2 dz \\ &\leq \int_{[0,1]^{d_Z}} \Big(\int_{[0,1]^{d_Y}} 1^2 \mathrm{d}y \Big) \Big(\int_{[0,1]^{d_Y}} \delta(y,z)^2 \mathrm{d}y \Big) \mathrm{d}z \\ &\leq \int_{[0,1]^{d_Y+d_Z}} (\hat{q}(y,z) - q(y,z))^2 \mathrm{d}y\mathrm{d}z, \\ &= \|\hat{q} - q\|^2_{L_2([0,1]^{d_Y+d_Z})}.\end{aligned}$$

Note that Eq. (3)-(4) *and $\int \tilde{q} = 1$ a.s. (Lemma D.3) implies $\hat{q} = \tilde{q}$, thus $\hat{q}$ is the wavelet (linear) projection of the form*

$$\hat{q} = \tilde{q} = \sum_{\zeta \in \Phi} \widehat{\theta}_\zeta \, \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi_j} \widehat{\theta}_\xi \, \xi.$$

To bound $\mathbb{E}\left[\|\hat{q} - q\|^2_{L_2([0,1]^{d_Y+d_Z})} \right]$, we introduce the following result

Proposition D.6 *Under Assumption 2.5-2.6, there exists a constant $C^\dagger$ depends on $q, \gamma, s, d_Y + d_Z$ such that*

$$\mathbb{E}\left[\int_{[0,1]^{d_Y+d_Z}} (\tilde{q}(y,z) - q(y,z))^2 \mathrm{d}y\mathrm{d}z \right] \leq C^\dagger n^{-\frac{2s}{2s+d_Y+d_Z}}.$$

Combining the above results implies the desired result.

D.2 Proof of Proposition 3.4

Proof D.7 (Proof of Proposition 3.4) *By Lin et al. [2026, Proposition 2],*

$$\mathsf{CW}_h(P,Q) = \int \mathsf{W}_h(P_Y^z, Q_Y^z) \mathrm{d}P_Z(z).$$

Therefore, we have

$$\begin{aligned}&\left| \mathsf{CW}_h(P,Q) - \mathsf{CW}_h(\widehat{P},\widehat{Q}) \right| \\ =& \left| \int \mathsf{W}_h \left(P_Y^z, Q_Y^z\right) \mathrm{d}P_Z(z) - \int \mathsf{W}_h \left(\widehat{P}_Y^z, \widehat{Q}_Y^z\right) \mathrm{d}\widehat{P}_Z(z) \right| \\ \leq& \underbrace{\left| \int \mathsf{W}_h \left(P_Y^z, Q_Y^z\right) \mathrm{d}P_Z(z) - \int \mathsf{W}_h \left(P_Y^z, \widehat{Q}_Y^z\right) \mathrm{d}\widehat{P}_Z(z) \right|}_{\textit{(Term A)}} \\ &+ \underbrace{\left| \int \mathsf{W}_h \left(P_Y^z, \widehat{Q}_Y^z\right) \mathrm{d}\widehat{P}_Z(z) - \int \mathsf{W}_h \left(\widehat{P}_Y^z, \widehat{Q}_Y^z\right) \mathrm{d}\widehat{P}_Z(z) \right|}_{\textit{(Term B)}}.\end{aligned}$$

*For **(Term A)**: we highlight that we consider the optimal coupling $\pi_{Z,Z'} \in \mathcal{P}(\mathcal{Z}^2)$ that is attained in the optimal transport $\mathsf{W}_1(P_Z, \widehat{P}_Z)$ to connect P_Z and $\widehat{P}_Z$, then*

$$
\begin{aligned}
(\textit{Term A}) &= \left| \int \mathsf{W}_h\left(P_Y^z, Q_Y^z\right) \mathrm{d}P_Z(z) - \int \mathsf{W}_h\left(P_Y^z, \widehat{Q}_Y^z\right) \mathrm{d}\widehat{P}_Z(z) \right| \\
&= \left| \int \mathsf{W}_h\left(P_Y^z, Q_Y^z\right) - \mathsf{W}_h\left(P_Y^{z'}, \widehat{Q}_Y^{z'}\right) \mathrm{d}\pi_{Z,Z'}(z, z') \right| \\
&\leq \int \left| \mathsf{W}_h\left(P_Y^z, Q_Y^z\right) - \mathsf{W}_h\left(P_Y^{z'}, \widehat{Q}_Y^{z'}\right) \right| \mathrm{d}\pi_{Z,Z'}(z, z') \\
&\leq L_h \int \mathsf{W}_1\left(P_Y^z, P_Y^{z'}\right) + \mathsf{W}_1\left(Q_Y^z, \widehat{Q}_Y^{z'}\right) \mathrm{d}\pi_{Z,Z'}(z, z') \\
&\leq L_h \int L_p \|z - z'\|_2 + \mathsf{W}_1\left(Q_Y^z, \widehat{Q}_Y^{z'}\right) \mathrm{d}\pi_{Z,Z'}(z, z') \\
&\leq L_h \int 2L_p \|z - z'\|_2 + \mathsf{W}_1\left(Q_Y^{z'}, \widehat{Q}_Y^{z'}\right) \mathrm{d}\pi_{Z,Z'}(z, z') \\
&\leq 2L_h L_p \mathsf{W}_1(P_Z, \widehat{P}_Z) + L_h \int \mathsf{W}_1\left(Q_Y^{z'}, \widehat{Q}_Y^{z'}\right) \mathrm{d}\widehat{P}_Z(z').
\end{aligned}
$$

Here, the second inequality is due to Lin et al. [2026, Lemma 5]; the third inequality is due to the condition $\mathsf{W}_1(P_Y^z, P_Y^{z'}) \leq L_p \|z - z'\|_2$ stated in Proposition 3.4; the last inequality is due to the definition of π, which achieves the optimality in $\mathsf{W}_1(P_Z, \widehat{P}_Z)$.

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

$$
\begin{aligned}
(\textit{Term B}) &= \left| \int \mathsf{W}_h\left(P_Y^z, \widehat{Q}_Y^z\right) \mathrm{d}\widehat{P}_Z(z) - \int \mathsf{W}_h\left(\widehat{P}_Y^z, \widehat{Q}_Y^z\right) \mathrm{d}\widehat{P}_Z(z) \right| \\
&\leq \int \left| \mathsf{W}_h\left(P_Y^z, \widehat{Q}_Y^z\right) - \mathsf{W}_h\left(\widehat{P}_Y^z, \widehat{Q}_Y^z\right) \right| \mathrm{d}\widehat{P}_Z(z) \\
&\leq L_h \int \mathsf{W}_1(P_Y^z, \widehat{P}_Y^z) \mathrm{d}\widehat{P}_Z(z),
\end{aligned}
$$

where the last inequality is due to Lin et al. [2026, Lemma 5].

Jointly, we get

$$
|\mathsf{CW}_h(\widehat{P}, \widehat{Q}) - \mathsf{CW}_h(P, Q)| \leq 2L_h L_p \mathsf{W}_1(P_Z, \widehat{P}_Z) + L_h \int \mathsf{W}_1(P_Y^z, \widehat{P}_Y^z) + \mathsf{W}_1(Q_Y^z, \widehat{Q}_Y^z)\, \widehat{P}_Z(\mathrm{d}z).
$$

D.3 Proof of Theorem 3.6

Proof D.8 (Proof of Theorem 3.6) *We first verify the conditions of Proposition 3.4 under Assumption 2.5 to 2.7.*

(i) Lipschitz continuity of h is assumed by Assumption 2.7.

(ii) Under Assumption 2.5-2.6, we have

$$\begin{aligned}
\mathsf{W}_1(Q_Y^z, Q_Y^{z'}) &\le \frac{\sqrt{d_Y}}{2} \int_{[0,1]^{d_Y}} |q(y|z) - q(y|z')| \mathrm{d}y \\
&\le \frac{\sqrt{d_Y}}{2} \left(\int_{[0,1]^{d_Y}} \frac{|q(y,z) - q(y,z')|}{q_Z(z)} \mathrm{d}y \right. \\
&\qquad \left. + \frac{|q_Z(z) - q_Z(z')|}{q_Z(z') q_Z(z)} \int_{[0,1]^{d_Y}} |q(y,z')| \mathrm{d}y \right) \\
&\le \frac{\sqrt{d_Y}}{2} \left(\gamma \int_{[0,1]^{d_Y}} |q(y,z) - q(y,z')| \mathrm{d}y + \gamma^3 |q_Z(z) - q_Z(z')| \right) \\
&\le \frac{\sqrt{d_Y}}{2} \left(\gamma \int_{[0,1]^{d_Y}} |q(y,z) - q(y,z')| \mathrm{d}y \right. \\
&\qquad \left. + \gamma^3 \int_{[0,1]^{d_Y}} |q(y,z) - q(y,z')| \mathrm{d}y \right) \\
&= \frac{\gamma(1+\gamma^2)\sqrt{d_Y}}{2} \int_{[0,1]^{d_Y}} |q(y,z) - q(y,z')| \mathrm{d}y \\
&\le \frac{\gamma(1+\gamma^2)\sqrt{d_Y}}{2} \left(\max_{(y,z) \in [0,1]^{d_Y + d_Z}} \|\partial_z q(y,z)\|_2 \right) \|z - z'\|_2.
\end{aligned}$$

Therefore, we set $L_p := \frac{\gamma(1+\gamma^2)\sqrt{d_Y}}{2} \max_{(y,z) \in [0,1]^{d_Y + d_Z}} \|\partial_z q(y,z)\|_2 \vee \|\partial_z p(y,z)\|_2 < \infty$.

Finally, apply Proposition 3.4, we have

$$\mathbb{E}[|V_{\mathrm{c}} - \widehat{V}_{\mathrm{c},n}|] \le 2 L_h L_p \mathbb{E}[\mathsf{W}_1(P_Z, \widehat{R}_n)] + L_h \mathbb{E}\left[\int \mathsf{W}_1(P_Y^z, \widehat{P}_{n,Y}^z) + \mathsf{W}_1(Q_Y^z, \widehat{Q}_{n,Y}^z) \, \widehat{R}_n(\mathrm{d}z) \right].$$

For the first term, applying Jensen's inequality, we get

$$\begin{aligned}
\mathbb{E}[\mathsf{W}_1(P_Z, \widehat{R}_n)] \le& \frac{1}{2} (\mathbb{E}[\mathsf{W}_1(P_Z, \widehat{P}_{n,Z})] + \mathbb{E}[\mathsf{W}_1(P_Z, \widehat{Q}_{n,Z})]) \\
\le& \frac{1}{2} \left((\mathbb{E}[\mathsf{W}_2^2(P_Z, \widehat{P}_{n,Z})])^{\frac{1}{2}} + (\mathbb{E}[\mathsf{W}_2^2(P_Z, \widehat{Q}_{n,Z})])^{\frac{1}{2}} \right).
\end{aligned}$$

Apply a similar reasoning as bounding the (Term B) in the proof of Proposition D.4: Under Assumption 2.5, applying Lemma D.1 - D.2, we get

$$\mathbb{E}[\mathsf{W}_2^2(P_Z, \widehat{P}_{n,Z})] \le C_{\mathrm{wb}}^2 C_{\mathrm{emb}}^2 \gamma \mathbb{E}[\|p - \hat{p}\|_2^2].$$

Then, applying Proposition D.6, we get

$$\mathbb{E}[\mathsf{W}_2^2(P_Z, \widehat{P}_{n,Z})] \lesssim n^{-\frac{2s}{2s + d_Y + d_Z}}.$$

For the second term, applying Jensen's inequality and Proposition D.4, we get

$$\mathbb{E}\left[\int \mathsf{W}_1(Q_Y^z, \widehat{Q}_{n,Y}^z) \, \widehat{R}_n(\mathrm{d}z) \right] \le \left(\mathbb{E}\left[\int \mathsf{W}_2^2(Q_Y^z, \widehat{Q}_{n,Y}^z) \, \widehat{R}_n(\mathrm{d}z) \right] \right)^{\frac{1}{2}} \lesssim n^{-\frac{s}{2s + d_Y + d_Z}}.$$

As a result, we get

$$\mathbb{E}[|V_{\mathrm{c}} - \widehat{V}_{\mathrm{c},n}|] \le C L_h n^{-\frac{s}{2s + d_Y + d_Z}},$$

where C is a constant that depends on $P, Q, \gamma, s, d_Y, d_Z$.

D.4 Proof of Theorem 3.7

As a preparation, we introduce the following result corresponding to Proposition D.6.

Proposition D.9 (High probability bound) *Under Assumption 2.5-2.6, there exists a constant $C^\dagger$ depending on $q, \gamma, s, d_Y + d_Z$ such that, for $1 \le t \le n$,*

$$\mathbb{P}\left(\|\tilde{q} - q\|_2^2 \le C^\dagger n^{-\frac{2s}{2s+d_Y+d_Z}} t\right) \ge 1 - e^{-t}.$$

Proof D.10 (Proof of Theorem 3.7) *Similar to the proof of Theorem 3.6, apply Proposition 3.4, we have*

$$|V_{\mathrm{c}} - \widehat{V}_{\mathrm{c},n}| \le 2L_h L_p \mathsf{W}_1(P_Z, \widehat{R}_n) + L_h \int \mathsf{W}_1(P_Y^z, \widehat{P}_{n,Y}^z) + \mathsf{W}_1(Q_Y^z, \widehat{Q}_{n,Y}^z)\, \widehat{R}_n(\mathrm{d}z).$$

Using the same argument as in the proof of Proposition D.4, we have

$$\mathsf{W}_2^2(P_Z, \widehat{R}_n) \lesssim \|\hat{p} - p\|_2^2 + \|\hat{q} - q\|_2^2,$$
$$\int \mathsf{W}_2^2(P_Y^z, \widehat{P}_{n,Y}^z) + \mathsf{W}_2^2(Q_Y^z, \widehat{Q}_{n,Y}^z)\, \widehat{R}_n(\mathrm{d}z) \lesssim \|\hat{p} - p\|_2^2 + \|\hat{q} - q\|_2^2.$$

By Lemma D.3, with probability at least $1 - c_b/n^2$,

$$\hat{p} = \tilde{p}, \quad \hat{q} = \tilde{q}.$$

For $1 \le t \le n$, applying Proposition D.9 to $\tilde{p}, \tilde{q}$, we obtain, with probability at least $1 - 2e^{-t}$,

$$\|\tilde{p} - p\|_2^2 + \|\tilde{q} - q\|_2^2 \lesssim n^{-\frac{2s}{2s+d_Y+d_Z}} t.$$

Hence, on the intersection of this event with the event from Lemma D.3,

$$\|\tilde{p} - p\|_2 + \|\tilde{q} - q\|_2 \lesssim n^{-\frac{s}{2s+d_Y+d_Z}} t^{1/2}.$$

Therefore

$$|V_{\mathrm{c}} - \widehat{V}_{\mathrm{c},n}| \lesssim L_h n^{-\frac{s}{2s+d_Y+d_Z}} t^{1/2}.$$

Combining the probability estimates by a union bound, we conclude that there exists a constant $C > 0$ independent of n, δ, such that

$$\mathbb{P}\left(|V_{\mathrm{c}} - \widehat{V}_{\mathrm{c},n}| \le C L_h n^{-\frac{s}{2s+d_Y+d_Z}} t^{1/2}\right) \ge 1 - 2e^{-t} - \frac{c_b}{n^2}.$$

For fixed $\delta \in (0,1)$, take $t = \log(4/\delta)$ and n sufficiently large that $t \le n$ and $c_b/n^2 \le \delta/2$. Then $2e^{-t} + c_b/n^2 \le \delta$, which proves the stated high-probability bound.

D.5 Proof of Proposition 3.8

The stability bound under quadratic objective stems from the case for unconditional Wasserstein-2 distance.

Proposition D.11 (Proposition 12 in [Manole et al., 2024]) *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$. Assume that the Brenier potential φ_0 between P, Q is a convex function such that $\varphi_0 \in \mathcal{C}^2([0,1]^d)$ and there exists $\lambda > 0$ satisfying the uniform Hessian bounds*

$$\frac{1}{\lambda} I_d \preceq \nabla^2 \varphi_0(x) \preceq \lambda I_d \quad \textit{for all } x \in [0,1]^d.$$

For any $\widehat{P}, \widehat{Q} \in \mathcal{P}([0,1]^d)$,

$$0 \le \mathsf{W}_2^2(\widehat{P}, \widehat{Q}) - \mathsf{W}_2^2(P, Q) - \int \phi_0\, d(\widehat{P} - P) - \int \psi_0\, d(\widehat{Q} - Q) \;\le\; \lambda\, (\mathsf{W}_2(\widehat{P}, P) + \mathsf{W}_2(\widehat{Q}, Q))^2, \tag{5}$$

where $\phi_0(x) = \|x\|^2 - 2\varphi_0(x)$ and $\psi_0(x) = \|x\|^2 - 2\varphi_0^*(x)$, where φ_0^* is the convex conjugate of φ_0.

Remark D.12 *We compare the conditional and the unconditional stability result. The upper bound in Proposition 3.8* $\int \mathsf{W}_2^2(\widehat{P}_Y^z, P_Y^z)\,\widehat{R}(\mathrm{d}z) + \int \mathsf{W}_2^2(\widehat{Q}_Y^z, Q_Y^z)\,\widehat{R}(\mathrm{d}z)$ *is no smaller than* $\mathsf{W}_2^2\big(\widehat{P}_Y^z \times \widehat{R},\ P_Y^z \times \widehat{R}\big) + \mathsf{W}_2^2\big(\widehat{Q}_Y^z \times \widehat{R},\ Q_Y^z \times \widehat{R}\big)$ *by Lemma D.13, where the latter is similar to the upper bound* $\mathsf{W}_2^2(\widehat{P}, P) + \mathsf{W}_2^2(\widehat{Q}, Q)$ *in Proposition D.11. A larger upper bound in Proposition 3.8 is consistent with the fact that COT is intrinsically more challenging than the unconditional counterpart.*

Lemma D.13 (Conditional Wasserstein distance V.S. Wasserstein distance) *Let* P *and* Q *be probability measures on* $\mathcal{Y} \times \mathcal{Z}$ *with the same marginal distribution* $P_Z = Q_Z$ *on* $\mathcal{Z}$*. For* $z \in \mathcal{Z}$*, let* P_Y^z *and* Q_Y^z *denote the corresponding conditional distributions on* $\mathcal{Y}$*. Consider the quadratic cost on* $\|y - y'\|^2$*. Then*

$$\mathbb{E}_{Z\sim P_Z}\big[\mathsf{W}_2^2\big(P_Y^Z,\, Q_Y^Z\big)\big] \ \geq\ \mathsf{W}_2^2(P, Q).$$

To obtain the strong convexity of the conditional potential functions in our setting, we shall leverage the following result.

Lemma D.14 (Gigli [2011, Corollary 3.2]) *Assume* $P, Q \in \mathcal{P}([0,1]^d)$ *with densities* p, q*. Assume the Brenier's potential* φ_0 *between* P, Q *satisfies* $\varphi_0 \in C^2([0,1]^d)$ *and* $\gamma^{-1} \leq p(x), q(x) \leq \gamma\ \forall x \in [0,1]^d$ *for a constant* $\gamma > 0$*. Then there exists a constant* $\lambda > 0$*, depending only on* γ *and* $\|\varphi_0\|_{C^2([0,1]^d)}$*, such that* φ_0 *is* λ*-strongly convex.*

Now we are ready to prove the proposition.

Proof D.15 (Proof of Proposition 3.8) *In the following, we denote* $\phi_z(y) := \phi(y, z)$*,* $\psi_z(y) := \psi(y, z)$*. In this proof only,* $\widehat{P}$ *and* $\widehat{Q}$ *denote the aligned measures* $\widehat{P}_n^\dagger$ *and* $\widehat{Q}_n^\dagger$*, respectively.*

First, applying Lemma D.14 under Assumption 2.5 (γ*) and also by the assumption that* $\|\varphi(\cdot, z)\|_{\mathcal{C}^2([0,1]^{d_Y})}$ *is bounded by* C_2 *uniformly for* z*, we get, there exists a constant* $\lambda > 0$ *that depends on* γ *and* C_2*, such that for any* $(y, z) \in [0,1]^{d_Y + d_Z}$*,*

$$\lambda^{-1} I_{d_Y} \preceq \nabla_y^2 \varphi(y, z) \preceq \lambda I_{d_Y}.$$

For all $z \in [0,1]^{d_Z}$*, under Assumption 2.5 and with constant* $\lambda > 0$*, apply Proposition D.11 to* (P_Y^z, Q_Y^z)*, we get*

$$\begin{aligned} 0 \leq& \underbrace{\mathsf{W}_2(\widehat{P}_Y^z, \widehat{Q}_Y^z)^2 - \mathsf{W}_2(P_Y^z, Q_Y^z)^2}_{:=\textit{Term (a)}} - \underbrace{\int \phi_z \mathrm{d}(\widehat{P}_Y^z - P_Y^z) + \int \psi_z \mathrm{d}(\widehat{Q}_Y^z - Q_Y^z)}_{:=\textit{Term (b)}} \\ \leq& \lambda \underbrace{\Big(\mathsf{W}_2(\widehat{P}_Y^z, P_Y^z) + \mathsf{W}_2(\widehat{Q}_Y^z, Q_Y^z)\Big)^2}_{:=\textit{Term (c)}}, \end{aligned}$$

for any $z \in [0,1]^{d_Z}$*. To derive the final result, we integrate the term (a) to (c) in the inequality above with respect to the empirical marginal distribution* $\widehat{R}(\mathrm{d}z) = \widehat{P}_Z(\mathrm{d}z) = \widehat{Q}_Z(\mathrm{d}z)$*. Here we use the alignment property that* $\widehat{P}$ *and* $\widehat{Q}$ *share the same marginal distribution of* Z*.*

Term (a). *By the alignment property of the estimated densities* $\widehat{R}(\mathrm{d}z) = \widehat{P}_Z(\mathrm{d}z) = \widehat{Q}_Z(\mathrm{d}z)$*, we have*

$$\int \mathsf{W}_2(\widehat{P}_Y^z, \widehat{Q}_Y^z)^2 \widehat{R}(\mathrm{d}z) = \mathsf{CW}_2(\widehat{P}, \widehat{Q})^2.$$

In addition, by the alignment property of the true densities $R(\mathrm{d}z) = P_Z(\mathrm{d}z) = Q_Z(\mathrm{d}z)$*,*

$$\begin{aligned} \int \mathsf{W}_2(P_Y^z, Q_Y^z)^2 \widehat{R}(\mathrm{d}z) &= \int \mathsf{W}_2(P_Y^z, Q_Y^z)^2 R(\mathrm{d}z) + \int \mathsf{W}_2(P_Y^z, Q_Y^z)^2 (\widehat{R}(\mathrm{d}z) - R(\mathrm{d}z)) \\ &= \mathsf{CW}_2(P, Q)^2 + \int \mathsf{W}_2(P_Y^z, Q_Y^z)^2 (\widehat{R}(\mathrm{d}z) - R(\mathrm{d}z)). \end{aligned} \tag{6}$$

***Term (b)**. We have*

$$\begin{aligned}
&\int\int \phi_z(y)(\widehat{P}_Y^z(\mathrm{d}y) - P_Y^z(\mathrm{d}y))\widehat{R}(\mathrm{d}z) \\
=&\int\int \phi_z(y)(\widehat{P}_Y^z(\mathrm{d}y)\widehat{R}(\mathrm{d}z) - P_Y^z(\mathrm{d}y)\widehat{R}(\mathrm{d}z)) \\
=&\int\int \phi_z(y)(\widehat{P}_Y^z(\mathrm{d}y)\widehat{R}(\mathrm{d}z) - P_Y^z(\mathrm{d}y)R(\mathrm{d}z)) - \int\int \phi_z(y)P_Y^z(\mathrm{d}y)(\widehat{R}(\mathrm{d}z) - R(\mathrm{d}z)) \\
=&\int \phi \mathrm{d}(\widehat{P} - P) - \int\int \phi_z(y)P_Y^z(\mathrm{d}y)(\widehat{R}(\mathrm{d}z) - R(\mathrm{d}z)),
\end{aligned}$$

where we use the definition of ϕ and ψ. Similarly,

$$\begin{aligned}
&\int\int \psi_z(y)(\widehat{Q}_Y^z(\mathrm{d}y) - Q_Y^z(\mathrm{d}y))\widehat{R}(\mathrm{d}z) \\
=&\int \psi \mathrm{d}(\widehat{Q} - Q) - \int\int \psi_z(y)Q_Y^z(\mathrm{d}y)(\widehat{R}(\mathrm{d}z) - R(\mathrm{d}z)).
\end{aligned}$$

By the optimality relation of COT,

$$\mathsf{W}_2(P_Y^z, Q_Y^z)^2 = \int \phi_z(y)P_Y^z(\mathrm{d}y) + \int \psi_z(y)Q_Y^z(\mathrm{d}y).$$

Therefore, we have

$$\begin{aligned}
&\int\int \phi_z(y)(\widehat{P}_Y^z(\mathrm{d}y) - P_Y^z(\mathrm{d}y))\widehat{R}(\mathrm{d}z) + \int\int \psi_z(y)(\widehat{Q}_Y^z(\mathrm{d}y) - Q_Y^z(\mathrm{d}y))\widehat{R}(\mathrm{d}z) \\
=&\int \phi \mathrm{d}(\widehat{P} - P) + \int \psi \mathrm{d}(\widehat{Q} - Q) - \int \mathsf{W}_2(P_Y^z, Q_Y^z)^2(\widehat{R}(\mathrm{d}z) - R(\mathrm{d}z)).
\end{aligned} \tag{7}$$

***Term (c)**. The integration of Term (c) over $\widehat{R}(dz)$ together with the inequality $(a+b)^2 \le 2a^2 + 2b^2$ give the right hand side of the inequality of Proposition 3.8.*

The marginal correction $\int \mathsf{W}_2(P_Y^z, Q_Y^z)^2(\widehat{R}(\mathrm{d}z) - R(\mathrm{d}z))$ cancels when the integrated linear term is subtracted from the integrated value difference, by (6) *and* (7)*. Combining term (a), term (b), and term (c), we get the desired result.*

D.6 Proof of Theorem 3.9

First we need a lemma to prove the smoothness of ϕ, ψ.

Lemma D.16 (Smooth Kantorovich potential) *Under Assumption 2.5-2.6, when $2s > d_Y + d_Z$, $\varphi \in \mathcal{C}^{s+1}([0,1]^{d_Y+d_Z})$, then $\phi, \psi \in \mathcal{C}^{s+1}([0,1]^{d_Y+d_Z})$.*

Proof D.17 (Proof of Theorem 3.9) *Since $2s > d_Y + d_Z \ge 2$, we have $s > 1$ and hence $\varphi \in \mathcal{C}^2([0,1]^{d_Y+d_Z})$. We may therefore apply Proposition 3.8 to obtain*

$$\begin{aligned}
0 \le& \mathsf{CW}_2(\widehat{P}_n^\dagger, \widehat{Q}_n^\dagger)^2 - \mathsf{CW}_2(P,Q)^2 - \int \phi \mathrm{d}(\widehat{P}_n^\dagger - P) - \int \psi \mathrm{d}(\widehat{Q}_n^\dagger - Q) \\
\le& 2\lambda \int \mathsf{W}_2^2(\widehat{P}_Y^z, P_Y^z) + \mathsf{W}_2^2(\widehat{Q}_Y^z, Q_Y^z)\widehat{R}_n(\mathrm{d}z),
\end{aligned}$$

where λ is a constant that depends on $\gamma, \|\varphi\|_{\mathcal{C}^2}$.

By Proposition D.4, we have

$$\mathbb{E}\left[\int \mathsf{W}_2^2(\widehat{P}_Y^z, P_Y^z)\,\widehat{R}(\mathrm{d}z)\right] \vee \mathbb{E}\left[\int \mathsf{W}_2^2(\widehat{Q}_Y^z, Q_Y^z)\,\widehat{R}(\mathrm{d}z)\right] \lesssim n^{-\frac{2s}{2s+d_Y+d_Z}}.$$

Therefore, when $2s > d_Y + d_Z$,

$$\int \mathsf{W}_2^2(\widehat{P}_Y^z, P_Y^z) + \mathsf{W}_2^2(\widehat{Q}_Y^z, Q_Y^z)\widehat{R}_n(\mathrm{d}z) = o_p(n^{-1/2}).$$

It remains to show that

$$\sqrt{n}\left(\int \phi \mathrm{d}(\widehat{P}_n^\dagger - P) + \int \psi \mathrm{d}(\widehat{Q}_n^\dagger - Q)\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2).$$

We use $\hat{r}(z)$ to denote the density of $\widehat{R}_n(z)$. We use $\mathcal{Y} \times \mathcal{Z}$ and $[0,1]^{d_Y+d_Z}$ interchangeably. Since $\mathcal{Y} \times \mathcal{Z}$ is compact and $p(y,z)$, $q(y,z)$, $\phi(y,z)$, $\psi(y,z) \in \mathcal{C}^{s+1}(\mathcal{Y} \times \mathcal{Z})$, there exists $M > 0$ such that $p(y,z)$, $q(y,z)$, $|\phi(y,z)|$, $|\psi(y,z)| < M$ for any $(y,z) \in \mathcal{Y} \times \mathcal{Z}$.

According to Definition 3.1, we can decompose $\int \phi \mathrm{d}(\widehat{P}_n^\dagger - P)$ as

$$\int \phi \mathrm{d}(\widehat{P}_n^\dagger - P) = \underbrace{\int \phi \mathrm{d}(\widehat{P}_n^\dagger - \widehat{P}_n)}_{:=\textit{Term (a)}} + \underbrace{\int \phi \mathrm{d}(\widehat{P}_n - P)}_{:=\textit{Term (b)}}.$$

***Term (a)**. In the following, we ignore the dependence on n.*

$$\begin{aligned}
\int \phi \mathrm{d}(\widehat{P}_n^\dagger - \widehat{P}_n) &= \int \phi(y,z)\left(\frac{\hat{p}(y,z)}{\hat{p}_Z(z)}\hat{r}(z) - \frac{\hat{p}(y,z)}{\hat{p}_Z(z)}\hat{p}_Z(z)\right)\mathrm{d}y\mathrm{d}z \quad (\textit{Definition 3.1}) \\
&= \int \phi(y,z)\frac{\hat{p}(y,z)}{\hat{p}_Z(z)}\left(\hat{r}(z) - \hat{p}_Z(z)\right)\mathrm{d}y\mathrm{d}z \\
&= \underbrace{\int \phi(y,z)\frac{p(y,z)}{p_Z(z)}\left(\hat{r}(z) - \hat{p}_Z(z)\right)\mathrm{d}y\mathrm{d}z}_{:=\textit{Term (a.1)}} \\
&\quad + \underbrace{\int \phi(y,z)\left(\frac{\hat{p}(y,z)}{\hat{p}_Z(z)} - \frac{p(y,z)}{p_Z(z)}\right)\left(\hat{r}(z) - \hat{p}_Z(z)\right)\mathrm{d}y\mathrm{d}z}_{:=\textit{Term (a.2)}}.
\end{aligned}$$

Term (a.1) is relatively straightforward as $\hat{r}(z)$ and $\hat{p}_Z(z)$ are based on marginalizing standard wavelet-based density estimators, and we postpone the analysis. For Term (a.2), by the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left(\int \phi(y,z)\left(\frac{\hat{p}(y,z)}{\hat{p}_Z(z)} - \frac{p(y,z)}{p_Z(z)}\right)\left(\hat{r}(z) - \hat{p}_Z(z)\right)\mathrm{d}y\mathrm{d}z\right)^2 \\
\leq & M^2 \underbrace{\int \left(\frac{\hat{p}(y,z)}{\hat{p}_Z(z)} - \frac{p(y,z)}{p_Z(z)}\right)^2 \mathrm{d}y\mathrm{d}z}_{:=\textit{Term (a.2.1)}} \underbrace{\int \left(\hat{r}(z) - \hat{p}_Z(z)\right)^2 \mathrm{d}z}_{:=\textit{Term (a.2.2)}}.
\end{aligned}$$

For Term (a.2.1), we decompose it as

$$\begin{aligned}
\textit{Term (a.2.1)} &= \int \left(\frac{\hat{p}(y,z)}{\hat{p}_Z(z)} - \frac{p(y,z)}{\hat{p}_Z(z)} + \frac{p(y,z)}{\hat{p}_Z(z)} - \frac{p(y,z)}{p_Z(z)}\right)^2 \mathrm{d}y\mathrm{d}z \\
&\leq 2\int \frac{1}{\hat{p}_Z(z)^2}\left(\hat{p}(y,z) - p(y,z)\right)^2 \mathrm{d}y\mathrm{d}z + 2\frac{M^2}{\hat{p}_Z(z)^2 p_Z(z)^2}\int \left(\hat{p}_Z(z) - p_Z(z)\right)^2 \mathrm{d}z.
\end{aligned} \tag{8}$$

Then, we apply the same reasoning as in the proof of Proposition D.4 that with probability at least $1 - c/n^2$ for some constant c, $\hat{p}_Z(z) \geq \gamma^{-1}/2 \, \forall z$. Thus,

$$\textit{Term (a.2.1)} \leq 2(2\gamma)^4 \int \left(\hat{p}(y,z) - p(y,z)\right)^2 \mathrm{d}y\mathrm{d}z + 2M^2 \int \left(\hat{p}_Z(z) - p_Z(z)\right)^2 \mathrm{d}z.$$

Again, by the same reasoning as in the proof of Proposition D.4, applying Proposition D.6, we get

$$\mathbb{E}[\textit{Term (a.2.1)}] \lesssim n^{-\frac{2s}{2s+d_Y+d_Z}}.$$

When $2s > d_Y + d_Z$, Term (a.2.1) $= o_p(n^{-1/2})$.

For Term (a.2.2), since

$$\begin{aligned}\hat{r}(z)-\hat{p}_Z(z) &= (\hat{r}(z)-p_Z(z))-(\hat{p}_Z(z)-p_Z(z)),\\ &= \frac{1}{2}(\hat{p}_Z(z)-p_Z(z))+\frac{1}{2}(\hat{q}_Z(z)-p_Z(z))-(\hat{p}_Z(z)-p_Z(z)) \quad (\textit{Definition 3.1}).\end{aligned}$$

Then, similarly, we have, when $2s > d_Y + d_Z$, Term (a.2.2) $= o_p(n^{-1/2})$.

Jointly, and we get Term (a.2) $= o_p(n^{-1/2})$, which implies that in Term (a.2) is ignorable.

Next, we bring Term (b), Term (a.1), as well as the counterparts for Q together. By the construction of $\hat{p}^\dagger(y,z)$ (Definition 3.1), we have $\hat{r}(z) = (\hat{p}_Z(z)+\hat{q}_Z(z))/2$, which implies

$$-(\hat{r}(z)-\hat{p}_Z(z)) = \hat{r}(z)-\hat{q}_Z(z) = (\hat{p}_Z(z)-\hat{q}_Z(z))/2. \tag{9}$$

Then

$$\begin{aligned}&\int \phi \mathrm{d}(\widehat{P}_n^\dagger - P) + \int \psi \mathrm{d}(\widehat{Q}_n^\dagger - Q)\\ &= \underbrace{\int \phi(y,z)\frac{p(y,z)}{p_Z(z)}\left(\hat{r}(z)-\hat{p}_Z(z)\right)\mathrm{d}y\mathrm{d}z}_{\textit{Term (a.1)}} + \underbrace{\int \phi(y,z)(\hat{p}(y,z)-p(y,z))\mathrm{d}y\mathrm{d}z}_{\textit{Term (b)}}\\ &\quad + \int \psi(y,z)\frac{q(y,z)}{q_Z(z)}\left(\hat{r}(z)-\hat{q}_Z(z)\right)\mathrm{d}y\mathrm{d}z + \int \psi(y,z)(\hat{q}(y,z)-q(y,z))\mathrm{d}y\mathrm{d}z + o_p(n^{-1/2})\\ &= \int \left(\phi(y,z)\frac{p(y,z)}{p_Z(z)} - \psi(y,z)\frac{q(y,z)}{q_Z(z)}\right)\left(\frac{\hat{q}_Z(z)-\hat{p}_Z(z)}{2}\right)\mathrm{d}y\mathrm{d}z \qquad (\textit{by Eq. (9)})\\ &\quad + \int \phi(y,z)(\hat{p}(y,z)-p(y,z))\mathrm{d}y\mathrm{d}z + \int \psi(y,z)(\hat{q}(y,z)-q(y,z))\mathrm{d}y\mathrm{d}z + o_p(n^{-1/2})\\ &= \underbrace{\int -\left(\phi(y,z)\frac{p(y,z)}{p_Z(z)} - \psi(y,z)\frac{q(y,z)}{q_Z(z)}\right)\left(\frac{\hat{p}_Z(z)-p_Z(z)}{2}\right) + \phi(y,z)(\hat{p}(y,z)-p(y,z))\mathrm{d}y\mathrm{d}z}_{:=\textit{Term (c)}}\\ &\quad + \underbrace{\int \left(\phi(y,z)\frac{p(y,z)}{p_Z(z)} - \psi(y,z)\frac{q(y,z)}{q_Z(z)}\right)\left(\frac{\hat{q}_Z(z)-q_Z(z)}{2}\right) + \psi(y,z)(\hat{q}(y,z)-q(y,z))\mathrm{d}y\mathrm{d}z}_{:=\textit{Term (d)}}\\ &\quad + o_p(n^{-1/2})\end{aligned}$$

$$(p_Z(z) = q_Z(z) \textit{ and } \hat{q}_Z(z)-\hat{p}_Z(z) = \hat{q}_Z(z)-q_Z(z)-(\hat{p}_Z(z)-p_Z(z)))$$

For Term (c), as $\hat{p}_Z(z)-p_Z(z) = \int \hat{p}(y,z)-p(y,z)\mathrm{d}y$,

$$\begin{aligned}\textit{term (c)} &= \int\left(\int -\left(\phi(y',z)\frac{p(y',z)}{p_Z(z)} - \psi(y',z)\frac{q(y',z)}{q_Z(z)}\right)\mathrm{d}y'\right)\\ &\qquad \times \left(\int \frac{\widehat{p}(y,z)-p(y,z)}{2}\,\mathrm{d}y\right)\mathrm{d}z\\ &\quad + \int \phi(y,z)(\hat{p}(y,z)-p(y,z))\mathrm{d}y\mathrm{d}z\\ &= \int \underbrace{\left(\phi(y,z) - \frac{1}{2}\left(\int\left(\phi(y',z)\frac{p(y',z)}{p_Z(z)} - \psi(y',z)\frac{q(y',z)}{q_Z(z)}\right)\mathrm{d}y'\right)\right)}_{=\eta(y,z)}\\ &\qquad \times (\hat{p}(y,z)-p(y,z))\,\mathrm{d}y\mathrm{d}z.\end{aligned}$$

As $\phi(y,z)$, $\psi(y,z)$, $p(y,z)$, $q(y,z) \in \mathcal{C}^{s+1}(\mathcal{Y}\times\mathcal{Z})$ and $p_Z, q_Z \geq \gamma^{-1}$, we have $\eta(y,z) \in \mathcal{C}^{s+1}(\mathcal{Y}\times\mathcal{Z})$. By Lemma 11 (control bias) and Lemma 64 (central limit theorem) of [Manole et al., 2024], we establish the central limit theorem for $\sqrt{n}\times$ term (c) with the variance

$$\textit{Var}_P\left(\eta(Y,Z)\right) = \textit{Var}_P\left(\phi(Y,Z) - \frac{\mathbb{E}_P[\phi(Y,Z)\mid Z] - \mathbb{E}_Q[\psi(Y,Z)\mid Z]}{2}\right).$$

The same argument applies to term (d). Note that term (c) and term (d) are based on different observations and thus independent, which yields a central limit theorem for their sum scaled by $\sqrt{n}$ *with the asymptotic variance* $Var_P(\eta(Y,Z)) + Var_Q(\kappa(Y,Z))$.

D.7 Proof of Theorem 3.14

The proof of fast convergence rate depends on the data-generating mechanism only through Lemma D.3 and Proposition D.6. More specifically, in the proof of Lemma D.3 (Manole et al. [2024, Lemma 27 and proof of Lemma 29]), the key step needs to be justified is the Bernstein inequality under the covariate-shift setting. In the proof of Proposition D.6, the key step needs to be justified is the L_2 estimation error of the wavelet coefficients. In the following, we state the proof of these two arguments. Then, the proof of Theorem 3.14 follows exactly as in the proof of Theorem 3.6.

Proposition D.18 (Conditional expectation gap) *Under Assumptions 3.11–3.12,*

$$\mathbb{E}\Big[\widehat{\theta}^{\mathrm{dr}}_{a,\xi} \mid \widehat{e}, \widehat{m}_{a,\xi}\Big] - \theta_{a,\xi} = \mathbb{E}_Z\left[\left(1 - \frac{\pi_a(Z)}{\widehat{\pi}_a(Z)}\right)\{\widehat{m}_{a,\xi}(Z) - m_{a,\xi}(Z)\}|\hat{e}, \widehat{m}_{a,\xi}\right].$$

Consequently,

$$\sup_{\xi} \mathbb{E}\left(\mathbb{E}[\widehat{\theta}^{\mathrm{dr}}_{a,\xi} \mid \widehat{e}, \widehat{m}_{a,\xi}] - \theta_{a,\xi}\right)^2 = o(n^{-1}).$$

Proof D.19 *By unconfoundedness,*

$$\mathbb{E}[I(W=a)\xi(Y,Z) \mid Z] = \pi_a(Z) m_{a,\xi}(Z),$$

and

$$\mathbb{E}[I(W=a)\widehat{m}_{a,\xi}(Z) \mid Z] = \pi_a(Z)\widehat{m}_{a,\xi}(Z).$$

Therefore,

$$\mathbb{E}[\widehat{\theta}^{\mathrm{dr}}_{a,\xi} \mid \widehat{e}, \widehat{m}_{a,\xi}] = \mathbb{E}\left[\widehat{m}_{a,\xi}(Z) + \frac{\pi_a(Z)}{\widehat{\pi}_a(Z)}\{m_{a,\xi}(Z) - \widehat{m}_{a,\xi}(Z)\}\right].$$

Subtracting $\theta_{a,\xi} = \mathbb{E}[m_{a,\xi}(Z)]$ *gives*

$$\mathbb{E}_Z\left[\left(1 - \frac{\pi_a(Z)}{\widehat{\pi}_a(Z)}\right)\{\widehat{m}_{a,\xi}(Z) - m_{a,\xi}(Z)\}|\hat{e}, \widehat{m}_{a,\xi}\right],$$

which is the claimed identity. Since $\widehat{\pi}_a(z) \geq \eta/2$*, Cauchy–Schwarz yields*

$$\mathbb{E}\left(\mathbb{E}[\widehat{\theta}^{\mathrm{dr}}_{a,\xi} \mid \widehat{e}, \widehat{m}_{a,\xi}] - \theta_{a,\xi}\right)^2 \leq \frac{4}{\eta^2}\mathbb{E}[\|\widehat{e} - e\|^2_{L^2(P_Z)}\|\widehat{m}_{a,\xi} - m_{a,\xi}\|^2_{L^2(P_Z)}] = o(n^{-1}).$$

Proposition D.20 (L^2 coefficient error) *Assume Assumptions 3.11–3.13. Then*

$$\sup_{\xi} \mathbb{E}\left[(\widehat{\theta}^{\mathrm{dr}}_{a,\xi} - \theta_{a,\xi})^2\right] \leq \frac{C}{n}$$

where $C > 0$ *depends on* $\eta, M, \|p\|_\infty, \|q\|_\infty$.

Proof D.21 *Let* $O_i = (Y_i, Z_i, W_i)$*, and*

$$A_{a,\xi}(O) := \widehat{m}_{a,\xi}(Z) + \frac{I(W=a)}{\widehat{\pi}_a(Z)}\{\xi(Y,Z) - \widehat{m}_{a,\xi}(Z)\}.$$

Then

$$\widehat{\theta}^{\mathrm{dr}}_{a,\xi} = \frac{1}{n}\sum_{i=1}^{n} A_{a,\xi}(O_i).$$

Hence

$$\mathbb{E}[(\widehat{\theta}^{\mathrm{dr}}_{a,\xi} - \theta_{a,\xi})^2 \mid \widehat{e}, \widehat{m}_{a,\xi}] = \frac{1}{n} Var(A_{a,\xi}(O) \mid \widehat{e}, \widehat{m}_{a,\xi}) + \left(\mathbb{E}[A_{a,\xi}(O) \mid \widehat{e}, \widehat{m}_{a,\xi}] - \theta_{a,\xi}\right)^2.$$

Since $\widehat{\pi}_a \geq \eta/2$*,* $\|\hat{m}_{a,\xi}\|_{L^2(P_Z)} \leq M, a.s.$ *(Assumption 3.12), and* $\mathbb{E}[\xi(Y,Z)^2] \leq \|p\|_\infty + \|q\|_\infty$ *(see* (15)*), there is a constant* C *depending on* $\eta, M, \|p\|_\infty, \|q\|_\infty$*, such that*

$$Var(A_{a,\xi}(O) \mid \widehat{e}, \widehat{m}_{a,\xi}) \leq C, \quad a.s.$$

The squared bias term is bounded by Proposition D.18. Combining the two bounds gives the claim.

Proposition D.22 (Bernstein-type coefficient bound) *Under the assumptions of Proposition D.20, there exist constants $C_1, C_2, C_3 > 0$ depending on $\eta, b, M, \bar{C}, \|p\|_\infty, \|q\|_\infty$, such that*

$$\mathbb{P}\left(\left|\widehat{\theta}^{\mathrm{dr}}_{a,\xi} - \theta_{a,\xi}\right| \geq u\right) \leq C_1 \exp\left(-\frac{nu^2}{2C_2 + C_3 2^{dj} u}\right)$$

Proof D.23 *For the conditional expectation gap, we have*

$$\left|\mathbb{E}[\widehat{\theta}^{\mathrm{dr}}_{a,\xi} \mid \widehat{e}, \widehat{m}_{a,\xi}] - \theta_{a,\xi}\right| \leq \frac{2}{\eta}\|\widehat{e} - e\|_{L^2(P_Z)}\|\widehat{m}_{a,\xi} - m_{a,\xi}\|_{L^2(P_Z)}$$

Then, under Assumption 3.13,

$$\mathbb{P}\left(\left|\mathbb{E}[\widehat{\theta}^{\mathrm{dr}}_{a,\xi} \mid \widehat{e}, \widehat{m}_{a,\xi}] - \theta_{a,\xi}\right| \geq \frac{2}{\eta}u\right) \leq \exp(-n\bar{C}u).$$

Under Assumption 3.12,

$$|A_{a,\xi}(O)| \leq C_{\eta,b} 2^{(d_Y + d_Z)j}.$$

From the proof of Proposition D.20,

$$\mathit{Var}(A_{a,\xi}(O) \mid \widehat{e}, \widehat{m}_{a,\xi}) \leq C_2.$$

By scalar Bernstein's inequality,

$$\mathbb{P}\left(\left|\widehat{\theta}^{\mathrm{dr}}_{a,\xi} - \mathbb{E}[\widehat{\theta}^{\mathrm{dr}}_{a,\xi} \mid \widehat{e}, \widehat{m}_{a,\xi}]\right| \geq u \mid \widehat{e}, \widehat{m}_{a,\xi}\right) \leq 2\exp\left(-\frac{nu^2}{2C_2 + C_{\eta,b} 2^{(d_Y+d_Z)j} u}\right).$$

As a result,

$$\mathbb{P}\left(\left|\widehat{\theta}^{\mathrm{dr}}_{a,\xi} - \theta_{a,\xi}\right| \geq u\right) \lesssim 2\exp\left(-\frac{nu^2}{2C_2 + C_{\eta,b} 2^{(d_Y+d_Z)j} u}\right) + \exp(-n\frac{\eta\bar{C}}{2}u).$$

Let $d = d_Y + d_Z$ and define

$$\widetilde{C}_2 := C_2, \qquad \widetilde{C}_{\eta,b} := C_{\eta,b} \vee \frac{2}{\eta\bar{C}}.$$

Since $2^{dj} \geq 1$,

$$\frac{u^2}{2\widetilde{C}_2 + \widetilde{C}_{\eta,b} 2^{dj} u} \leq \frac{u^2}{\frac{2}{\eta\bar{C}}u} = \frac{\eta\bar{C}}{2}u.$$

Hence

$$\exp\left(-n\frac{\eta\bar{C}}{2}u\right) \leq \exp\left(-\frac{nu^2}{2\widetilde{C}_2 + \widetilde{C}_{\eta,b} 2^{dj} u}\right).$$

Moreover,

$$2\widetilde{C}_2 + \widetilde{C}_{\eta,b} 2^{dj} u \geq 2C_2 + C_{\eta,b} 2^{dj} u,$$

so

$$\exp\left(-\frac{nu^2}{2C_2 + C_{\eta,b} 2^{dj} u}\right) \leq \exp\left(-\frac{nu^2}{2\widetilde{C}_2 + \widetilde{C}_{\eta,b} 2^{dj} u}\right).$$

Consequently,

$$2\exp\left(-\frac{nu^2}{2C_2 + C_{\eta,b} 2^{dj} u}\right) + \exp\left(-n\frac{\eta\bar{C}}{2}u\right) \leq 3\exp\left(-\frac{nu^2}{2\widetilde{C}_2 + \widetilde{C}_{\eta,b} 2^{dj} u}\right).$$

D.8 Proof of Proposition C.3

Proof D.24 (Proof of Proposition C.3) *The first part $\mathcal{C}^s([0,1]^d) \subseteq \mathcal{B}^s_{\infty,\infty}([0,1]^d)$ has been proved in Manole et al. [2024, Lemma 25]. We focus on the second part: $\mathcal{C}^s \subseteq \mathcal{B}^{s-\epsilon}_{2,2}$ for any $\epsilon > 0$.*

Lemma D.25 (Hölder to Sobolev/Besov with ε-loss) *Let $d, s > 0$. For any $\varepsilon > 0$, there exists a constant $C = C(d, s, \varepsilon)$ such that*

$$\|f\|_{B^{s-\varepsilon}_{2,2}([0,1]^d)} \le C\,\|f\|_{\mathcal{C}^s([0,1]^d)} \qquad \forall f \in \mathcal{C}^s([0,1]^d).$$

Equivalently, $\mathcal{C}^s([0,1]^d) \hookrightarrow B^{s-\varepsilon}_{2,2}([0,1]^d)$.

Proof D.26 (Proof of Lemma D.25) *Fix $\varepsilon > 0$ and set $t = s - \varepsilon$. Let $\{\phi_k\}_{k\in\mathcal{I}_0}$ be the (boundary-corrected) scaling functions at the coarsest resolution and $\{\psi_{j,k}\}_{j\ge j_0,\, k\in\mathcal{I}_j}$ be a compactly supported, boundary-corrected wavelet basis on $[0,1]^d$ with regularity strictly larger than s. Then every $f \in L^2([0,1]^d)$ admits the wavelet expansion*

$$f = \sum_{k\in\mathcal{I}_0} \alpha_k\,\phi_k \;+\; \sum_{j\ge j_0}\sum_{k\in\mathcal{I}_j} \beta_{j,k}\,\psi_{j,k}, \qquad \alpha_k = \langle f, \phi_k\rangle,\ \beta_{j,k} = \langle f, \psi_{j,k}\rangle.$$

We use the standard wavelet characterizations of Besov norms on $[0,1]^d$ (with boundary correction):

$$\|f\|^2_{B^t_{2,2}([0,1]^d)} \lesssim \sum_{k\in\mathcal{I}_0} |\alpha_k|^2 + \sum_{j\ge j_0} 2^{2jt} \sum_{k\in\mathcal{I}_j} |\beta_{j,k}|^2. \tag{10}$$

Let $M := \|f\|_{\mathcal{C}^s([0,1]^d)}$. By the proof of Manole et al. [2024, Lemma 25], we have

$$\sup_{k\in\mathcal{I}_j} |\beta_{j,k}| \le C_1\, M\, 2^{-j(s+d/2)} \qquad \forall j \ge j_0, \tag{11}$$

for some constant C_1 depending only on (d, s) and the chosen wavelet system.

Next, recall that the number of wavelets at level j satisfies $|\mathcal{I}_j| \lesssim 2^{jd}$. Using (11)*,*

$$\sum_{k\in\mathcal{I}_j} |\beta_{j,k}|^2 \le |\mathcal{I}_j|\Big(\sup_{k\in\mathcal{I}_j} |\beta_{j,k}|\Big)^2 \lesssim 2^{jd}\cdot\big(M^2 2^{-2j(s+d/2)}\big) = C_2\, M^2\, 2^{-2js}, \tag{12}$$

for some constant C_2.

Multiplying (12) *by $2^{2jt} = 2^{2j(s-\varepsilon)}$ yields*

$$2^{2jt}\sum_{k\in\mathcal{I}_j} |\beta_{j,k}|^2 \le C_2\, M^2\, 2^{2j(s-\varepsilon)}\cdot 2^{-2js} = C_2\, M^2\, 2^{-2j\varepsilon}.$$

Summing over $j \ge j_0$ and using $\sum_{j\ge j_0} 2^{-2j\varepsilon} < \infty$, we obtain

$$\sum_{j\ge j_0} 2^{2jt} \sum_{k\in\mathcal{I}_j} |\beta_{j,k}|^2 \le C_3\, M^2, \tag{13}$$

for a constant $C_3 = C_3(d, s, \varepsilon)$.

Finally, again by Manole et al. [2024, Lemma 25], $\sup_{k\in\mathcal{I}_0} |\alpha_k| \le M$, thus

$$\sum_{k\in\mathcal{I}_0} |\alpha_k|^2 \le |\mathcal{I}_0|\Big(\sup_{k\in\mathcal{I}_0} |\alpha_k|\Big)^2 \le C_4\, M^2. \tag{14}$$

Combining (10)–(14) *gives*

$$\|f\|^2_{B^t_{2,2}([0,1]^d)} \le C_5\, M^2,$$

hence $\|f\|_{B^{s-\varepsilon}_{2,2}([0,1]^d)} \le \sqrt{C_5}\,\|f\|_{\mathcal{C}^s([0,1]^d)}$, proving the claim.

D.9 Proof of Proposition D.6

In this section, we prove Proposition D.6 in a more general way (see Theorem D.29). By Assumption 2.6 and Proposition C.3, $p, q \in \mathcal{B}^s_{2,2}([0,1]^{d_Y+d_Z})$ and $\gamma^{-1} \le p, q \le \gamma$, thus satisfying Assumptions D.27–D.28.

Assumption D.27 (Smoothness and boundedness) *Let p be a density on $[0,1]^d$ such that $p \in B^s_{2,2}([0,1]^d)$ for some $s > 0$. Assume also $0 \le p \le M$ a.e. for some $M < \infty$.*

Assumption D.28 (Wavelet basis on $[0,1]^d$) *Let*

$$\{\phi^{\mathrm{bc}}_{j_0,k}\}_{k\in\mathcal{K}_0(j_0)} \cup \{\psi^{\mathrm{bc}}_{j,k,\ell}\}_{j\ge j_0,\, k\in\mathcal{K}_0(j),\, \ell\in\{0,1\}^d\setminus\{0\}}$$

be an orthonormal (boundary-corrected) wavelet basis of $L^2([0,1]^d)$. Write the coefficients of p as

$$\alpha_k := \langle p, \phi^{\mathrm{bc}}_{j_0,k}\rangle, \qquad \beta_{j,k,\ell} := \langle p, \psi^{\mathrm{bc}}_{j,k,\ell}\rangle.$$

(Linear projection estimator). Let $X_1, \dots, X_n \sim p$ i.i.d. Fix an integer $J \ge j_0$ and define the empirical coefficients

$$\hat\alpha_k := \frac{1}{n}\sum_{i=1}^n \phi^{\mathrm{bc}}_{j_0,k}(X_i), \qquad \hat\beta_{j,k,\ell} := \frac{1}{n}\sum_{i=1}^n \psi^{\mathrm{bc}}_{j,k,\ell}(X_i), \quad j_0 \le j \le J.$$

Define the projection estimator

$$\tilde p_J(x) := \sum_{k\in\mathcal{K}(j_0)} \hat\alpha_k\, \phi^{\mathrm{bc}}_{j_0,k}(x) \;+\; \sum_{j=j_0}^{J}\sum_{\ell\neq 0}\sum_{k\in\mathcal{K}(j)} \hat\beta_{j,k,\ell}\, \psi^{\mathrm{bc}}_{j,k,\ell}(x).$$

Theorem D.29 (L^2 risk of the truncated projection estimator) *Under Assumptions D.27–D.28,*

$$\mathbb{E}\|\tilde p_J - p\|^2_{L^2([0,1]^d)} \;\lesssim\; 2^{-2Js} \;+\; \frac{2^{Jd}}{n}.$$

Consequently, choosing $2^J \asymp n^{1/(2s+d)}$ yields

$$\mathbb{E}\|\tilde p_J - p\|_2^2 \;\lesssim\; n^{-\frac{2s}{2s+d}}.$$

The hidden constants depend only on (s,d), the wavelet family, M, and $\|p\|_{B^s_{2,2}([0,1]^d)}$.

Proof D.30 (Proof of Theorem D.29) *By orthonormality,*

$$\|\tilde p_J - p\|_2^2 = \|\tilde p_J - p_J\|_2^2 + \|p_J - p\|_2^2,$$

and taking expectations gives

$$\mathbb{E}\|\tilde p_J - p\|_2^2 = \underbrace{\mathbb{E}\|\tilde p_J - p_J\|_2^2}_{\text{(Term b)}} + \underbrace{\|p_J - p\|_2^2}_{\text{(Term a)}}.$$

(Term a). *Since $p \in B^s_{2,2}([0,1]^d)$, the wavelet characterization yields the tail energy bound*

$$\|p - p_J\|_2^2 = \sum_{j>J}\sum_{\ell\neq 0}\sum_{k\in\mathcal{K}(j)} \beta^2_{j,k,\ell} \;\lesssim\; 2^{-2Js}\,\|p\|^2_{B^s_{2,2}}.$$

(Term b). *By orthonormality,*

$$\mathbb{E}\|\tilde p_J - p_J\|_2^2 = \sum_{k\in\mathcal{K}(j_0)} \mathbb{E}(\hat\alpha_k - \alpha_k)^2 \;+\; \sum_{j=j_0}^{J}\sum_{\ell\neq 0}\sum_{k\in\mathcal{K}(j)} \mathbb{E}(\hat\beta_{j,k,\ell} - \beta_{j,k,\ell})^2.$$

Each empirical coefficient is an average of i.i.d. terms, so

$$\begin{aligned}\mathbb{E}(\hat\beta_{j,k,\ell} - \beta_{j,k,\ell})^2 &= \frac{1}{n}\mathrm{Var}\big(\psi^{\mathrm{bc}}_{j,k,\ell}(X_1)\big) \le \frac{1}{n}\mathbb{E}\big[\psi^{\mathrm{bc}}_{j,k,\ell}(X_1)^2\big] \\ &\le \frac{\|p\|_\infty}{n}\int_{[0,1]^d} \big(\psi^{\mathrm{bc}}_{j,k,\ell}(x)\big)^2\,dx = \frac{\|p\|_\infty}{n}, \end{aligned} \tag{15}$$

using $\|p\|_\infty \le M$ and $\|\psi^{\mathrm{bc}}_{j,k,\ell}\|_2 = 1$. The same bound holds for the scaling coefficients. Hence

$$\mathbb{E}\|\tilde{p}_J - p_J\|_2^2 \;\lesssim\; \frac{1}{n}\Big(|\mathcal{K}(j_0)| + \sum_{j=j_0}^{J}\sum_{\ell\neq 0}|\mathcal{K}(j)|\Big).$$

On $[0,1]^d$, $|\mathcal{K}(j)| \lesssim 2^{jd}$ and there are (2^d-1) wavelet types $\ell \neq 0$, so

$$\mathbb{E}\|\tilde{p}_J - p_J\|_2^2 \;\lesssim\; \frac{1}{n}\sum_{j=j_0}^{J} 2^{jd} \;\lesssim\; \frac{2^{Jd}}{n}.$$

Finally, combining (Term a) and (Term b), we get

$$\mathbb{E}\|\tilde{p}_J - p\|_2^2 \;\lesssim\; 2^{-2Js} + \frac{2^{Jd}}{n}.$$

Choosing $2^J \asymp n^{1/(2s+d)}$ balances the two terms and yields $\mathbb{E}\|\tilde{p}_J - p\|_2^2 \lesssim n^{-2s/(2s+d)}$.

D.10 Proof of Proposition D.9

First we introduce the following lemma.

Lemma D.31 ([Manole et al., 2024, Lemma 24 (i)(ii)(v)]) *There exist constants $C_1, C_2 \ge 1$ depending only on d, such that the following properties hold.*

1. *The cardinalities of Φ and Ψ_j satisfy*

$$|\Phi| \le C_1, \qquad |\Psi_j| \le C_2 2^{dj}$$

   for all $j \ge j_0$.

2. *For all $j \ge j_0$ and $\xi \in \Psi_j$, there exists a rectangle $I_\xi \subseteq [0,1]^d$ such that*

$$\mathrm{diam}(I_\xi) \le C_1 2^{-j}, \qquad \mathrm{supp}(\xi) \subseteq I_\xi,$$

   and

$$\Big\| \sum_{\xi\in\Psi_j} \mathbf{1}(\cdot \in I_\xi) \Big\|_{L^\infty} \le C_2.$$

3.

$$\sup_{\zeta\in\Phi} \|\zeta\|_{L^\infty} \le C_1,$$

   and

$$\sup_{j\ge j_0}\sup_{\xi\in\Psi_j} 2^{-\frac{jd}{2}}\|\xi\|_{L^\infty} \le C_2.$$

As a result, let

$$\Theta_J(x) := (\psi_\lambda(x))_{\lambda\in\Lambda_J}$$

denotes the vector of basis functions up to level J, including the scaling functions, then

$$\begin{aligned}\sup_{x\in[0,1]^d}\sum_{\lambda\in\Lambda_J}\psi_\lambda(x)^2 \le& \sum_{j=j_0}^{J} C_2 \sup_{\xi\in\Psi_j}\|\xi\|_{L^\infty}^2 + |\Phi|\sup_{\zeta\in\Phi}\|\zeta\|_{L^\infty}^2\\ \le& C_2^3\sum_{j=j_0}^{J}2^{jd} + C_1^3\\ \lesssim& 2^{Jd}.\end{aligned}$$

Theorem D.32 (L^2 high-probability bound) *Under Assumptions D.27–D.28, suppose in addition that $\|p\|_\infty \le M$ and that the boundary-corrected wavelet projection kernel satisfies*

$$\sup_{x\in[0,1]^d} \sum_{\lambda\in\Lambda_J} \psi_\lambda(x)^2 \le C_\Psi 2^{Jd},$$

where Λ_J denotes all scaling and wavelet indices up to level J. Then, for every $t \ge 1$, with probability at least $1 - e^{-t}$,

$$\|\tilde{p}_J - p\|_2^2 \le C\left\{2^{-2Js} + \frac{2^{Jd}t}{n} + \frac{2^{Jd}t^2}{n^2}\right\}.$$

In particular, if $1 \le t \le n$, then with probability at least $1 - e^{-t}$,

$$\|\tilde{p}_J - p\|_2^2 \le C\left\{2^{-2Js} + \frac{2^{Jd}t}{n}\right\}.$$

Taking $2^J \asymp n^{1/(2s+d)}$ and $1 \le t \le n$ gives

$$\|\tilde{p}_J - p\|_2^2 \lesssim n^{-\frac{2s}{2s+d}} t$$

up to constants, with probability at least $1 - e^{-t}$.

Proof D.33 *Let V_J be the linear span of the scaling functions and wavelets up to level J, and let Λ_J be the corresponding index set. Let p_J be the L^2 projection of p onto V_J. By orthonormality,*

$$\|\tilde{p}_J - p\|_2^2 = \|\tilde{p}_J - p_J\|_2^2 + \|p_J - p\|_2^2.$$

By the analysis of (term a) in the proof of Theorem D.29, we have

$$\|p_J - p\|_2^2 \lesssim 2^{-2Js}\|p\|_{B^s_{2,2}}^2.$$

It remains to control the stochastic term. Recall that

$$\Theta_J(x) = (\psi_\lambda(x))_{\lambda\in\Lambda_J}$$

denotes the vector of basis functions up to level J, including the scaling functions.

Using the same notation as in the previous proof, the vector of empirical coefficient errors is

$$\begin{aligned}\widehat{\theta}_J - \theta_J =& \{\hat{\alpha}_k - \alpha_k : k \in \mathcal{K}(j_0)\} \cup \{\hat{\beta}_{j,k,\ell} - \beta_{j,k,\ell}\ j = j_0, ..., J, k \in \mathcal{K}(j), \ell \ne 0\} \\ =& \frac{1}{n}\sum_{i=1}^n \{\Theta_J(X_i) - \mathbb{E}\Theta_J(X_i)\}.\end{aligned}$$

By orthonormality,

$$\|\tilde{p}_J - p_J\|_2 = \|\widehat{\theta}_J - \theta_J\|_{\ell_2}.$$

Set $\xi_i := \Theta_J(X_i) - \mathbb{E}\Theta_J(X_i)$. The assumed projection-kernel bound implies

$$\sup_{x\in[0,1]^d} \|\Theta_J(x)\|_{\ell_2}^2 = \sup_{x\in[0,1]^d} \sum_{\lambda\in\Lambda_J} \psi_\lambda(x)^2 \le C_\Psi 2^{Jd},$$

and hence

$$\|\xi_i\|_{\ell_2} \le 2C_\Psi^{1/2} 2^{Jd/2}.$$

Moreover,

$$\mathbb{E}\|\xi_i\|_{\ell_2}^2 \le \mathbb{E}\|\Theta_J(X_i)\|_{\ell_2}^2 = \int \sum_{\lambda\in\Lambda_J} \psi_\lambda(x)^2 p(x)\,dx \le M|\Lambda_J| \lesssim M2^{Jd}.$$

Applying a Hilbert-space Bernstein inequality to $n^{-1}\sum_{i=1}^n \xi_i$ gives that, with probability at least $1 - e^{-t}$,

$$\|\widehat{\theta}_J - \theta_J\|_{\ell_2} \le C\left(\sqrt{\frac{2^{Jd}t}{n}} + \frac{2^{Jd/2}t}{n}\right).$$

Therefore, using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\|\tilde{p}_J - p_J\|_2^2 \le C\left(\frac{2^{Jd}t}{n} + \frac{2^{Jd}t^2}{n^2}\right)$$

with probability at least $1 - e^{-t}$. Combining this stochastic bound with the deterministic bias bound proves the theorem.

D.11 Proof of Lemmas

D.11.1 Proof of Lemma D.1

Proof D.34 (Proof of Lemma D.1) *Plug $p = 2$ in Theorem 4 in [Niles-Weed and Berthet, 2022] and we finish the proof.*

D.11.2 Proof of Lemma D.2

Proof D.35 (Proof of Lemma D.2) *Let $\{\psi_\lambda\}_{\lambda\in\Lambda}$ be an orthonormal wavelet basis of $L^2([0,1]^d)$, and let $\Lambda = \bigsqcup_{j\geq j_0}\Lambda_j$ denote the decomposition of indices by scale. Fix $f \in L^2([0,1]^d)$, write its wavelet expansion*

$$f \;=\; f_{<j_0} \;+\; \sum_{j\geq j_0} f_j, \qquad f_j \;=\; \sum_{\lambda\in\Lambda_j} \alpha(\lambda)\,\psi_\lambda,$$

where $f_{<j_0}$ is the coarse (scaling) component.

For each $j \geq j_0$, since $\{\psi_\lambda\}_{\lambda\in\Lambda_j}$ is an orthonormal basis at level j, Parseval's identity then gives

$$\|f_j\|^2_{L^2([0,1]^d)} \;=\; \sum_{\lambda\in\Lambda_j} |\alpha(\lambda)|^2, \qquad \textit{hence} \qquad \|f_j\|_{L^2([0,1]^d)} \;=\; \Big(\sum_{\lambda\in\Lambda_j} |\alpha(\lambda)|^2\Big)^{1/2}. \tag{16}$$

Therefore the left-hand side can be rewritten as

$$\left(\sum_{j\geq j_0} 2^{-j}\Big(\sum_{\lambda\in\Lambda_j} |\alpha(\lambda)|^2\Big)^{1/2}\right)^2 \;=\; \left(\sum_{j\geq j_0} 2^{-j}\,\|f_j\|_{L^2([0,1]^d)}\right)^2.$$

Apply the Cauchy–Schwarz inequality to the sequences $a_j = 2^{-j}$ and $b_j = \|f_j\|_{L^2([0,1]^d)}$:

$$\sum_{j\geq j_0} 2^{-j}\,\|f_j\|_{L^2([0,1]^d)} \;\leq\; \left(\sum_{j\geq j_0} 2^{-2j}\right)^{1/2}\left(\sum_{j\geq j_0} \|f_j\|^2_{L^2([0,1]^d)}\right)^{1/2}. \tag{17}$$

Squaring (17) *yields*

$$\left(\sum_{j\geq j_0} 2^{-j}\,\|f_j\|_{L^2([0,1]^d)}\right)^2 \;\leq\; \left(\sum_{j\geq j_0} 2^{-2j}\right)\left(\sum_{j\geq j_0} \|f_j\|^2_{L^2([0,1]^d)}\right). \tag{18}$$

Next, by the orthogonal decomposition

$$\begin{aligned}\|f\|_{L^2([0,1]^d)} &= (\|f_{<j_0}\|^2_{L^2([0,1]^d)} + \sum_{j\geq j_0} \|f_j\|^2_{L^2([0,1]^d)})^{\frac{1}{2}} \\ &\geq\; \frac{1}{2}\left(\|f_{<j_0}\|_{L^2([0,1]^d)} + \left(\sum_{j\geq j_0} \|f_j\|^2_{L^2([0,1]^d)}\right)^{\frac{1}{2}}\right).\end{aligned}$$

Substituting this bound into (18) *gives*

$$\left(\sum_{j\geq j_0} 2^{-2j}\right)^{\frac{1}{2}} \|f_{<j_0}\|_{L^2([0,1]^d)} + \sum_{j\geq j_0} 2^{-j}\,\|f_j\|_{L^2([0,1]^d)} \;\leq\; 2\left(\sum_{j\geq j_0} 2^{-2j}\right)^{\frac{1}{2}} \|f\|_{L^2([0,1]^d)}.$$

Finally, combining this with

$$\|f\|_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} = \|f_{<j_0}\|_{L^2([0,1]^d)} + \sum_{j\geq j_0} 2^{-j}\,\|f_j\|_{L^2([0,1]^d)}$$

concludes the proof.

D.11.3 Proof of Lemma D.13

Proof D.36 (Proof of Lemma D.13) *For each $z \in \mathcal{Z}$, let π_Z^* be an optimal coupling between P_Y^z and Q_Y^z for the cost $\|y - y'\|^2$. Define a coupling π^* on $(\mathcal{Y} \times \mathcal{Z})^2$ via*

$$\pi^*(dy, dz, dy', dz') := P_Z(dz)\, \pi_Z^*(dy, dy')\, \delta_Z(dz').$$

By construction, the first marginal of π^* is P and the second marginal is Q, hence $\pi^* \in \Pi(P, Q)$. Under the cost $c\big((y, z), (y', z')\big) = \|y - y'\|^2$, the transport cost of π is

$$\int \|y - y'\|^2 \, d\pi^* = \int \left(\int \|y - y'\|^2 \, d\pi_Z^*(y, y') \right) P_Z(dz) = \mathbb{E}_{Z \sim P_Z} \left[\mathsf{W}_2^2\big(P(Y \mid Z),\, Q(Y \mid Z)\big) \right].$$

Since $\mathsf{W}_2^2(P, Q)$ is the infimum of $\int \|y - y'\|^2 \, d\pi$ over all $\pi \in \Pi(P, Q)$, we conclude that

$$\mathsf{W}_2^2(P, Q) \le \int \|y - y'\|^2 \, d\pi^* = \mathbb{E}_{Z \sim P_Z} \left[\mathsf{W}_2^2\big(P(Y \mid Z),\, Q(Y \mid Z)\big) \right].$$

D.11.4 Proof of Lemma D.16

Proof D.37 (Proof of Lemma D.16) *Since $\phi(y, z) = \|y\|_2^2 - 2\varphi(y, z)$, then $\phi \in \mathcal{C}^{s+1}([0,1]^{d_Y + d_Z})$. As for ψ, note that $\psi(y, z) = \|y\|_2^2 - 2\varphi^*(y, z)$, and*

$$\nabla_y^2 \varphi^*(y, z) = \nabla_y^2 \varphi(\nabla_y \varphi^*(y, z), z)^{-1} \qquad (\star)$$

where the convex conjugate is defined with respect to y, i.e., $\varphi^*(y, z) := \sup_{y' \in [0,1]^{d_Y}} \{y^\top y' - \varphi(y', z)\}$.

By Lemma D.14, there is a constant $\lambda > 0$, such that

$$\frac{1}{\lambda} I_{d_Y} \preceq \nabla_y^2 \varphi(y, z) \preceq \lambda I_{d_Y} \quad \forall y, z \in [0,1]^{d_Y + d_Z}.$$

Therefore, by $(\star)$ and $\varphi \in \mathcal{C}^{s+1}([0,1]^{d_Y + d_Z})$, we get $\varphi^* \in \mathcal{C}^{s+1}([0,1]^{d_Y + d_Z})$, so does ψ.

E Algorithms

E.1 Wavelet-based Density Estimation

We provide a detailed Algorithm 1 for the wavelet-based density estimation described in Definition 2.4.

E.2 Optimal Transport Problem with Estimated Densities

Given the estimated joint densities $\widehat{P}$ and $\widehat{Q}$ of $(Y_i(0), Z_i)$ and $(Y_i(1), Z_i)$, together with an estimate of the marginal density of Z, we generate synthetic data through a hierarchical sampling scheme designed to mimic the joint structure of the population.

1. Draw N_1 samples of Z from its estimated marginal distribution.
2. Conditional on each sampled value of Z, draw N_2 outcomes from the conditional densities $\widehat{P}(Y(0) \mid Z)$ and $\widehat{Q}(Y(1) \mid Z)$.

We emphasize that these generated samples are not the original observations, but are Monte Carlo draws from the estimated density. To estimate the optimal objective value, we solve an optimal transport problem separately at each sampled Z, and average the resulting objective values over the sampled covariates to obtain the final estimate. Details are summarized in Algorithm 2.

There are a variety of approaches for sampling from an estimated density [Robert et al., 1999], including rejection sampling, importance sampling, Markov chain Monte Carlo (MCMC) methods such as Metropolis–Hastings and Gibbs sampling. In Algorithm 2, we adopt a discretization-based sampler for its favorable practical stability. Specifically, we first discretize the support of the estimated density into a fine grid. The estimated density is then evaluated on this grid and normalized to form

Input: Samples $(Y_i, Z_i) \in [0,1]^{d_Y+d_Z}$ for $i = 1, \dots, n$; wavelet family with smoothness s; coarse level $J_0 \geq \lfloor \log_2(s/0.18 + 1) \rfloor + 1$.

Output: Estimated joint density $\widehat{f}_{Y,Z}$ on $[0,1]^{d_Y+d_Z}$.

Notation. Let $d := d_Y + d_Z$ and $X_i := (Y_i, Z_i) \in [0,1]^d$. Let $\{\phi^{\mathrm{bc}}_{J_0,k}\}_{k \in \mathcal{K}_{J_0}}$ be the *boundary-corrected* scaling functions at level J_0, and for each $j \geq J_0$ let $\{\psi^{\mathrm{bc}}_{j,k,\ell}\}_{k \in \mathcal{K}_j,\, \ell \in \mathcal{L}}$ be the boundary-corrected wavelets (where ℓ indexes the $2^d - 1$ multivariate wavelet types in d dimensions). Basis functions are supported on $[0,1]^d$ and form an orthonormal basis of $L^2([0,1]^d)$.

Step 1: choose resolution levels.

Pick (J_0, J) such that J_0 fixed small, $J \geq J_0$, and $J \asymp \frac{\log_2(n)}{2s + d_Y + d_Z}$.

Step 2: compute empirical wavelet coefficients.

foreach $k \in \mathcal{K}_{J_0}$ **do**
  $\widehat{\alpha}_{J_0,k} \leftarrow \frac{1}{n} \sum_{i=1}^n \phi^{\mathrm{bc}}_{J_0,k}(X_i)$;
end
for $j = J_0$ *to* J **do**
  foreach $\ell \in \mathcal{L}$ **do**
    foreach $k \in \mathcal{K}_j$ **do**
      $\widehat{\beta}_{j,k,\ell} \leftarrow \frac{1}{n} \sum_{i=1}^n \psi^{\mathrm{bc}}_{j,k,\ell}(X_i)$;
    end
  end
end

Step 3 (optional): threshold / shrink detail coefficients.

for $j = J_0$ *to* J **do**
  foreach $\ell \in \mathcal{L}$ **do**
    foreach $k \in \mathcal{K}_j$ **do**
      $\widetilde{\beta}_{j,k,\ell} \leftarrow \mathcal{T}(\widehat{\beta}_{j,k,\ell}; \lambda_j)$;
      `// Examples: hard` $\mathcal{T}(b;\lambda) = b\mathbf{1}\{|b| \geq \lambda\}$`, soft` $\mathcal{T}(b;\lambda) = \mathrm{sign}(b)\,(|b| - \lambda)_+$.
    end
  end
end
if *no thresholding* **then**
  $\widetilde{\beta}_{j,k,\ell} \leftarrow \widehat{\beta}_{j,k,\ell}$ for all (j,k,ℓ);
end

Step 4: assemble the estimator.

Define, for any $x \in [0,1]^d$,

$$\widehat{f}_{Y,Z}(x) \;:=\; \sum_{k \in \mathcal{K}_{J_0}} \widehat{\alpha}_{J_0,k}\, \phi^{\mathrm{bc}}_{J_0,k}(x) \;+\; \sum_{j=J_0}^{J} \sum_{\ell \in \mathcal{L}} \sum_{k \in \mathcal{K}_j} \widetilde{\beta}_{j,k,\ell}\, \psi^{\mathrm{bc}}_{j,k,\ell}(x).$$

Step 5: enforce density constraints.

if *nonnegativity flag* **then**
  `// Simple projection (may introduce bias); other options: log-density wavelets, constrained optimization.`
  $\widehat{f}_{Y,Z}(x) \leftarrow \max\{\widehat{f}_{Y,Z}(x), 0\}$;
end
if *normalization flag* **then**
  `// Renormalize to integrate to one; numerical quadrature or exact if basis integrates analytically.`
  $c \leftarrow \int_{[0,1]^d} \widehat{f}_{Y,Z}(x)\, dx$;
  $\widehat{f}_{Y,Z}(x) \leftarrow \widehat{f}_{Y,Z}(x)/c$;
end
return $\widehat{f}_{Y,Z}$;

Algorithm 1: Boundary-corrected wavelet density estimator for the joint density of (Y, Z)

a discrete probability distribution. Samples are subsequently drawn from this grid according to the discrete probabilities. We remark that this discretization step differs from that used by [Lin et al., 2026], where the discretization is used as part of the estimation procedure and applied to the observed dataset. In contrast, here the discretization is used as a post-estimation device for sampling from the already constructed continuous density estimate. This numerical approximation is distinct from the continuous-density estimator analyzed in our theory; its discretization and Monte Carlo errors must be controlled separately. The implementation settings below also use periodic boundary handling and a separately chosen marginal resolution, whereas the theoretical estimator uses boundary-corrected wavelets and the marginals induced by the joint density estimators.

E.3 Computational Complexity

For our Algorithm, let N_Z denote the number of sampled Z values and let N_Y denote the number of sampled Y values per sampled Z. The overall computational complexity is

$$O\left((n+m)d2^{d_Z+d_Y}(n\vee m)^{\frac{d_Z+d_Y}{2s+d_Z+d_Y}} + N_Z\left(d_Y N_Y^2 + N_Y^3\log N_Y\right)\right).$$

The first term comes from estimating the marginal densities using wavelets. The second term comes from sampling from the estimated conditional sampler and then solving the corresponding empirical optimal-transport problems. For the second step, to achieve an additional error of order $O(n^{-1/2})$, $N_Z = n$ and $N_Y = n^{d_Y}$ are needed, which dominates the first term.

Author check: The preceding prescription $N_Y = n^{d_Y}$ differs from Algorithm 2, which uses $N_Y = \lfloor (n\vee m)^{(d_Y/4)\vee 1}\log(n\vee m)\rfloor$. Please reconcile these choices and justify the numerical-error rate. Preserving the CLT for the continuous-density estimator requires numerical error $o_p(n^{-1/2})$, or a separate analysis of its contribution.

This suggests that the main computational burden is not the wavelet estimation step: even with an oracle sampler for the conditional distributions, one would still need to solve the empirical optimal-transport problem for each sampled Z, incurring the same computational cost.

We remark that various acceleration techniques can be applied, such as parallel computing.

F Additional Numerical Experiment

F.1 Simulation details of estimation

F.1.1 Wavelet Algorithm Parameters

- `wavelet = "db4"`: Daubechies-4 wavelet basis, chosen for compact support and good smoothness properties.
- `mode = "periodization"`: boundary handling rule that treats the support as periodic; it differs from the boundary-corrected wavelet system used in the theory.
- `J_joint = 4`: wavelet resolution level for estimating the joint density $\widehat{p}_{Y,Z}(y,z)$. Larger values lead to finer resolution (lower bias, higher variance) and constitute the primary smoothing parameter.
- `J_z = 6`: wavelet resolution level for estimating the marginal density $\widehat{p}_Z(z)$. A higher level is used since Z is lower dimensional and easier to estimate accurately.
- `threshold = None`: no coefficient thresholding is applied, preserving the plug-in estimator form and simplifying theoretical analysis.
- `nonnegativity = True`: enforces nonnegativity of the reconstructed density, preventing oscillations from producing negative values.
- `renormalize = True`: rescales the estimated density to integrate to one, ensuring it is a valid probability density.
- `(i)` $N_z = 200$, $N_y = 300$ `(location model)`, `(ii)` $N_z = 50$, $N_y = 300$ `(quadratic model)`, `(iii)` $N_z = 100$, $N_y = 600$ `(scale model)`: number of grid points used to discretize the supports of Z and Y when numerically evaluating the conditional Wasserstein functional.

Input: i.i.d. sample $((Y_i(0), Z_i(0)), 1 \le i \le n)$ drawn from P, i.i.d. sample $((Y_j(1), Z_j(1)), 1 \le j \le m)$ drawn from Q

Output: $\widehat{\mathsf{CW}}(\widehat{P}_n^\dagger, \widehat{Q}_n^\dagger)$

Step 1:

Compute the wavelet estimator $\widehat{P}$ based on $((Y_i(0), Z_i(0)), 1 \le i \le n)$ using Algorithm 1.

Compute the wavelet estimator $\widehat{Q}$ based on $((Y_j(1), Z_j(1)), 1 \le j \le m)$ using Algorithm 1.

Step 2:

Compute the wavelet estimator of $\widehat{R} = (\widehat{P}_Z + \widehat{Q}_Z)/2$.

Denote $\widehat{P}_n^\dagger = \widehat{P}_n^\dagger(\mathrm{d}y, \mathrm{d}z) = \widehat{P}_Y^z(\mathrm{d}y)\widehat{R}(\mathrm{d}z)$.

Denote $\widehat{Q}_n^\dagger = \widehat{Q}_n^\dagger(\mathrm{d}y, \mathrm{d}z) = \widehat{Q}_Y^z(\mathrm{d}y)\widehat{R}(\mathrm{d}z)$.

Step 3:

Sample $(\widetilde{Z}_\tau, 1 \le \tau \le N_Z)$, where $N_Z := \lfloor (n \vee m)\log(n \vee m) \rfloor$, from $\widehat{R}$.

Step 4:

for $\tau = 1$ *to* N_Z **do**

Let $z = \tilde{Z}_\tau$, $N_Y = \lfloor (n \vee m)^{(d_Y/4)\vee 1} \log(n \vee m) \rfloor$;

Sample $(Y_k^\tau(0), 1 \le k \le N_Y)$ from $\widehat{P}_Y^z$;

Sample $(Y_l^\tau(1), 1 \le l \le N_Y)$ from $\widehat{Q}_Y^z$;

Compute

$$\widehat{\mathsf{W}}_2^2(\widehat{P}_{\widetilde{Z}_\tau}, \widehat{Q}_{\widetilde{Z}_\tau}) := \mathsf{W}_2^2\left(\frac{1}{N_Y}\sum_{k=1}^{N_Y}\delta_{Y_k^\tau(0)}, \frac{1}{N_Y}\sum_{l=1}^{N_Y}\delta_{Y_l^\tau(1)}\right)$$

using the Python Optimal Transport (`POT`[3]) library [Flamary et al., 2021];

end

Step 5:

Define

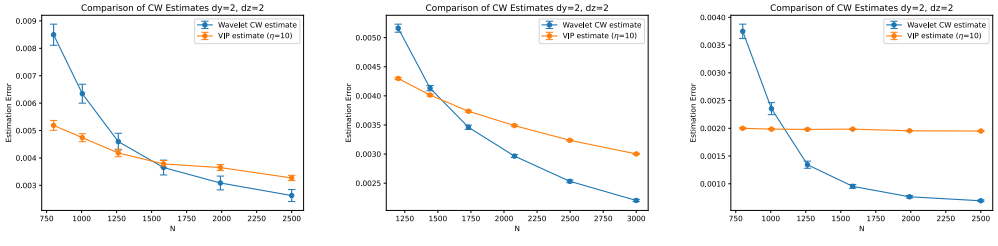
$$\widehat{\mathsf{CW}}(\widehat{P}_n^\dagger, \widehat{Q}_n^\dagger) := \frac{1}{N_Z}\sum_{\tau=1}^{N_Z}\widehat{\mathsf{W}}_2^2(\widehat{P}_{\widetilde{Z}_\tau}, \widehat{Q}_{\widetilde{Z}_\tau}).$$

return $\widehat{\mathsf{CW}}(\widehat{P}_n^\dagger, \widehat{Q}_n^\dagger)$

Algorithm 2: Smooth conditional Wasserstein distance

F.1.2 Additional Estimation Result

In Figure 5, we show the plots of convergence rate corresponding to Figure 2.



(a) Location model (b) Quadratic model (c) Scale model

Figure 5: Plots of estimation error comparison of our method (*Wavelet CW estimate*) and [Lin et al., 2025] (*VIP estimate* with $\eta = 10$) with $d_Y = 2$, $d_Z = 2$. The estimation error is defined by $\mathbb{E}[|V_{\mathrm{c}} - \widehat{V}_{\mathrm{c}}|]$. The mean error curve and the corresponding 90% standard-error confidence bands are computed by aggregating results over 300 Monte Carlo repetitions.

F.2 Simulation details of inference

F.2.1 Data generation mechanism.

In all scenarios,

$$Y(w) \mid Z = z \sim \mathcal{N}\big(\mu_w(z), \Sigma_w\big), \quad w \in \{0, 1\}.$$

Scenario 1: $d_Y = 2,\ d_Z = 1$

$$\mu_0(z) = \begin{pmatrix} 0.35 \\ 0.55 \end{pmatrix} + \begin{pmatrix} 0.10 \\ -0.05 \end{pmatrix} z,$$

$$\mu_1(z) = \mu_0(z) + \begin{pmatrix} 0.12 \\ -0.08 \end{pmatrix} + \begin{pmatrix} 0.10 \\ 0.02 \end{pmatrix} z.$$

$$\Sigma_0 = \begin{pmatrix} 0.05^2 & 0 \\ 0 & 0.03^2 \end{pmatrix}, \qquad \Sigma_1 = \begin{pmatrix} 0.07^2 & 0.01 \cdot 0.07 \cdot 0.04 \\ 0.01 \cdot 0.07 \cdot 0.04 & 0.04^2 \end{pmatrix}.$$

Scenario 2: $d_Y = 2,\ d_Z = 2$ **(default)**

$$\mu_0(z) = \begin{pmatrix} 0.35 \\ 0.55 \end{pmatrix} + \begin{pmatrix} 0.20 & -0.10 \\ 0.05 & 0.15 \end{pmatrix} z,$$

$$\mu_1(z) = \mu_0(z) + \begin{pmatrix} 0.12 \\ -0.08 \end{pmatrix} + \begin{pmatrix} 0.10 & 0.04 \\ 0.02 & 0.06 \end{pmatrix} z.$$

$$\Sigma_0 = \begin{pmatrix} 0.05^2 & 0 \\ 0 & 0.03^2 \end{pmatrix}, \qquad \Sigma_1 = \begin{pmatrix} 0.07^2 & 0.01 \cdot 0.07 \cdot 0.04 \\ 0.01 \cdot 0.07 \cdot 0.04 & 0.04^2 \end{pmatrix}.$$

Scenario 3: $d_Y = 3,\ d_Z = 2$

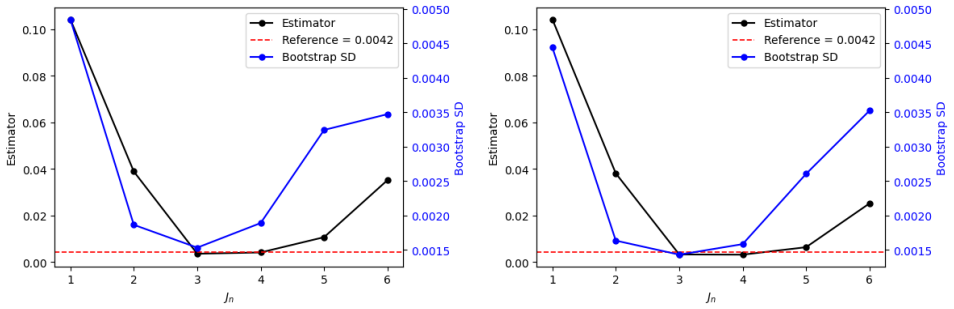
$$\mu_0(z) = \begin{pmatrix} 0.35 \\ 0.55 \\ 0.45 \end{pmatrix} + \begin{pmatrix} 0.20 & -0.10 \\ 0.05 & 0.15 \\ -0.12 & 0.08 \end{pmatrix} z,$$

$$\mu_1(z) = \mu_0(z) + \begin{pmatrix} 0.12 \\ -0.08 \\ 0.05 \end{pmatrix} + \begin{pmatrix} 0.10 & 0.04 \\ 0.02 & 0.06 \\ -0.03 & 0.01 \end{pmatrix} z.$$

$$\Sigma_0 = \begin{pmatrix} 0.05^2 & 0 & 0 \\ 0 & 0.03^2 & 0 \\ 0 & 0 & 0.04^2 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 0.07^2 & 0.01 \cdot 0.07 \cdot 0.04 & 0 \\ 0.01 \cdot 0.07 \cdot 0.04 & 0.04^2 & 0 \\ 0 & 0 & 0.05^2 \end{pmatrix}.$$

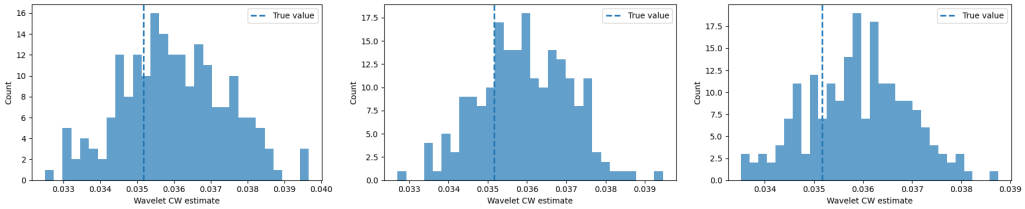
F.2.2 Wavelet Algorithm Parameters

- `wavelet = "db4"`: Daubechies-4 wavelet basis, chosen for compact support and good smoothness properties.
- `mode = "periodization"`: boundary handling rule that treats the support as periodic; it differs from the boundary-corrected wavelet system used in the theory.
- `J_joint = 4`: wavelet resolution level for estimating the joint density $\widehat{p}_{Y,Z}(y, z)$. Larger values lead to finer resolution (lower bias, higher variance) and constitute the primary smoothing parameter.
- `J_z = 6`: wavelet resolution level for estimating the marginal density $\widehat{p}_Z(z)$. A higher level is used since Z is lower dimensional and easier to estimate accurately.
- `threshold = None`: no coefficient thresholding is applied, preserving the plug-in estimator form and simplifying theoretical analysis.
- `nonnegativity = True`: enforces nonnegativity of the reconstructed density, preventing oscillations from producing negative values.
- `renormalize = True`: rescales the estimated density to integrate to one, ensuring it is a valid probability density.
- $N_z = 120,\ N_y = 120$: number of grid points used to discretize the supports of Z and Y when numerically evaluating the conditional Wasserstein functional.



(a) $n = 3000$ (b) $n = 5000$

Figure 6: Tuning of the wavelet parameter J_n for (a) $n = 3000$ and (b) $n = 5000$. The black solid line shows the estimator, the red dashed line marks the true value, and the blue solid line shows the bootstrap standard deviation.



(a) $n = 1000$ (b) $n = 2000$ (c) $n = 3000$

Figure 7: Histogram of the proposed estimator with the true value marked by a vertical dashed line. Here the default setting $d_Y = 2$, $d_Z = 1$ is adopted. The results are aggregated over 200 Monte Carlo repetitions.

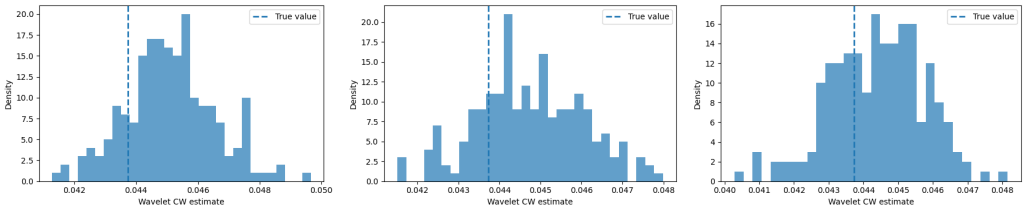
F.2.3 Tuning of Smoothness parameter

The smoothness parameter affects the wavelet resolution level J_n, corresponding to J_joint in our implementation. Here we describe how to tune this parameter. We consider the grid $\{1, 2, 3, 4, 5, 6\}$. For each choice, we plot the estimator and the bootstrap standard deviation (subsampled 200 times). The true reference value is shown only for evaluation purposes, since in practice the true value is unknown and the bias cannot be directly observed.

We see that when J_n is small, both the bias and the bootstrap standard deviation are relatively large, due to boundary effects: at low resolution, the wavelet approximation relies heavily on boundary extension, so edge artifacts can affect a large fraction of the coefficients. As J_n becomes larger, the bias and the bootstrap standard deviation decreases. However, when J_n is overly large, the bootstrap standard deviation starts to increase again because the estimator includes more wavelet coefficients and hence greater variance. Moreover, the additional variability in $\widehat{P}$ and $\widehat{Q}$ does not cancel after applying the nonlinear Wasserstein-2 functional; instead, it is converted into additional transportation cost, thereby inducing upward bias. Based on the plot, we recommend choosing J_n by minimizing either the estimator value or its bootstrap standard deviation.

F.3 Inference

In Figure 7, we illustrate the asymptotic distribution of our estimator $\mathsf{CW}_2^2(\widehat{P}_n^\dagger, \widehat{Q}_n^\dagger)$ for $d_Y = 2$, $d_Z = 1$. In Figure 8, we present the corresponding asymptotic distribution for $d_Y = 3$, $d_Z = 2$. The results are consistent with those in Section 4.2: larger sample sizes decrease the bias and improve the normality.



(a) $n = 4000$ (b) $n = 5000$ (c) $n = 6000$

Figure 8: Histogram of the proposed estimator with the true value marked by a vertical dashed line. Here the default setting $d_Y = 3$, $d_Z = 2$ is adopted. The results are aggregated over 200 Monte Carlo repetitions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: For each theoretical result, the paper provides the full set of assumptions and a complete and correct proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For each experiment, the paper provides sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited and the license and terms of use explicitly are mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets are introduced in the paper well documented and is the documentation provided alongside the assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.