GENERALIZATION BOUNDS AND REPRESENTATION LEARNING FOR DISTRIBUTION-VALUED POTENTIAL OUTCOMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneous treatment effect estimation is central to understanding how causal effects vary across units with different observed characteristics. In many applications, however, each unit is observed through repeated post-treatment measurements that are more naturally represented by a probability distribution. Treatment can then change dispersion, tails, or shape even when the conditional mean is unchanged. In this paper, we study treatment-specific conditional Wasserstein barycenters for distribution-valued potential outcomes and define their conditional contrast through the difference of barycenter quantile functions. We develop a shared-representation estimator for learning these treatment-specific barycenters from observational data. We characterize the approximation induced by the learned representation, establish counterfactual generalization and finite-sample guarantees for target-population barycenter and contrast risks, and clarify how finite repeated measurements affect identification and estimation of the latent distribution-valued targets. Experimental results show that (i) shared fitting achieves lower mean contrast risk than separate fitting in all reported settings and lower treatment-specific barycenter risk in each synthetic family, and (ii) controlled experiments support the theoretically predicted effects of repeated measurements, quantile discretization, and representation learning.

1 INTRODUCTION

Heterogeneous treatment effect estimation is central to understanding how causal effects vary across units with different observed characteristics. Most existing methods are developed for scalar outcomes and summarize heterogeneity through conditional mean potential outcomes, leading to the conditional average treatment effect. Flexible estimators based on forests, meta-learners, and representation learning are now well developed for this setting (Athey & Imbens, 2016; Wager & Athey, 2018; Künzel et al., 2019; Shalit et al., 2017). Many applications, however, record repeated post-treatment measurements for each unit that are more naturally summarized by an entire response distribution. Examples include activity measurements over a monitoring window and spending across repeated transactions (Lin et al., 2023; Leung et al., 2025). Treatment can change dispersion, tails, or distributional shape even when the conditional mean is unchanged. Mean effects therefore provide an incomplete description of treatment heterogeneity in such settings.

Fréchet and Wasserstein regression provide conditional prediction methods for distribution-valued responses, including flexible regression and neural approaches (Petersen & Müller, 2019; Fan & Müller, 2025; Iao et al., 2025; Kim et al., 2025; Zhou et al., 2026; Zhang et al., 2025). Some causal inference work has studied population-level effects for distribution-valued potential outcomes (Lin et al., 2023), matching-based estimation of treatment-specific conditional Wasserstein barycenters (Katta et al., 2024), and population effects under continuous treatments (Leung et al., 2025). These developments provide estimands and estimation tools for causal analysis with distribution-valued outcomes. What remains less focused is flexible representation learning for treatment-specific conditional Wasserstein barycenters under observational treatment assignment, together with generalization theory for distribution-valued potential outcomes observed through repeated measurements.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This research focus raises two challenges beyond standard scalar counterfactual representation learning. First, each unit reveals only the potential outcome measure under its assigned treatment, while both treatment-specific conditional barycenters are needed over a common target population. A shared representation can borrow information across treatment groups and reduce differences between their covariate representations, but compression or aggressive balancing can discard covariate information needed for outcome prediction (Shalit et al., 2017; Johansson et al., 2019; 2022). Second, each unit-level outcome distribution is typically observed through finitely many repeated measurements, while practical estimators use a finite-dimensional quantile representation. Representation compression, repeated measurements, and quantile discretization therefore contribute distinct errors to the same counterfactual prediction problem. These challenges motivate a representation-learning framework that estimates treatment-specific conditional Wasserstein barycenters and their contrast over the target population with explicit generalization guarantees.

In this paper, we develop a shared-representation estimator with a common encoder and treatment-specific monotone quantile heads. The estimator combines factual Wasserstein prediction with optional treatment-group discrepancy and covariate reconstruction penalties. We characterize the approximation error induced by the learned representation and show how reconstruction controls this error under regularity conditions. We then derive a counterfactual generalization bound relating factual prediction to target-population barycenter and contrast risks, and show that factual fit and representation balance alone do not uniformly control target risk under weak overlap. For the fitted estimator, we establish a finite-sample bound that separates quantile discretization, statistical estimation, repeated-measurement error, numerical transport approximation, and optimization error. We further characterize the role of repeated measurements by showing that fixed bounded measurement counts do not identify unrestricted latent conditional barycenters in general and by giving sufficient conditions for consistency as within-unit measurement counts increase.

Our **contributions** are threefold:

1. We formulate counterfactual representation learning for distribution-valued potential outcomes and develop a shared-representation approach for estimating treatment-specific conditional Wasserstein barycenters and their contrast from observational data.
2. We establish generalization and finite-sample theory that links factual prediction and learned representations to target-population barycenter and contrast risks, while clarifying how repeated measurement affects identification and estimation of the latent distribution-valued targets.
3. Shared fitting has lower mean risk than Separate in every reported comparison. Controlled experiments also support the predicted behavior of representation discrepancy, reconstruction, repeated measurements, and quantile discretization.

2 Related Work

Distribution-valued regression and causal outcomes. Fréchet regression and conditional Wasserstein barycenters provide regression targets for random distributions, with nonlinear and neural extensions for flexible prediction (Petersen & Müller, 2019; Fan & Müller, 2025; Iao et al., 2025; Kim et al., 2025; Zhou et al., 2026). For causal inference with distribution-valued potential outcomes, Lin et al. (2023) develop population causal effects and doubly robust estimation, Katta et al. (2024) estimate treatment-specific conditional barycenters through matching, and Leung et al. (2025) study continuous-treatment population effects using quantile-function regression. Related work extends causal inference to outcomes in general geodesic metric spaces (Kurisu et al., 2024). Our focus is flexible prediction of both conditional barycenter functions from binary observational data with risk evaluated over a common target covariate population.

Distributional causal targets for scalar outcomes. A different literature studies distributional features of scalar or vector potential outcomes. Kallus & Oprescu (2023) learn conditional quantile, super-quantile, and related treatment effects for scalar outcomes. Counterfactual mean embeddings estimate intervention-specific outcome distributions in a reproducing kernel Hilbert space (Muandet et al., 2021), while recent generative methods learn conditional counterfactual laws (Xie & Huang, 2026; Tamano & Imaizumi, 2026). In those settings, the distribution is induced across units by scalar

or vector potential outcomes. Here the potential outcome for each unit is itself a probability measure, and the treatment-specific conditional barycenter averages unit-level quantile functions given covariates. This distinction determines both the estimand and the repeated-measurement observation model.

Counterfactual representation learning and repeated measurements. The Treatment-Agnostic Representation Network (TARNet) shares an encoder across treatment heads, counterfactual regression (CFR) adds treatment-group discrepancy regularization, and subsequent work studies support, invertibility, and covariate-preserving balancing (Shalit et al., 2017; Johansson et al., 2019; 2022; Du et al., 2021; Zhou et al., 2022). Our analysis uses the one-dimensional Wasserstein isometry to obtain function-valued prediction losses, while also tracking approximation caused by representation compression and a finite quantile grid. Repeated observation of latent distributions creates a second sampling layer. Wasserstein regression with empirical-measure responses and sparse barycenter estimation study this issue outside the present conditional causal prediction problem (Zhou & Müller, 2024; Peng et al., 2026). Wasserstein policy learning addresses the downstream decision problem once distribution-valued causal information is available (Huang et al., 2026).

3 Setup and Identification

Units, outcome distributions, and measurements. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a compact metric covariate space, let $I = [\underline{y}, \overline{y}]$ be a known bounded measurement interval, and write $D_Y = \overline{y} - \underline{y} > 0$. Let $\mathcal{P}_2(I)$ denote the probability measures on I. The latent tuples $(X_i, A_i, Y_i(0), Y_i(1))$, $i = 1, \ldots, n$, are independent and identically distributed, where $A_i \in \{0, 1\}$ and $Y_i(a) \in \mathcal{P}_2(I)$ is the distribution of a post-treatment measurement under treatment a. The target population is the marginal covariate law P_X.

The factual outcome measure is $Y_i = Y_i(A_i)$, following the standard potential outcome framework (Rubin, 1974; Imbens & Rubin, 2015). In the fully observed model, Y_i is observed directly. In the repeated-measurement model, we observe $(X_i, A_i, M_i, Z_{i1:M_i})$, where $M_i \geq 1$ is determined before the repeated draws and

$$Z_{ij} \text{ are conditionally independent with law } Y_i, \qquad \widetilde{Y}_i = \frac{1}{M_i} \sum_{j=1}^{M_i} \delta_{Z_{ij}}. \tag{1}$$

The counts may vary across units, but every unit receives equal weight within its treatment arm. Data splitting is performed by unit.

For $\nu \in \mathcal{P}_2(I)$, define its quantile function $Q_\nu(u) = \inf\{t : F_\nu(t) \geq u\}$. Let $\mathsf{H} = L^2(0, 1)$ with norm $\|g\|_{\mathsf{H}}^2 = \int_0^1 g(u)^2 \, \mathrm{d}u$, and let $\mathcal{Q}(I) \subset \mathsf{H}$ be the closed convex set of nondecreasing, I-valued functions. In one dimension,

$$W_2^2(\nu, \nu') = \|Q_\nu - Q_{\nu'}\|_{\mathsf{H}}^2. \tag{2}$$

Hence conditional Wasserstein barycenter estimation can be written as squared-loss regression with quantile-function responses (Panaretos & Zemel, 2020). The index u denotes quantile rank.

Causal target and prediction risk. For treatment $a \in \{0, 1\}$ and covariate value x, define the conditional 2-Wasserstein barycenter

$$\mu_a^\star(x) = \operatorname*{arg\,min}_{\nu \in \mathcal{P}_2(I)} \mathbb{E}[W_2^2\{Y(a), \nu\} \mid X = x]$$

(Petersen & Müller, 2019; Fan & Müller, 2025; Lin et al., 2023). By (2), its quantile function and the conditional barycenter contrast are

$$q_a^\star(x, u) = \mathbb{E}\{Q_{Y(a)}(u) \mid X = x\}, \qquad \Delta^\star(x, u) = q_1^\star(x, u) - q_0^\star(x, u), \qquad u \in (0, 1). \tag{3}$$

Integrating $\Delta^\star(x, \cdot)$ over rank gives the conditional treatment effect on the unit-level measurement mean. The full barycenter contrast can also reflect treatment differences in dispersion, tails, and shape.

Assumption 3.1 (Causal conditions)**.** Treatments are well defined, there is no interference between units, and $Y = Y(A)$. Furthermore, $(Y(0), Y(1)) \perp A \mid X$ and $0 < \mathbb{P}(A = 1 \mid X) < 1$ almost surely (Rosenbaum & Rubin, 1983; Imbens & Rubin, 2015).

Proposition 3.2 (Identification from fully observed factual measures)**.** *Under Assumption 3.1, the law of (X, A, Y) with Y fully observed identifies the conditional barycenter $\mu_a^\star(x) = \arg\min_{\nu \in \mathcal{P}_2(I)} \mathbb{E}[W_2^2\{Y(a), \nu\} \mid X = x]$, which is unique for P_X-almost every x, and satisfies*

$$Q_{\mu_a^\star(x)} = q_a^\star(x) = \mathbb{E}(Q_Y \mid X = x, A = a) \quad \textit{in } \mathsf{H}. \tag{4}$$

If $Y(a) = \delta_{S(a)}$ for scalar potential outcomes $S(a)$, then $\Delta^\star(x, u) = \mathbb{E}\{S(1) - S(0) \mid X = x\}$ for every u.

Proposition 3.2 identifies the latent conditional barycenter contrast when the factual outcome measure is observed. Appendix A gives the proof and the associated risk identities. Repeated measurements provide only a finite sample from each latent measure. With a fixed bounded number of draws per unit, the unrestricted latent barycenters need not be identified from the observed data. Proposition B.1 in Appendix B gives an explicit construction. We use the latent contrast in (3) as the scientific estimand, treat finite repeated measurements as an approximation source, and give consistency conditions as within-unit information increases.

For candidate quantile predictors $f_a : \mathcal{X} \to \mathcal{Q}(I)$, our primary risk is

$$\mathcal{E}_\Delta(f_0, f_1) = \mathbb{E}\,\|f_1(X) - f_0(X) - \Delta^\star(X)\|_{\mathsf{H}}^2 . \tag{5}$$

Linear functionals of the contrast, including the conditional mean effect and averages over upper quantile ranks, are controlled by this risk through Cauchy–Schwarz. Appendix A gives the corresponding identities.

4 Counterfactual Generalization

4.1 Prediction from a learned representation

For $a, b \in \{0, 1\}$, let

$$\mathcal{E}_a^b(f_a) = \mathbb{E}[\|f_a(X) - q_a^\star(X)\|_{\mathsf{H}}^2 \mid A = b], \tag{6}$$

where the lower index is the predicted treatment and the upper index specifies the covariate population. Define $\mathcal{E}_{\mu,a}(f_a) = \mathbb{E}\,\|f_a(X) - q_a^\star(X)\|_{\mathsf{H}}^2$. For $p_a = \mathbb{P}(A = a)$, the total target-population barycenter risk is

$$\mathcal{E}_\mu(f_0, f_1) = \sum_{a=0}^{1} \mathcal{E}_{\mu,a}(f_a) = \sum_{a=0}^{1} \{p_a \mathcal{E}_a^a + p_{1-a} \mathcal{E}_a^{1-a}\}. \tag{7}$$

The contrast risk satisfies

$$\mathcal{E}_\Delta \le 2\mathcal{E}_\mu = 2 \sum_{a=0}^{1} \{p_a \mathcal{E}_a^a + p_{1-a} \mathcal{E}_a^{1-a}\}. \tag{8}$$

Under Assumption 3.1, the factual arm-specific squared quantile loss equals $\mathcal{E}_a^a$ plus an arm-specific irreducible residual term that does not depend on f_a. Thus factual training targets the factual component of (8). The remaining problem is prediction over the other treatment group's covariate distribution (Shalit et al., 2017; Johansson et al., 2022).

Consider $\theta = (\Phi, h_0, h_1, \Psi)$, where $(\mathcal{Z}, d_{\mathcal{Z}})$ is a metric representation space, $\Phi : \mathcal{X} \to \mathcal{Z}$ is an encoder, $h_a : \mathcal{Z} \to \mathcal{Q}(I)$ are treatment-specific heads, $f_{\theta,a} = h_a \circ \Phi$, and $\Psi : \mathcal{Z} \to \mathcal{X}$ is a training-only decoder. The exact approximation error induced by a representation is

$$\inf_h \mathbb{E}\,\|q_a^\star(X) - h(\Phi(X))\|_{\mathsf{H}}^2 = \mathbb{E}\,\|q_a^\star(X) - \mathbb{E}\{q_a^\star(X) \mid \Phi(X)\}\|_{\mathsf{H}}^2 , \tag{9}$$

where the infimum is over measurable square-integrable heads. Thus sharing can help estimate covariate features used by both treatment-specific barycenters, while a representation that removes predictive variation creates irreducible approximation error. Appendix C proves (9).

Proposition 4.1 (A weak-overlap example)**.** *For each $\varepsilon \in (0, 1/2)$ there is a law satisfying Assumption 3.1 and a constant encoder such that*

$$W_1(P_0^\Phi, P_1^\Phi) = 0, \qquad \mathcal{E}_0^0 = \mathcal{E}_1^1 = 4\varepsilon, \qquad \mathcal{E}_\Delta = 4, \tag{10}$$

where $P_a^\Phi = \mathcal{L}(\Phi(X) \mid A = a)$. Hence factual errors and representation discrepancy alone cannot yield a multiplicative risk bound with a constant uniform over overlap lower bounds.

Appendix C proves Proposition 4.1 and gives the fixed-overlap comparison. The example isolates the role of overlap: exact representation balance can coexist with large target contrast error when the representation removes outcome-relevant covariate variation.

4.2 Transfer with reconstruction

Let $P_a^X = \mathcal{L}(X \mid A = a)$. Define the treatment-group representation discrepancy and reconstruction error by

$$D_\theta = W_1(P_0^\Phi, P_1^\Phi), \qquad \rho_{\theta,a} = \mathbb{E}[d_{\mathcal{X}}\{X, \Psi(\Phi(X))\} \mid A = a], \qquad R_\theta = \rho_{\theta,0} + \rho_{\theta,1}. \tag{11}$$

Assumption 4.2 (Regularity)**.** The target admits an $L_{q,a}$-Lipschitz version $q_a^\star : \mathcal{X} \to \mathsf{H}$. Every candidate head is $L_{h,a}$-Lipschitz into H and every decoder is L_Ψ-Lipschitz into $\mathcal{X}$. These finite constants are uniform over the candidate class.

Assumption 4.2 implies that representation-induced approximation in (9) is controlled by reconstruction:

$$\mathbb{E}\,\|q_a^\star(X) - \mathbb{E}\{q_a^\star(X) \mid \Phi(X)\}\|_{\mathsf{H}}^2 \le L_{q,a}^2\,\mathrm{diam}(\mathcal{X})\mathbb{E}d_{\mathcal{X}}(X, \Psi\Phi(X)).$$

Appendix C proves this reconstruction bound. This yields the following counterfactual comparison.

Theorem 4.3 (Counterfactual risk bound)**.** *Under Assumption 4.2, let $B_a = 2D_Y(L_{h,a} + L_{q,a}L_\Psi)$. Every candidate satisfies*

$$\mathcal{E}_a^{1-a}(f_{\theta,a}) \le \mathcal{E}_a^a(f_{\theta,a}) + B_a D_\theta + 2D_Y L_{q,a} R_\theta, \tag{12}$$

$$\mathcal{E}_\mu(\theta) \le \sum_a \mathcal{E}_a^a(f_{\theta,a}) + C_D D_\theta + C_R R_\theta, \tag{13}$$

$$\mathcal{E}_\Delta(\theta) \le 2\sum_a \mathcal{E}_a^a(f_{\theta,a}) + 2C_D D_\theta + 2C_R R_\theta, \tag{14}$$

where $C_D = p_1 B_0 + p_0 B_1$ and $C_R = 2D_Y(p_1 L_{q,0} + p_0 L_{q,1})$.

Theorem 4.3 connects target-population contrast error to factual prediction, treatment-group discrepancy, and reconstruction jointly. Appendix C gives the proof. The bound does not imply that lowering discrepancy or reconstruction alone lowers target risk because factual prediction and the other term can change with the fitted representation. The decoder also implies

$$W_1(P_0^X, P_1^X) \le R_\theta + L_\Psi D_\theta. \tag{15}$$

When covariates predictive of treatment assignment are retained for reconstruction, reducing R_θ can conflict with reducing D_θ. Appendix C proves (15), which motivates evaluating the two quantities jointly.

5 Estimator and Finite-Sample Analysis

5.1 Quantile outputs and objective

Partition $(0, 1]$ into K equal bins $B_j = ((j-1)/K, j/K]$. Define $[T_K q]_j = K \int_{B_j} q(u)\,\mathrm{d}u$ and let $\iota_K(v)(u) = v_j$ on B_j. Then $P_K = \iota_K T_K$ is the orthogonal projection onto binwise-constant functions. With $\mathcal{C}_K = \{v \in I^K : v_1 \le \cdots \le v_K\}$, a head with raw output $r_a(z) \in \mathbb{R}^K$ predicts

$$h_a(z) = \iota_K \Pi_{\mathcal{C}_K} r_a(z), \qquad \|\iota_K(v) - \iota_K(w)\|_{\mathsf{H}}^2 = K^{-1}\,\|v - w\|_2^2\,. \tag{16}$$

The Euclidean projection $\Pi_{\mathcal{C}_K}$ is nonexpansive and is computed by isotonic regression followed by clipping to I (Barlow et al., 1972). Bin averages of empirical quantile functions are computed exactly.

For the neural implementation, $\mathcal{X} = [0, 1]^d$ and $\mathcal{Z} = [-1, 1]^r$ use Euclidean metrics. For a fixed architecture, the class Θ_K contains finite feed-forward encoder, head, and decoder networks with bounded matrix spectral norms and bias norms. Encoder and decoder outputs are clipped to their respective cubes. These constraints give the uniform Lipschitz and covering-number bounds used

below. Appendix F gives the constants and the finite-catalogue extension used for predeclared architecture selection.

Let $n_a = \sum_i \mathbf{1}(A_i = a) \geq 1$ and $y_i^{(K)} = T_K Q_{\widetilde{Y}_i}$, replacing $\widetilde{Y}_i$ by Y_i when the latent outcome measure is directly observed. We train with

$$\widehat{L}_K(\theta) = \sum_{a=0}^{1} \frac{1}{n_a} \sum_{i:A_i=a} \frac{1}{K} \left\| y_i^{(K)} - T_K f_{\theta,a}(X_i) \right\|_2^2, \tag{17}$$

$$\widehat{R}_\theta = \sum_{a=0}^{1} \frac{1}{n_a} \sum_{i:A_i=a} d_{\mathcal{X}}(X_i, \Psi\Phi(X_i)), \tag{18}$$

$$\widetilde{J}_{\lambda,K}(\theta) = \widehat{L}_K(\theta) + \lambda_D \widetilde{D}_\theta + \lambda_R \widehat{R}_\theta, \qquad \lambda_D, \lambda_R \geq 0. \tag{19}$$

Here $\widetilde{D}_\theta$ approximates the full-sample empirical W_1 discrepancy $\widehat{D}_\theta$ between the two empirical representation distributions. With transport cost $c_{ij} = \|\Phi(X_i) - \Phi(X_j)\|_2$, uniform treatment-arm marginals, entropy regularization ϵ, and a feasible plan with entropic primal-dual gap at most τ_{OT} (Cuturi, 2013; Altschuler et al., 2017),

$$0 \leq \widetilde{D}_\theta - \widehat{D}_\theta \leq b_{\mathrm{OT}} := \epsilon \log \min(n_0, n_1) + \tau_{\mathrm{OT}}. \tag{20}$$

Appendix G gives the certified transport construction, including the unregularized case $\epsilon = 0$. Every training transport solve must satisfy this certificate before its update is accepted.

We optimize (19) with projected first-order updates. Model selection uses only unit-held-out factual prediction loss. Appendix J gives the optimization procedure, including finite predeclared candidate catalogues, and Appendix K gives the experimental protocol.

5.2 Error of the fitted estimator

Write $q^\star_{a,K} = P_K q^\star_a$ and define

$$J_{\lambda,K}(\theta) = \sum_a \mathbb{E}[\left\| f_{\theta,a}(X) - q^\star_{a,K}(X) \right\|_{\mathsf{H}}^2 \mid A = a] + \lambda_D D_\theta + \lambda_R R_\theta. \tag{21}$$

Three terms describe the approximation and estimation errors beyond $J_{\lambda,K}$ in (21). The quantile discretization error is $G_K = \mathbb{E}\,\|\Delta^\star - P_K\Delta^\star\|_{\mathsf{H}}^2$. The uniform empirical deviation $\Gamma_{\lambda,K,n}(\delta)$ is given explicitly in (48). Appendix F states its dependence on network size, sample size, and representation dimension. Repeated-measurement error is bounded by

$$\overline{\Xi}_M(\delta_M) = D_Y^2 \sum_{a=0}^{1} \frac{1}{n_a} \sum_{i:A_i=a} M_i^{-1/2} + 2D_Y^2 \sqrt{\frac{\log(1/\delta_M)}{2} \sum_{a=0}^{1} \frac{1}{n_a^2} \sum_{i:A_i=a} M_i^{-1}}. \tag{22}$$

For fully observed outcome measures, $\overline{\Xi}_M = 0$. Appendix E proves (22).

Theorem 5.1 (Finite-sample contrast-risk bound)**.** *Suppose Assumptions 3.1 and 4.2, the sampling model in Section 3, and the bounded network class of Appendix F hold. For $\delta, \delta_M \in (0,1)$, conditional on any treatment-label sequence with $n_0, n_1 \geq 1$, with probability at least $1 - \delta - \delta_M$, the following statements hold simultaneously for all $\theta \in \Theta_K$ and all nonnegative $\lambda = (\lambda_D, \lambda_R)$:*

$$\begin{aligned} \mathcal{E}_\Delta(\theta) \leq\ & G_K + 2\widetilde{J}_{\lambda,K}(\theta) + 2\Gamma_{\lambda,K,n}(\delta) + 2\overline{\Xi}_M(\delta_M) \\ & + 2(C_D - \lambda_D)_+ D_\theta + 2(C_R - \lambda_R)_+ R_\theta. \end{aligned} \tag{23}$$

Here $t_+ = \max(t, 0)$. For any returned model $\widehat{\theta}$, let $\epsilon_{\mathrm{opt}} = \widetilde{J}_{\lambda,K}(\widehat{\theta}) - \inf_{\theta\in\Theta_K} \widetilde{J}_{\lambda,K}(\theta)$. Then

$$\begin{aligned} \mathcal{E}_\Delta(\widehat{\theta}) \leq\ & G_K + 2 \inf_{\theta\in\Theta_K} J_{\lambda,K}(\theta) + 4\Gamma_{\lambda,K,n}(\delta) + 4\overline{\Xi}_M(\delta_M) \\ & + 2\lambda_D b_{\mathrm{OT}} + 2\epsilon_{\mathrm{opt}} + 2(C_D - \lambda_D)_+ D_{\widehat{\theta}} + 2(C_R - \lambda_R)_+ R_{\widehat{\theta}}. \end{aligned} \tag{24}$$

For bounded monotone quantiles, $G_K \leq D_Y^2/K$. If $q^\star_a(x,\cdot)$ is uniformly $L_{u,a}$-Lipschitz, then $G_K \leq (L_{u,0}^2 + L_{u,1}^2)/(6K^2)$.

Theorem 5.1 decomposes the bound into quantile discretization, the population objective $J_{\lambda,K}$, uniform empirical deviation, repeated-measurement error, transport approximation, optimization error, and remaining discrepancy or reconstruction terms. When $\lambda_D \geq C_D$ and $\lambda_R \geq C_R$, the residual terms vanish. Otherwise they remain explicit in (24). The decomposition separates errors reduced by more units, larger within-unit measurement counts, and finer quantile discretization, clarifying the distinct information needed for the latent target. Appendix D proves the quantile discretization bounds. The proof in Appendix H connects the implemented empirical loss to the projected population target and establishes the uniform event over the candidate networks and regularization weights.

For common $M_i = M$, (22) is of order $M^{-1/2}$, with concentration term of order $(n_{\min}M)^{-1/2}$ at fixed confidence. More generally, $n_a^{-1}\sum_{i:A_i=a} M_i^{-1/2} \to 0$ in each arm is sufficient for the measurement term to vanish.

5.3 Consistency under a fixed overlap lower bound

Let $\mathcal{A}_K$ denote the best armwise projected-target prediction risk in the candidate class, and let $\Gamma_{0,K,n}$ denote $\Gamma_{\lambda,K,n}$ with $\lambda_D = \lambda_R = 0$. Appendix I gives their definitions and the finite-sample comparison used below.

Corollary 5.2 (Consistency with increasing measurement counts)**.** *Under the conditions of Theorem 5.1, suppose* $\mathbb{P}(A = a \mid X) \geq \eta > 0$ *almost surely for both treatments. Along a sequence with* $\delta_n, \delta_{M,n} \to 0$*, assume* $K \to \infty$*,* $\mathcal{A}_K \to 0$*,* $\Gamma_{0,K,n}(\delta_n) \to 0$*,* $\overline{\Xi}_M(\delta_{M,n}) \to 0$*,* $\lambda_D\{\mathrm{diam}(\mathcal{Z}) + b_{\mathrm{OT}}\} + 2\lambda_R \,\mathrm{diam}(\mathcal{X}) \to 0$*, and* $\epsilon_{\mathrm{opt}} \to 0$ *in probability. Then* $\mathcal{E}_\mu(\widehat{\theta}) \to 0$ *and* $\mathcal{E}_\Delta(\widehat{\theta}) \to 0$ *in probability.*

Appendix I gives the proof. The fixed overlap lower bound converts arm-conditional projected prediction error into target-population error. The displayed conditions then make every term in the bound of Appendix I vanish.

Theorems 4.3 and 5.1 and Corollary 5.2 motivate three empirical checks: estimation of both conditional barycenters and their contrast, the relation of discrepancy and reconstruction to target risk, and the dependence of measurement and quantile discretization errors on M and K.

6 Experiments

In this section, we mainly investigate the following questions: **Q1.** Does shared fitting improve treatment-specific conditional Wasserstein barycenter estimation relative to Separate, and how does its contrast estimation compare with Separate, a doubly robust learner, and cycle-balanced representation learning? **Q2.** Does reducing empirical representation discrepancy or reconstruction error by itself imply lower target-population contrast risk? **Q3.** Do repeated-measurement and quantile discretization errors exhibit the dependence on M and K predicted by the theory? Appendix J gives implementation details, and Appendix K gives the data-generating processes and additional results.

6.1 Experimental setup

Benchmarks and protocol. We use three synthetic outcome families: smooth quantiles, equal means with a rank-dependent contrast, and two-point distributions with discontinuous quantiles. The primary experiments use $n_{\mathrm{train}} = 2000$, $n_{\mathrm{val}} = 1000$, $n_{\mathrm{test}} = 5000$, $M = 256$, and $K = 128$, with ten independently regenerated data sets. A semi-synthetic study uses covariates from the Wisconsin Diagnostic Breast Cancer (WDBC) data over five repeated splits after discarding the diagnosis label and newly simulating treatment and distribution-valued outcomes. Analytic conditional barycenters are used only for test evaluation. Appendix K gives the synthetic data-generating processes and WDBC protocol.

Comparisons and training. Separate fits two independent treatment-specific quantile networks. Shared uses one encoder with two treatment-specific quantile heads. Shared + discrepancy adds representation-discrepancy regularization, and Shared + discrepancy + reconstruction adds both dis-

Table 1: Target-population prediction errors (mean ± Monte Carlo SE. Lower is better). Panel A reports the contrast risk $10^4\mathcal{E}_\Delta$ for the three synthetic families and WDBC. Panel B reports the treatment-specific barycenter risks $10^4\mathcal{E}_{\mu,0}$ and $10^4\mathcal{E}_{\mu,1}$ for Separate and Shared on the synthetic families. Synthetic results use ten independently regenerated data sets, and WDBC uses five repeated splits. In Panel A, bold indicates the lowest mean and underline the second lowest.

Panel A: target-population contrast risk

Method	Smooth	Equal mean	Two-point	WDBC
Separate	3.853 ± 0.168	3.907 ± 0.215	20.660 ± 0.433	2.573 ± 0.456
Shared	$\underline{1.068 \pm 0.048}$	$\underline{0.436 \pm 0.022}$	$\mathbf{17.169 \pm 0.137}$	$\mathbf{1.769 \pm 0.147}$
Shared + discrepancy	1.178 ± 0.055	0.442 ± 0.041	17.720 ± 0.299	1.907 ± 0.217
Shared + discrepancy + reconstruction	$\mathbf{0.976 \pm 0.050}$	$\mathbf{0.416 \pm 0.039}^\dagger$	18.601 ± 0.497	$\underline{1.895 \pm 0.314}$
DR learner	1.706 ± 0.140	0.688 ± 0.094	17.892 ± 0.185	2.761 ± 1.034
Cycle-balanced	1.406 ± 0.069	0.463 ± 0.038	$\underline{17.441 \pm 0.171}$	2.188 ± 0.509

Panel B: treatment-specific target-population barycenter risk

	$10^4\mathcal{E}_{\mu,0}$		$10^4\mathcal{E}_{\mu,1}$	
Family	Separate	Shared	Separate	Shared
Smooth	3.192 ± 0.239	$\mathbf{0.842 \pm 0.039}$	3.493 ± 0.216	$\mathbf{1.286 \pm 0.071}$
Equal mean	3.245 ± 0.229	$\mathbf{0.806 \pm 0.029}$	3.296 ± 0.183	$\mathbf{0.856 \pm 0.035}$
Two-point	11.058 ± 0.339	$\mathbf{7.729 \pm 0.070}$	13.209 ± 0.347	$\mathbf{10.647 \pm 0.147}$

† Factual validation selects $\lambda_D = \lambda_R = 0$ in the equal-mean setting.

crepancy and covariate-reconstruction penalties. Shared + reconstruction is used only in the controlled ablation. We additionally compare with a cross-fitted quantile-function doubly robust learner (Kennedy, 2023; Chernozhukov et al., 2018), denoted DR learner, and a quantile-output adaptation of cycle-balanced representation learning (Zhou et al., 2022), denoted Cycle-balanced. Separate and Shared use closely matched predictive parameter counts for each candidate representation dimension. All methods use the same raw covariates, unit splits, and factual-only tuning budget. For each method and each synthetic outcome family, 20 pilot fits are run on an independent pilot data set using factual validation loss only. The selected hyperparameters are frozen before the reported experiments, and checkpoints are selected by held-out factual loss. Models are trained with full-batch Adam for at most 3000 updates. Architectures and search ranges are reported in Appendix J, with the frozen settings in Appendix J.3.

Evaluation metrics. The primary metric is the target-population contrast risk $\mathcal{E}_\Delta$ in (5). We also report the treatment-specific barycenter risks $\mathcal{E}_{\mu,0}$ and $\mathcal{E}_{\mu,1}$ to determine whether changes in contrast risk coincide with changes in both treatment-specific predictions. Tables report mean ± Monte Carlo standard error (SE) across independently regenerated data sets or repeated WDBC splits.

6.2 Conditional barycenter and contrast estimation

Shared fitting improves barycenter and contrast estimation. Table 1 shows that Shared has lower mean contrast risk than Separate, DR learner, and Cycle-balanced in all four settings. Shared also lowers both treatment-specific barycenter risks relative to Separate in every synthetic family. Shared + discrepancy + reconstruction has the lowest mean in Smooth. In Equal mean, the configuration selected from the Shared + discrepancy + reconstruction catalogue has $\lambda_D = \lambda_R = 0$, so it is an unregularized shared model with a separately selected architecture. Its mean contrast risk is the lowest numerical mean in that family. Shared has the lowest mean in Two-point and WDBC. Equal mean has $\int_0^1 \Delta^\star(x,u)\,\mathrm{d}u = 0$ with a nonzero rank-dependent contrast, so it evaluates distributional changes absent from the conditional mean effect.

6.3 Diagnostics for the theoretical error terms

Contrast risk depends on the joint transfer terms. Panels (a) and (b) of Figure 1 study empirical counterparts of the population quantities D_θ and R_θ in Theorem 4.3. Under stronger covariate-dependent treatment selection ($\kappa = 3$), Shared + discrepancy lowers $\widetilde{D}_\theta$ from 0.952 to 0.416, while

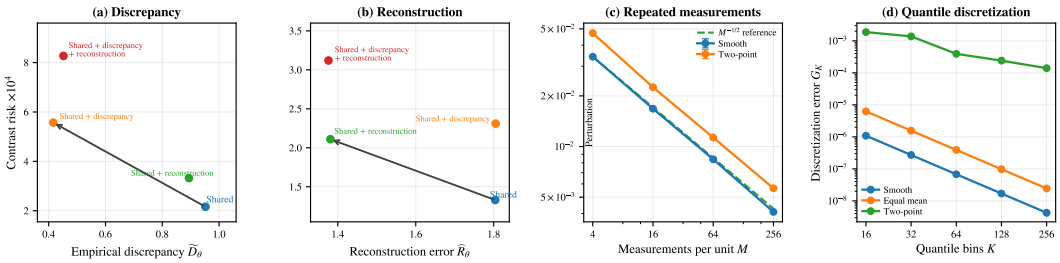

Figure 1: Diagnostics for the theoretical error terms. (a) Target contrast risk versus empirical representation discrepancy under stronger covariate-dependent treatment selection ($\kappa = 3$). (b) Target contrast risk versus empirical reconstruction error when one covariate affects treatment assignment but neither conditional barycenter. In panels (a) and (b), points show means over five independently regenerated data sets, and arrows mark the comparisons discussed in the text. (c) Fixed-predictor repeated-measurement loss perturbation as M varies, with an $M^{-1/2}$ reference slope. (d) Exact population quantile discretization error G_K as K varies.

$10^4\mathcal{E}_\Delta$ increases from 2.16 to 5.57. In a second setting where one covariate affects treatment assignment but neither conditional barycenter, Shared + reconstruction lowers $\widehat{R}_\theta$ from 1.804 to 1.381, while $10^4\mathcal{E}_\Delta$ increases from 1.33 to 2.11. These comparisons illustrate that decreasing empirical discrepancy or reconstruction error alone need not decrease target risk. Theorem 4.3 controls target risk through factual prediction, population representation discrepancy, and population reconstruction jointly. Numerical values for all four configurations are reported in Appendix K.4.

Measurement and quantile discretization errors exhibit the predicted scaling. Panels (c) and (d) of Figure 1 isolate the approximation terms due to repeated measurements and quantile discretization. For repeated measurements, we evaluate the absolute perturbation of the continuous factual squared-L^2 quantile loss using fixed data-independent predictors $f_0 \equiv -0.25$ and $f_1 \equiv 0.25$, held unchanged across M. As M increases from 4 to 256, the fitted log-log slope is -0.509 for both Smooth and Two-point, consistent with the $M^{-1/2}$ dependence in (22). For quantile discretization, we compute the exact population error G_K over $K \in \{16, 32, 64, 128, 256\}$. The quantity $K^2 G_K$ is constant for Smooth and Equal mean, while KG_K remains bounded for Two-point. These calculations recover K^{-2} behavior when $q_a^\star(x, \cdot)$ is Lipschitz and the general K^{-1} bound for bounded monotone quantiles. Additional values and the heterogeneous measurement-count diagnostic are reported in Appendix K.4.

7 Conclusion

We study counterfactual prediction when each potential outcome is a probability distribution observed through repeated measurements. Our estimator learns a shared covariate representation with treatment-specific quantile heads to estimate conditional Wasserstein barycenters and their contrast. The theory characterizes representation approximation and counterfactual transfer, then gives finite-sample risk bounds that separate statistical estimation, repeated measurements, quantile discretization, numerical transport, and optimization. It also distinguishes identification of the latent barycenter target from approximation based on finitely many within-unit measurements. Experiments show lower mean contrast risk for Shared than Separate in every reported setting, lower treatment-specific barycenter risk in every synthetic family, and behavior consistent with the predicted measurement and discretization terms. Together, these results provide a generalization framework for representation learning with distribution-valued potential outcomes under observational treatment assignment.

Limitations and future work. The current theory uses univariate distributions with bounded support and binary treatment. Extending it to multivariate distribution-valued outcomes is an interesting next step because the quantile isometry used here is one-dimensional. Multivalued and continuous treatments require transfer analysis over a richer treatment index. Another direction is to allow dependent or irregular within-unit measurements, which requires concentration results adapted to the observation process.

AI USE STATEMENT

Generative AI tools were used for language polishing. The authors retain full responsibility for the correctness of theoretical derivations, the experimental design, and results analysis.

REFERENCES

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Richard E. Barlow, David J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, 1972.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4):1713–1738, 2021.

Jianing Fan and Hans-Georg Müller. Conditional wasserstein barycenters and interpolation/extrapolation of distributions. *IEEE Transactions on Information Theory*, 71(5):3835–3853, 2025.

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.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Yiyan Huang, Cheuk Hang Leung, Qi Wu, and Zhiheng Zhang. Wasserstein policy learning for distributional outcomes. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 3349–3350. PMLR, 2026.

Su I Iao, Yidong Zhou, and Hans-Georg Müller. Deep fréchet regression. *Journal of the American Statistical Association*, 120(551):1437–1448, 2025.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 527–536. PMLR, 2019.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23(166):1–50, 2022.

Nathan Kallus and Miruna Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6037–6060. PMLR, 2023.

Srikar Katta, Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. Interpretable causal inference for analyzing wearable, sensor, and distributional data. In *International Conference on Artificial Intelligence and Statistics*, pp. 3340–3348. PMLR, 2024.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Kyum Kim, Yaqing Chen, and Paromita Dubey. Dfnn: A deep fréchet neural network framework for learning metric-space-valued responses. *arXiv preprint arXiv:2510.17072*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Daisuke Kurisu, Yidong Zhou, Taisuke Otsu, and Hans-Georg Müller. Geodesic causal inference. *arXiv preprint arXiv:2406.19604*, 2024.

Cheuk Hang Leung, Yijun Li, and Qi Wu. Distribution-valued causal machine learning: Implications of credit on spending patterns. *arXiv preprint arXiv:2509.03063*, 2025.

Zhenhua Lin, Dehan Kong, and Linbo Wang. Causal inference on distribution functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):378–398, 2023.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989.

Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

James Peng, Florian Stijven, Linbo Wang, and Peter B Gilbert. Estimating the wasserstein barycenter of one-dimensional distributions under sparse sampling. *arXiv preprint arXiv:2606.10096*, 2026.

Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019. doi: 10.1214/17-AOS1624.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data sciences. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Shu Tamano and Masaaki Imaizumi. Extended wasserstein-gan approach to causal distribution learning: Density-free estimation and minimax optimality. *arXiv preprint arXiv:2605.10206*, 2026.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.

Yifei Xie and Jian Huang. Repflow: Representation enhanced flow matching for causal effect estimation. *arXiv preprint arXiv:2605.05890*, 2026.

Kaicheng Zhang, Sinian Zhang, Doudou Zhou, and Yidong Zhou. Wasserstein transfer learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 110016–110049, 2025. doi: 10.52202/085713-3315.

Guanglin Zhou, Lina Yao, Xiwei Xu, Chen Wang, and Liming Zhu. Cycle-balanced representation learning for counterfactual inference. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pp. 442–450. SIAM, 2022.

Yidong Zhou and Hans-Georg Müller. Wasserstein regression with empirical measures and density estimation for sparse data. *Biometrics*, 80(4):ujae127, 2024.

Yidong Zhou, Su I Iao, and Hans-Georg Müller. End-to-end deep learning for predicting metric space-valued outputs. *Journal of Machine Learning Research*, 27(132):1–38, 2026.

A IDENTIFICATION AND PREDICTION RISK

The quantile representation and one-dimensional transport identities used below are standard (Panaretos & Zemel, 2020; Peyré & Cuturi, 2019). We give the conditional-expectation and risk arguments explicitly.

A.1 MEASURABILITY AND THE CONDITIONAL BARYCENTER

Equip $\mathcal{P}_2(I)$ with its Borel sigma field under W_2. The quantile map $\nu \mapsto Q_\nu$ is an isometry into the separable Hilbert space H. Consequently, a random probability measure induces a strongly measurable H-valued random variable. Its norm is bounded by $\max(|\underline{y}|, |\overline{y}|)$, so all conditional expectations below exist as Bochner integrals. Compact metric spaces are standard Borel, allowing regular conditional distributions and measurable versions of these expectations.

The set $\mathcal{Q}(I)$ is convex. It is also closed in H: from an L^2-convergent sequence choose a subsequence converging almost everywhere. On the common full-measure set of convergence, order and the bounds in I pass to the limit. Extending this ordered limit across the null set gives a nondecreasing I-valued representative of the same L^2 class. A Bochner average of elements of a closed convex set remains in that set, by approximation with simple random variables. Hence the conditional mean of a random quantile belongs to $\mathcal{Q}(I)$ almost surely.

Proof of Proposition 3.2. For fixed a, let $m_a(X) = \mathbb{E}[Q_{Y(a)} \mid X]$. For any X-measurable candidate $q(X)$, expand the squared norm as

$$\begin{aligned}\left\|Q_{Y(a)} - q(X)\right\|_{\mathsf{H}}^2 = \ & \left\|Q_{Y(a)} - m_a(X)\right\|_{\mathsf{H}}^2 + \left\|m_a(X) - q(X)\right\|_{\mathsf{H}}^2 \\ & + 2\langle Q_{Y(a)} - m_a(X), m_a(X) - q(X)\rangle_{\mathsf{H}}. \end{aligned} \tag{25}$$

Conditional expectation of the last term is zero, since the first factor has conditional mean zero and the second is X-measurable. Thus the unique minimizer in H, for P_X-almost every x, is $m_a(x)$. It lies in $\mathcal{Q}(I)$, so it corresponds to a unique probability measure and also minimizes the conditional Wasserstein criterion. Consistency and conditional exchangeability give

$$\mathbb{E}[Q_Y \mid X, A = a] = \mathbb{E}[Q_{Y(a)} \mid X, A = a] = \mathbb{E}[Q_{Y(a)} \mid X].$$

Positivity makes this identity informative for both arms at P_X-almost every covariate value. This proves (4). If $Y(a) = \delta_{S(a)}$, its quantile is the constant function $S(a)$. Taking conditional expectations and subtracting the two arms proves the scalar reduction. □

A.2 EXCESS RISK, CONTRAST RISK, AND SCALAR SUMMARIES

For a candidate f_a, set $e_a(X) = f_a(X) - q_a^\star(X)$. The pointwise Hilbert inequality

$$\|e_1 - e_0\|_{\mathsf{H}}^2 \le 2\,\|e_1\|_{\mathsf{H}}^2 + 2\,\|e_0\|_{\mathsf{H}}^2$$

follows by expanding the left side and applying Cauchy–Schwarz and $2st \le s^2 + t^2$. Integrate over P_X and apply total expectation over A to obtain (8). Conditional exchangeability identifies the factual conditional mean as $q_a^\star(X)$. The same orthogonal expansion as above yields

$$\mathbb{E}[\|Q_Y - f_a(X)\|_{\mathsf{H}}^2 \mid A = a] = \sigma_a^2 + \mathcal{E}_a^a(f_a), \qquad \sigma_a^2 = \mathbb{E}[\|Q_Y - q_a^\star(X)\|_{\mathsf{H}}^2 \mid A = a]. \tag{26}$$

The variance term is arm specific and independent of the candidate.

For $w \in \mathsf{H}$, define the linear functional $T_w(g) = \int_0^1 w(u)g(u)\,\mathrm{d}u$. In particular, $T_1(g)$ is the integral of g over quantile rank. Bounded support allows Fubini's theorem, and the quantile integral of a probability distribution equals its mean. Therefore

$$\int_0^1 \Delta^\star(x,u)\,\mathrm{d}u = \mathbb{E}\left[\int_I t\,Y(1)(\,\mathrm{d}t) - \int_I t\,Y(0)(\,\mathrm{d}t) \mid X = x\right].$$

For $w \in \mathsf{H}$, the pointwise Cauchy–Schwarz inequality gives $|T_w(g(X))|^2 \le \|w\|_{\mathsf{H}}^2\,\|g(X)\|_{\mathsf{H}}^2$. Set $g = f_1 - f_0 - \Delta^\star$ and integrate to prove the summary-risk bound in Section 3. This establishes the link between the stated contrast risk and mean or upper-quantile-rank summaries without changing the estimand.

B NONIDENTIFICATION WITH A FIXED BOUNDED MEASUREMENT COUNT

The following construction establishes nonidentification in the repeated-measurement observation model with a fixed bounded measurement count. Related sparse-sampling issues in barycenter estimation are studied by Peng et al. (2026).

Proposition B.1 (Nonidentification with a fixed measurement count)**.** *For every integer $M \geq 1$, there are two laws satisfying Assumption 3.1, with fixed covariates, that induce the same observed law of $(A, Z_1, \ldots, Z_M)$ but different $\Delta^\star$ in H. The assertion also holds if an independent observed count takes values only in $\{1, \ldots, M\}$.*

Proof. Let T be a random variable on $[0,1]$ and take the latent measure $Y_T = (1-T)\delta_0 + T\delta_1$. For an ordered binary vector of M draws having k ones, its probability is $\mathbb{E}[T^k(1-T)^{M-k}]$. This depends on the law of T only through its moments up to degree M.

Define the nonzero polynomial

$$p(t) = \frac{\mathrm{d}^{M+1}}{\mathrm{d}t^{M+1}}\{t^{M+1}(1-t)^{M+1}\}.$$

For each integer $j \leq M$, integration by parts $M+1$ times gives $\int_0^1 t^j p(t)\,\mathrm{d}t = 0$. Every boundary term vanishes because the polynomial inside braces has a zero of multiplicity $M+1$ at each endpoint. The final integral is zero because the $(M+1)$st derivative of t^j is zero. Choose $0 < \zeta < 1/\sup_{[0,1]}|p|$. Then $g_\pm(t) = 1 \pm \zeta p(t)$ are distinct positive probability densities with identical first M moments. They induce the same ordered M-draw law and, by marginalization, the same law of any smaller number of draws.

The latent quantile is $Q_{Y_T}(u) = \mathbf{1}(u > 1-T)$ apart from an immaterial endpoint. Therefore its mean is $\mathbb{P}(T > 1-u)$. The two mean quantiles differ by $2\zeta\int_{1-u}^1 p(t)\,\mathrm{d}t$. Its derivative is $2\zeta p(1-u)$, which is nonzero somewhere because p is nonzero. The two mean quantiles therefore differ on an interval and hence in H norm.

To embed these two laws in the causal model, take X constant, $A \sim \mathrm{Bernoulli}(1/2)$ independent of all potential outcomes, $Y(0) = \delta_0$, and $Y(1) = Y_T$ with density g_+ or g_-. Both worlds satisfy consistency, exchangeability, and positivity. Their observed repeated-draw laws are identical in each treatment arm, but their treated barycenter quantiles, and therefore their contrasts, differ. An independent count bounded by M simply mixes the identical smaller-draw laws, proving the final assertion. □

The proposition shows that bounded repeated measurements identify only finitely many mixture moments in this unrestricted model, which is insufficient to determine the latent conditional barycenter. The consistency result therefore allows within-unit measurement counts to increase, so the empirical within-unit measure approaches the latent outcome distribution.

C REPRESENTATION APPROXIMATION AND COUNTERFACTUAL TRANSFER

C.1 THE EXACT APPROXIMATION FROM A REPRESENTATION

Fix a measurable encoder and write $Z = \Phi(X)$ and $m(Z) = \mathbb{E}[q_a^\star(X) \mid Z]$. A measurable function of Z representing this conditional expectation exists because the spaces are standard Borel. For every square-integrable measurable head h,

$$\mathbb{E}\,\|q_a^\star(X) - h(Z)\|_{\mathsf{H}}^2 = \mathbb{E}\,\|q_a^\star(X) - m(Z)\|_{\mathsf{H}}^2 + \mathbb{E}\,\|m(Z) - h(Z)\|_{\mathsf{H}}^2\,. \tag{27}$$

Indeed, expansion leaves a cross term whose conditional expectation given Z is zero. The second term is nonnegative and equals zero for $h = m$, proving (9). The minimizing m is a quantile function almost surely, so restricting this infimum to quantile-valued heads leaves it unchanged. Restricting it to a particular neural head class may increase it.

To bound the infimum by reconstruction, choose $h(z) = q_a^\star(\Psi(z))$. This is measurable and bounded. Assumption 4.2 gives

$$\inf_h \mathbb{E}\,\|q_a^\star(X) - h(\Phi(X))\|_{\mathsf{H}}^2 \leq L_{q,a}^2\,\mathbb{E}d_{\mathcal{X}}^2(X, \Psi\Phi(X)).$$

Since $0 \le d_{\mathcal{X}} \le \operatorname{diam}(\mathcal{X})$, its square is at most $\operatorname{diam}(\mathcal{X}) d_{\mathcal{X}}$. This proves the connection to the unsquared reconstruction regularizer. The same proof works under either arm-specific covariate law.

C.2 Weak overlap and the fixed-overlap comparison

Proof of Proposition 4.1. Let $I = [-1, 1]$, let X be uniform on $\{-1, 1\}$, and set $Y(0) = Y(1) = \delta_X$. Assign treatment with $\mathbb{P}(A = 1 \mid X = 1) = 1 - \varepsilon$ and $\mathbb{P}(A = 1 \mid X = -1) = \varepsilon$. Then conditional exchangeability holds by construction, positivity holds for every $\varepsilon \in (0, 1/2)$, and $p_0 = p_1 = 1/2$. Bayes' rule gives

$$\mathbb{P}(X = -1 \mid A = 1) = \varepsilon, \qquad \mathbb{P}(X = 1 \mid A = 0) = \varepsilon.$$

The true arm quantiles are both $q_a^\star(x, u) = x$ and the true contrast is zero. Under $f_1 = 1$ and $f_0 = -1$, each factual squared error equals 4ε. The predicted contrast is the constant 2, so its risk is 4. A constant encoder has the same point-mass representation distribution in each arm, yielding zero discrepancy. Hence

$$\frac{\mathcal{E}_\Delta}{\mathcal{E}_0^0 + \mathcal{E}_1^1 + D_\theta} = \frac{1}{2\varepsilon} \longrightarrow \infty.$$

For any finite proposed constant, take ε small enough to contradict the proposed bound. The absolute risk stays bounded while the ratio diverges. □

For the fixed-overlap comparison, let $e_a(x) = \mathbb{P}(A = a \mid X = x) \ge \eta$. For nonnegative $r(X)$,

$$p_a \mathbb{E}[r(X) \mid A = a] = \mathbb{E}[e_a(X) r(X)] \ge \eta \mathbb{E} r(X).$$

Use $r(X) = \|f_a(X) - q_a^\star(X)\|_{\mathsf{H}}^2$, sum over the two treatments, and apply the contrast inequality to obtain $\mathcal{E}_\Delta \le (2/\eta) \sum_a p_a \mathcal{E}_a^a$. Thus non-injectivity alone is compatible with the stated fixed-overlap risk comparison.

C.3 Transfer bound and the two regularizers

Proof of Theorem 4.3. For any $v, w, t \in \mathcal{Q}(I)$, all pairwise H distances are at most D_Y. The identity

$$\big| \, \|v - t\|_{\mathsf{H}}^2 - \|w - t\|_{\mathsf{H}}^2 \, \big| = |\langle v - w, v + w - 2t \rangle_{\mathsf{H}}| \le 2 D_Y \, \|v - w\|_{\mathsf{H}}$$

follows by Cauchy–Schwarz. Fix an arm a and define

$$\ell_a(x) = \|h_a(\Phi(x)) - q_a^\star(x)\|_{\mathsf{H}}^2, \qquad g_a(z) = \|h_a(z) - q_a^\star(\Psi(z))\|_{\mathsf{H}}^2 .$$

The preceding inequality and target regularity imply

$$|\ell_a(x) - g_a(\Phi(x))| \le 2 D_Y L_{q,a} d_{\mathcal{X}}(x, \Psi\Phi(x)). \tag{28}$$

Changing both arguments of the squared distance, one at a time, gives

$$\begin{aligned} |g_a(z) - g_a(z')| &\le 2 D_Y \{ \|h_a(z) - h_a(z')\|_{\mathsf{H}} + \|q_a^\star(\Psi(z)) - q_a^\star(\Psi(z'))\|_{\mathsf{H}} \} \\ &\le 2 D_Y (L_{h,a} + L_{q,a} L_\Psi) d_{\mathcal{Z}}(z, z') = B_a d_{\mathcal{Z}}(z, z'). \end{aligned}$$

For any coupling (Z_0, Z_1) with marginals P_0^Φ, P_1^Φ, $|\mathbb{E} g_a(Z_1) - \mathbb{E} g_a(Z_0)| \le B_a \mathbb{E} d_{\mathcal{Z}}(Z_0, Z_1)$. Taking the infimum over couplings proves

$$|\mathbb{E}[g_a(\Phi(X)) \mid A = 1] - \mathbb{E}[g_a(\Phi(X)) \mid A = 0]| \le B_a D_\theta.$$

This is the Lipschitz transport inequality, which also follows from Kantorovich–Rubinstein duality (Peyré & Cuturi, 2019). Applying (28) on both sides yields

$$\mathcal{E}_a^{1-a} \le \mathbb{E}[g_a(\Phi(X)) \mid A = 1 - a] + 2 D_Y L_{q,a} \rho_{\theta, 1-a} \le \mathcal{E}_a^a + B_a D_\theta + 2 D_Y L_{q,a} (\rho_{\theta,a} + \rho_{\theta,1-a}).$$

Multiply this inequality by p_{1-a}, add $p_a \mathcal{E}_a^a$, and sum over a. By (7) and $p_a + p_{1-a} = 1$, this gives $\mathcal{E}_\mu \le \sum_a \mathcal{E}_a^a + C_D D_\theta + C_R R_\theta$, proving (13). Equation (8) then gives (14). This proves all conclusions. □

To prove (15), let $\Psi_\# P$ denote the law of $\Psi(Z)$ when Z has law P. The joint pair $(X, \Psi\Phi(X))$ in arm a is a coupling, so $W_1(P_a^X, \Psi_\# P_a^\Phi) \le \rho_{\theta,a}$. Pushing any representation coupling through Ψ gives $W_1(\Psi_\# P_0^\Phi, \Psi_\# P_1^\Phi) \le L_\Psi D_\theta$. The triangle inequality proves (15).

D Quantile Projection and Discretization Error

D.1 Projection and exact loss identities

For a closed convex subset C of Euclidean space, its least-squares projection $p = \Pi_C v$ satisfies $\langle v - p, t - p\rangle \le 0$ for every $t \in C$. To see this, compare p to $p + s(t - p)$ for $0 < s \le 1$, expand their squared distances from v, divide by s, and let $s \downarrow 0$. Apply the inequality twice to $p = \Pi_C v$ and $p' = \Pi_C v'$ to obtain

$$\|p - p'\|_2^2 \le \langle p - p', v - v'\rangle \le \|p - p'\|_2 \|v - v'\|_2 .$$

Canceling the first factor when nonzero proves nonexpansiveness. The same variational inequality implies $\|p - t\|_2^2 \le \|v - t\|_2^2$ for $t \in C$. These statements apply to $\mathcal{C}_K$.

For any $q \in \mathsf{H}$, the integral of $q - P_K q$ is zero on each bin. Thus it is orthogonal to every binwise-constant function and

$$\|q - \iota_K v\|_{\mathsf{H}}^2 = \|q - P_K q\|_{\mathsf{H}}^2 + K^{-1} \|T_K q - v\|_2^2 . \tag{29}$$

The bin averages of a nondecreasing function are nondecreasing, and averages of I-valued functions lie in I. Consequently $P_K \mathcal{Q}(I) \subset \mathcal{Q}(I)$. Orthogonal projection is a contraction in H. It is also a contraction in L^1: on each bin, the absolute integral of a function is no larger than the integral of its absolute value. Summing gives

$$\|P_K q - P_K q'\|_{L^1} \le \|q - q'\|_{L^1} . \tag{30}$$

Since P_K is bounded and linear, it commutes with conditional expectation. Thus

$$\mathbb{E}[P_K Q_Y \mid X, A = a] = P_K q_a^\star(X) = q_{a,K}^\star(X).$$

The projected target remains $L_{q,a}$-Lipschitz into H. The factual projected loss has the exact decomposition

$$\mathbb{E}[\|P_K Q_Y - f_{\theta,a}(X)\|_{\mathsf{H}}^2 \mid A = a] = \mathbb{E}[\left\|f_{\theta,a}(X) - q_{a,K}^\star(X)\right\|_{\mathsf{H}}^2 \mid A = a] + v_{a,K}, \tag{31}$$

where $v_{a,K} = \mathbb{E}[\left\|P_K Q_Y - q_{a,K}^\star(X)\right\|_{\mathsf{H}}^2 \mid A = a]$ is independent of θ. This follows by the conditional orthogonal expansion, not by treating the latent distribution as noise-free. Finally, the estimated contrast is binwise constant, so

$$\mathcal{E}_\Delta(\theta) = \mathbb{E}\,\|f_{\theta,1} - f_{\theta,0} - P_K \Delta^\star\|_{\mathsf{H}}^2 + G_K. \tag{32}$$

D.2 A bound permitting jumps

Let $q \in \mathcal{Q}(I)$. On bin B_j, let a_j, b_j be its essential lower and upper limits and set $d_j = b_j - a_j$. For a random variable V in $[a_j, b_j]$ with mean m, the inequality $(V - a_j)(b_j - V) \ge 0$ gives $\mathrm{Var}(V) \le (m - a_j)(b_j - m) \le d_j^2/4$. Apply this to $q(U)$ for U uniform on B_j:

$$\int_{B_j} (q - [T_K q]_j)^2 \,\mathrm{d}u \le \frac{d_j^2}{4K}.$$

Monotonicity implies $\sum_j d_j \le D_Y$, including when there are jumps at bin boundaries. Therefore

$$\|q - P_K q\|_{\mathsf{H}}^2 \le \frac{1}{4K} \sum_j d_j^2 \le \frac{D_Y^2}{4K}. \tag{33}$$

Apply the squared triangle inequality to the two target residuals to obtain $G_K \le D_Y^2/K$. A single jump of size D_Y at a bin midpoint attains $D_Y^2/(4K)$ in (33), demonstrating the order of this uniform approximation bound. A jump exactly on a bin boundary contributes no error. Thus the approximation error for a fixed discontinuous quantile can vary irregularly with K.

If q is L_u-Lipschitz, put $h = 1/K$ and use the exact variance identity on each bin:

$$\int_{B_j} (q - [T_K q]_j)^2 \,\mathrm{d}u = \frac{1}{2h} \int_{B_j} \int_{B_j} (q(u) - q(v))^2 \,\mathrm{d}u \,\mathrm{d}v \le \frac{L_u^2 h^3}{12}.$$

Here $\int_0^h \int_0^h (u - v)^2 \,\mathrm{d}u \,\mathrm{d}v = h^4/6$. Summing over bins gives $L_u^2/(12K^2)$. Applying the squared triangle inequality to the arm residuals yields the smoothness refinement in Theorem 5.1. An atom creates a flat portion of a quantile, not necessarily a jump. Gaps in distribution support can create quantile jumps. Bound (33) requires neither continuity nor a density.

E Repeated-Measurement Error

E.1 Uniform perturbation of the implemented factual loss

Fix the latent measures and measurement counts. For each unit write $q_i = Q_{Y_i}$ and $\widetilde{q}_i = Q_{\widetilde{Y}_i}$. For every bounded valid grid prediction f,

$$\begin{aligned}\left|\|P_K q_i - f\|_{\mathsf{H}}^2 - \|P_K \widetilde{q}_i - f\|_{\mathsf{H}}^2\right| &\le 2D_Y \|P_K(q_i - \widetilde{q}_i)\|_{L^1} \\ &\le 2D_Y \|q_i - \widetilde{q}_i\|_{L^1} = 2D_Y W_1(Y_i, \widetilde{Y}_i). \end{aligned} \tag{34}$$

The first step factors the difference of squares pointwise. The second uses (30), and the last is the one-dimensional quantile formula for W_1. Average with weight $w_i = 1/n_{A_i}$. The perturbation of the entire factual objective, uniformly in θ, is at most

$$\Xi_M = 2D_Y \sum_i w_i W_1(Y_i, \widetilde{Y}_i). \tag{35}$$

The same assertion holds without a grid by omitting P_K.

E.2 Concentration using all measurements

For a fixed measure Y_i, use the equivalent distribution-function formula for W_1. Conditional on Y_i, M_i, each empirical distribution-function value is a binomial average. Jensen's inequality gives

$$\begin{aligned}\mathbb{E}[W_1(Y_i, \widetilde{Y}_i) \mid Y_i, M_i] &= \int_I \mathbb{E}|F_{Y_i}(t) - F_{\widetilde{Y}_i}(t)| \,\mathrm{d}t \\ &\le \int_I \sqrt{F_{Y_i}(t)(1 - F_{Y_i}(t))/M_i} \,\mathrm{d}t \le \frac{D_Y}{2\sqrt{M_i}}. \end{aligned} \tag{36}$$

Tonelli's theorem justifies exchanging expectation and the nonnegative integral. If one repeated measurement in unit i is replaced, the two empirical measures can be coupled by keeping their other $M_i - 1$ observations matched. Their W_1 distance is at most D_Y/M_i. The reverse triangle inequality then shows that $V = \sum_i w_i W_1(Y_i, \widetilde{Y}_i)$ changes by at most $w_i D_Y/M_i$. There are M_i such independent coordinates in that unit. Conditional bounded differences (McDiarmid et al., 1989) consequently gives

$$\mathbb{P}\left(V - \mathbb{E}[V \mid (Y_i, M_i)_i] > t \mid (X_i, A_i, Y_i(0), Y_i(1), M_i)_i\right) \le \exp\left(-\frac{2t^2}{D_Y^2 \sum_i w_i^2/M_i}\right).$$

Combining this with (36), taking $t = D_Y\sqrt{\frac{1}{2}\log(1/\delta_M)\sum_i w_i^2/M_i}$, and multiplying by $2D_Y$ proves $\Xi_M \le \overline{\Xi}_M(\delta_M)$ with conditional probability at least $1 - \delta_M$.

The required independence is among repeated draws conditional on the latent unit tuple and measurement count. Measurement counts may depend on the latent unit tuples as long as the stipulated conditional draw law holds. The across-unit generalization event remains conditional only on treatment labels. The two events are combined by taking conditional probabilities over measurements and then averaging over the latent tuples and counts.

For $M_i = M$, the mean term is $2D_Y^2/\sqrt{M}$ and the concentration term is $2D_Y^2\sqrt{\log(1/\delta_M)(1/n_0 + 1/n_1)/(2M)}$. For varying counts, let $a_a = n_a^{-1}\sum_{i:A_i=a} M_i^{-1/2}$. Since $M_i^{-1} \le M_i^{-1/2}$, the concentration variance factor is at most $a_0/n_0 + a_1/n_1$. Thus $a_0, a_1 \to 0$ suffices to make both terms vanish at a fixed confidence level, even when some individual measurement counts remain small.

F Explicit Uniform Bounds for the Neural Class

This appendix specifies a finite neural class and derives a deviation bound that also covers the encoder in the transport term. The proof uses an elementary parameter cover, a multiscale transport bound, Hoeffding's inequality (Hoeffding, 1963), and bounded differences (McDiarmid et al., 1989). All complexity terms below follow from the stated bounded neural class.

F.1 Architecture and parameter sensitivity

Assume $d, r \geq 1$. Each of the encoder, two raw heads, and decoder has a fixed feed-forward architecture with at most L affine layers. Hidden activations are coordinatewise rectified linear units, which are 1-Lipschitz and vanish at zero. Every weight matrix W satisfies $\|W\|_{\mathrm{op}} \leq s$ for $s \geq 1$, and every bias vector has Euclidean norm at most $b \geq 0$. The encoder and decoder final outputs are clipped to $[-1,1]^r$ and $[0,1]^d$, respectively. Raw heads are projected as in (16). The parameter vector $\theta \in \mathbb{R}^P$ contains all scalar entries of these four networks. The set Θ_K consists of the parameters satisfying these constraints. In particular, each coordinate belongs to $[-B, B]$ with $B = \max(s, b) \geq 1$.

Let $R_0 = \max(\sqrt{d}, \sqrt{r})$ and define

$$C_{\mathrm{par}} = Ls^{L-1}(R_0 + Lb + 1). \tag{37}$$

Each individual network, including its output operation, changes uniformly by at most $C_{\mathrm{par}} \|\theta - \theta'\|_2$ when parameters are changed from θ to θ', in Euclidean output norm for the encoder and decoder and H norm for the heads. To verify this, first consider the raw affine/ReLU network. With input norm at most R_0, the norm after ℓ layers is at most $s^\ell(R_0 + \ell b)$, by induction using $s \geq 1$. Put $t = \|\theta - \theta'\|_2$. The difference of corresponding outputs at layer ℓ is at most

$$s \times \text{previous-layer difference} + t\{s^{\ell-1}(R_0 + (\ell-1)b) + 1\}.$$

Here a matrix perturbation has operator norm at most its Frobenius norm, which in turn is at most t. The same bound holds for a bias perturbation. Iterating over at most L layers bounds the difference by $Ls^{L-1}(R_0 + Lb + 1)t$. Output clipping is nonexpansive. The head projection and embedding have combined Lipschitz constant $1/\sqrt{K} \leq 1$, proving the stated claim. The network Lipschitz constants in their inputs are at most s^L. One may use $L_{h,a} = L_\Psi = s^L$ or valid smaller uniform bounds.

Partition the parameter cube into cells of side at most $t/\sqrt{P}$, and choose one member of Θ_K from each occupied cell. This gives a Euclidean t-cover with

$$N(t) \leq (1 + 2B\sqrt{P}/t)^P.$$

Take $t = (nC_{\mathrm{par}})^{-1}$ and write $N = N(t)$. For any model and its cover representative, all four component functions are uniformly within $u = 1/n$ in their respective output norms. Set $L_h = \max_a L_{h,a}$. The composed predictor and reconstruction function then differ by at most

$$(1 + L_h)u \quad \text{in prediction norm,} \qquad (1 + L_\Psi)u \quad \text{in reconstruction distance.} \tag{38}$$

For example, compare the heads first at a common encoder output, then compare the two encoder outputs under a fixed head.

F.2 Expected empirical transport distance on a cube

For $m \geq 1$ independent points from any probability distribution P_Z on $[-1,1]^r$, let $\widehat{P}_{Z,m}$ be their empirical law. Put $J_m = \lceil \log_2 m \rceil$, taking $J_1 = 0$, and define

$$\beta_r(m) = \min\left\{ 2\sqrt{r},\ 2\sqrt{r}\,2^{-J_m} + \sqrt{r}\sum_{j=1}^{J_m} 2^{-j} \min\left(2, \sqrt{2^{rj}/m}\right) \right\}. \tag{39}$$

Then

$$\mathbb{E}W_1(P_Z, \widehat{P}_{Z,m}) \leq \beta_r(m). \tag{40}$$

Here is a direct proof. Partition the cube into 2^{rj} dyadic subcubes at level j and move all mass in each finest subcube to its center. The sum of the two rounding costs is at most $2\sqrt{r}\,2^{-J_m}$. Connect every subcube center to its parent center. The length of an edge from level j is $\sqrt{r}\,2^{-j}$. For discrete masses at the leaves, cancel positive and negative mass differences within each subtree and pass its remaining signed mass to its parent. The absolute flow across an edge is the absolute difference between the two measures' masses in that subcube. This constructs a feasible transport along the

tree: splitting its flows into paths from surplus leaves to deficit leaves preserves all masses. Euclidean distance between leaf centers is at most the path length. Thus the transport cost is bounded by

$$\sqrt{r}\sum_{j=1}^{J_m} 2^{-j} \sum_{C \text{ at level } j} |P_Z(C) - \widehat{P}_{Z,m}(C)|$$

plus the rounding costs. For each partition with k cells, Jensen and Cauchy–Schwarz imply

$$\mathbb{E}\sum_C |P_Z(C) - \widehat{P}_{Z,m}(C)| \le \sum_C \sqrt{P_Z(C)(1-P_Z(C))/m} \le \sqrt{k/m}.$$

The sum is also at most 2. This proves the second expression inside the minimum in (39). The cube diameter proves the first. The construction works for arbitrary, including singular, distributions.

The bound has the dimensional rates stated in (49), consistent with classical empirical-Wasserstein convergence behavior (Fournier & Guillin, 2015). To see this, split the sum at $j_0 = \lfloor \log_2(m)/r \rfloor$. Up to j_0, its terms are bounded by $m^{-1/2}2^{(r/2-1)j}$. Thereafter their sum is bounded by a constant times 2^{-j_0}. For $r = 1$, the first sum is a convergent geometric sum and the remaining rounding cost is $O(m^{-1})$. For $r = 2$, there are $O(\log(m+1))$ terms of order $m^{-1/2}$. For $r > 2$, the increasing geometric sum and the tail are both $O(m^{-1/r})$. All constants depend only on r. This analysis makes the cost of high latent dimension explicit.

F.3 Uniform factual, reconstruction, and discrepancy deviations

Condition on any label sequence with both arms nonempty. The latent observations within arm a are independent with law $\mathcal{L}(X, Q_Y \mid A = a)$. Define

$$H_{K,n,\delta} = P\log(1 + 2B\sqrt{P}\, nC_{\text{par}}) + \log(12/\delta), \qquad D_X = \sqrt{d}, \quad D_Z = 2\sqrt{r}. \tag{41}$$

For $a \in \{0, 1\}$, set

$$F_a(\delta) = D_Y^2\sqrt{\frac{H_{K,n,\delta}}{2n_a}} + \frac{4D_Y(1+L_h)}{n}, \tag{42}$$

$$U_a(\delta) = D_X\sqrt{\frac{H_{K,n,\delta}}{2n_a}} + \frac{2(1+L_\Psi)}{n}, \tag{43}$$

$$V_a(\delta) = \beta_r(n_a) + D_Z\sqrt{\frac{H_{K,n,\delta}}{2n_a}} + \frac{2}{n}. \tag{44}$$

With conditional probability at least $1-\delta$, simultaneously for all $\theta \in \Theta_K$,

$$\left| \frac{1}{n_a} \sum_{i:A_i=a} \|P_K Q_{Y_i} - f_{\theta,a}(X_i)\|_{\mathsf{H}}^2 - \mathbb{E}[\|P_K Q_Y - f_{\theta,a}(X)\|_{\mathsf{H}}^2 \mid A = a] \right| \le F_a, \tag{45}$$

$$|\widehat{\rho}_{\theta,a} - \rho_{\theta,a}| \le U_a, \tag{46}$$

$$W_1(P_a^\Phi, \widehat{P}_a^\Phi) \le V_a. \tag{47}$$

In (46), $\widehat{\rho}_{\theta,a}$ is the arm-a summand in (18).

At a fixed cover representative, the factual losses lie in $[0, D_Y^2]$ and reconstruction distances lie in $[0, D_X]$. Hoeffding's inequality gives the first terms of F_a and U_a, with failure probability at most $2\exp(-H_{K,n,\delta})$ each. A perturbation by a nearby model changes a factual loss by at most $2D_Y(1+L_h)/n$, so changes the difference of empirical and population means by at most twice this amount. The analogous reconstruction change is $2(1+L_\Psi)/n$. These are the remaining terms in F_a, U_a.

For a fixed encoder, changing one observation changes its empirical law in W_1 by at most D_Z/n_a. By the reverse triangle inequality, the same bounds the change in $W_1(P_a^\Phi, \widehat{P}_a^\Phi)$. Apply bounded differences above its expectation and use (40). This gives $\beta_r(n_a) + D_Z\sqrt{H_{K,n,\delta}/(2n_a)}$ with failure probability at most $\exp(-H_{K,n,\delta})$. Replacing a cover encoder by a uniformly $1/n$-close encoder changes both the population and empirical pushforward by at most $1/n$ in W_1, giving the additional

$2/n$. The union bound over all N cover representatives, four two-sided deviations and two one-sided deviations, has failure probability at most $10N\exp(-H_{K,n,\delta}) \le \delta$. This proves (45)–(47).

The triangle inequality now gives

$$\sup_{\theta\in\Theta_K} |D_\theta - \widehat{D}_\theta| \le V_0 + V_1, \qquad \sup_{\theta\in\Theta_K} |R_\theta - \widehat{R}_\theta| \le U_0 + U_1.$$

The explicit term used in the main theorem is

$$\Gamma_{\lambda,K,n}(\delta) = \sum_{a=0}^{1} F_a(\delta) + \lambda_R \sum_{a=0}^{1} U_a(\delta) + \lambda_D \sum_{a=0}^{1} V_a(\delta). \tag{48}$$

For fixed depth, norm bounds, covariate dimension, and latent dimension, the displayed terms imply

$$\Gamma_{\lambda,K,n}(\delta) = O\Bigg((1+\lambda_D+\lambda_R)\sqrt{\frac{P\log\{n(P+1)\}+\log(1/\delta)}{n_{\min}}} + \lambda_D b_r(n_{\min}) + \frac{1+\lambda_D+\lambda_R}{n}\Bigg), \tag{49}$$

where P is the number of scalar network parameters, $n_{\min} = \min(n_0, n_1)$, and $b_r(m) = m^{-1/2}$ for $r = 1$, $b_r(m) = m^{-1/2}\log(m+1)$ for $r = 2$, and $b_r(m) = m^{-1/r}$ for $r \ge 3$. Let $\widehat{J}^{\mathrm{lat}}_{\lambda,K}$ use the latent projected factual loss, exact empirical transport, and empirical reconstruction. On the same event, (31) implies

$$\sup_{\theta\in\Theta_K} \left|\widehat{J}^{\mathrm{lat}}_{\lambda,K}(\theta) - \{J_{\lambda,K}(\theta) + v_{0,K} + v_{1,K}\}\right| \le \Gamma_{\lambda,K,n}(\delta). \tag{50}$$

The same base events apply to every λ, so this conclusion holds simultaneously for every pair of finite nonnegative weights. Furthermore,

$$D_\theta \le \widetilde{D}_\theta + V_0 + V_1, \qquad R_\theta \le \widehat{R}_\theta + U_0 + U_1, \tag{51}$$

using the one-sided transport approximation. Substituting these into (23) replaces its remaining population discrepancies by observable quantities, although the target regularity constants remain assumptions.

For a fixed finite architecture and dimension, $F_a, U_a, V_a \to 0$ as $n_0, n_1 \to \infty$. With growing architectures, the exact sufficient condition for the square-root terms to vanish is $H_{K,n,\delta}/n_{\min} \to 0$, together with vanishing displayed $1/n$ terms and transport term. Its magnitude depends on the specified network class. If S architectures or grid sizes are declared before observing training data, apply the result at level δ/S to each one. This covers their data-driven selection. Equation (35) is uniform over all bounded predictions and all K, so the same repeated-measurement event covers every declared grid size.

G Certified Approximation of Empirical Transport

G.1 The precise transport objective

Index control units by $i = 1, \dots, n_0$ and treated units by $j = 1, \dots, n_1$. Let $\alpha_i = 1/n_0$, $\beta_j = 1/n_1$, and let $\mathcal{U}(\alpha, \beta)$ be the nonnegative matrices with row sums α and column sums β. For a cost matrix $c_{ij} = d_{\mathcal{Z}}(\Phi(X_i), \Phi(X_j))$, define

$$F_\epsilon(c) = \min_{\pi\in\mathcal{U}(\alpha,\beta)} \left\{ \sum_{ij} c_{ij}\pi_{ij} + \epsilon \sum_{ij} \pi_{ij} \log \frac{\pi_{ij}}{\alpha_i\beta_j} \right\}, \tag{52}$$

with $0\log 0 = 0$. Then $F_0(c) = \widehat{D}_\theta$. This is the entropy-regularized transport cost with a *positive* relative-entropy term. The objective differs from a debiased Sinkhorn divergence and from evaluating only the transport cost of an entropic plan. This sign and normalization yield (20).

For $\epsilon > 0$, any vectors $f \in \mathbb{R}^{n_0}$ and $g \in \mathbb{R}^{n_1}$ give a dual lower bound

$$d_\epsilon(f, g) = \alpha^\top f + \beta^\top g - \epsilon \sum_{ij} \alpha_i \beta_j \left\{ \exp\left((f_i + g_j - c_{ij})/\epsilon \right) - 1 \right\} \le F_\epsilon(c). \tag{53}$$

To verify weak duality, write the entropy term equivalently as $\epsilon \sum_{ij} [\pi_{ij} \log(\pi_{ij}/m_{ij}) - \pi_{ij} + m_{ij}]$ with $m_{ij} = \alpha_i \beta_j$. Both masses sum to one on the feasible set. Add the row and column constraints with multipliers f, g and minimize each entry over $\pi_{ij} \ge 0$. Its minimizer is $m_{ij} \exp((f_i + g_j - c_{ij})/\epsilon)$, and substitution gives (53). Evaluating at any feasible plan proves the lower bound. Compactness of the feasible set and continuity of the entropy convention ensure existence of a primal minimizer.

G.2 Feasible plans and a computable error bound

At a finite Sinkhorn iterate, the positive matrix generally has only approximate marginals. Use the following rounding procedure, as in Altschuler et al. (2017). First scale each row down by the smaller of 1 and its target-to-current row-mass ratio. Then scale each column down analogously. Let π'' denote the resulting matrix. Its row and column masses are at most their targets. Let

$$u = \alpha - \pi'' \mathbf{1}, \quad v = \beta - (\pi'')^\top \mathbf{1}, \quad t = \mathbf{1}^\top u = \mathbf{1}^\top v.$$

If $t > 0$, return $\overline{\pi} = \pi'' + uv^\top / t$. If $t = 0$, return $\overline{\pi} = \pi''$. The residuals are nonnegative, and direct summation shows that the returned plan has exactly the required marginals. Zero current rows or columns can be left unchanged during downscaling. The final residual update supplies their missing mass.

Compute its primal value $F_\epsilon(c, \overline{\pi})$ and a dual lower bound from (53). A gap at most τ_{OT} implies

$$F_\epsilon(c) \le F_\epsilon(c, \overline{\pi}) \le F_\epsilon(c) + \tau_{\mathrm{OT}}.$$

Set $\widetilde{D}_\theta = F_\epsilon(c, \overline{\pi})$. Relative entropy is nonnegative. In addition, for every coupling with the specified marginals,

$$\mathrm{KL}(\pi \| \alpha \otimes \beta) = H(\alpha) + H(\beta) - H(\pi) \le \min\{H(\alpha), H(\beta)\} = \log \min(n_0, n_1).$$

Here H is finite discrete entropy. The inequality follows from $H(\pi) = H(\alpha) + \sum_i \alpha_i H(\pi_{i\cdot}/\alpha_i) \ge H(\alpha)$, and the analogous column decomposition. Evaluating the entropic objective at an optimal unregularized coupling proves

$$\widehat{D}_\theta \le F_\epsilon(c) \le \widehat{D}_\theta + \epsilon \log \min(n_0, n_1).$$

Combining these inequalities proves (20). When $\epsilon = 0$, use an unregularized feasible primal plan and dual variables satisfying $f_i + g_j \le c_{ij}$. Their gap gives the same assertion with $b_{\mathrm{OT}} = \tau_{\mathrm{OT}}$. These arguments separate objective approximation from neural optimization error.

For $\epsilon > 0$, log-domain alternating updates are

$$f_i \leftarrow -\epsilon \log \sum_j \beta_j \exp((g_j - c_{ij})/\epsilon),$$

$$g_j \leftarrow -\epsilon \log \sum_i \alpha_i \exp((f_i - c_{ij})/\epsilon).$$

These are Sinkhorn scaling expressed through dual potentials (Cuturi, 2013; Peyré & Cuturi, 2019). Reconstruct the positive matrix from the potentials, round it, and check the primal-dual gap. A maximum-iteration exit is recorded as a failed certified solve whenever the requested gap is not attained. Floating-point implementations must check feasibility and add numerical tolerance to the reported bound. The displayed objective-gap certificate requires more than a small marginal residual.

H Proof of the Learned-Estimator Bound

Proof of Theorem 5.1. First apply Theorem 4.3 to $q^\star_{a,K}$ in place of $q^\star_a$. Projection preserves the target range and cannot increase its covariate Lipschitz constant. Thus the same C_D, C_R are valid. For the grid-projected contrast risk, followed by (32), this gives

$$\mathcal{E}_\Delta(\theta) \le G_K + 2J_{\lambda,K}(\theta) + 2(C_D - \lambda_D)_+ D_\theta + 2(C_R - \lambda_R)_+ R_\theta. \tag{54}$$

The positive parts only discard favorable negative differences when a weight exceeds its sufficient threshold.

Let $c_K = v_{0,K} + v_{1,K} \geq 0$ and abbreviate $\Gamma = \Gamma_{\lambda,K,n}(\delta)$ and $\Xi = \overline{\Xi}_M(\delta_M)$. Let $\widehat{J}^{\mathrm{obs}}_{\lambda,K}$ use the observed empirical-measure factual loss and *exact* empirical W_1. On the event in Appendix F,

$$\sup_\theta |\widehat{J}^{\mathrm{lat}}_{\lambda,K}(\theta) - J_{\lambda,K}(\theta) - c_K| \leq \Gamma.$$

On the conditional measurement event in Appendix E,

$$\sup_\theta |\widehat{J}^{\mathrm{obs}}_{\lambda,K}(\theta) - \widehat{J}^{\mathrm{lat}}_{\lambda,K}(\theta)| \leq \Xi.$$

For every candidate whose transport routine returns a certified value,

$$\widehat{J}^{\mathrm{obs}}_{\lambda,K}(\theta) \leq \widetilde{J}_{\lambda,K}(\theta) \leq \widehat{J}^{\mathrm{obs}}_{\lambda,K}(\theta) + \lambda_D b_{\mathrm{OT}}. \tag{55}$$

The theoretical statement assumes a certified transport value for every candidate. A finite implementation records certificate failures. The conditional measurement event has failure probability at most δ_M for every collection of latent tuples and counts. Integrating over these variables, conditional only on labels, and combining with the across-unit event proves the stated probability $1 - \delta - \delta_M$. The across-unit event remains conditional only on treatment labels, including when measurement counts are informative.

On the intersection of the events,

$$\begin{aligned} J_{\lambda,K}(\theta) &\leq \widehat{J}^{\mathrm{lat}}_{\lambda,K}(\theta) + \Gamma - c_K \\ &\leq \widehat{J}^{\mathrm{obs}}_{\lambda,K}(\theta) + \Xi + \Gamma - c_K \\ &\leq \widetilde{J}_{\lambda,K}(\theta) + \Xi + \Gamma. \end{aligned}$$

Substitute this into (54) to prove (23). The one-sided sandwich in (55) absorbs transport approximation in this upper bound, while b_{OT} enters the oracle comparison below.

For the comparison with the best candidate, retain the constant c_K and use the definition of ϵ_{opt}:

$$\begin{aligned} J_{\lambda,K}(\widehat{\theta}) + c_K &\leq \widehat{J}^{\mathrm{lat}}_{\lambda,K}(\widehat{\theta}) + \Gamma \\ &\leq \widetilde{J}_{\lambda,K}(\widehat{\theta}) + \Xi + \Gamma \\ &\leq \inf_\theta \widetilde{J}_{\lambda,K}(\theta) + \epsilon_{\mathrm{opt}} + \Xi + \Gamma \\ &\leq \inf_\theta \widehat{J}^{\mathrm{obs}}_{\lambda,K}(\theta) + \lambda_D b_{\mathrm{OT}} + \epsilon_{\mathrm{opt}} + \Xi + \Gamma \\ &\leq \inf_\theta \widehat{J}^{\mathrm{lat}}_{\lambda,K}(\theta) + \lambda_D b_{\mathrm{OT}} + \epsilon_{\mathrm{opt}} + 2\Xi + \Gamma \\ &\leq \inf_\theta J_{\lambda,K}(\theta) + c_K + \lambda_D b_{\mathrm{OT}} + \epsilon_{\mathrm{opt}} + 2\Xi + 2\Gamma. \end{aligned}$$

All infima are over the same Θ_K. These inequalities remain valid if an infimum is not attained, by taking an approximating sequence. Cancel c_K, multiply by two, and apply (54) at $\widehat{\theta}$ to obtain (24). Appendix D proves the two bounds on G_K. The base statistical events hold for all finite nonnegative weights, which proves the simultaneous assertion. □

The theorem conditions on the realized treatment labels and therefore allows random treatment-arm sizes. Unconditionally, the same conclusion on the event that both arms are observed has failure probability at most $\delta + \delta_M + p_0^n + p_1^n$, because these last two terms are exactly the probabilities of observing no treated units or no control units. A run with an empty treatment arm cannot fit this estimator.

I Consistency of the Same Estimator with Vanishing Regularization

For a candidate class Θ_K, define

$$\mathcal{A}_K = \inf_{\theta \in \Theta_K} \sum_a \mathbb{E}[\big\| f_{\theta,a}(X) - q^\star_{a,K}(X) \big\|^2_{\mathsf{H}} \mid A = a], \tag{56}$$

$$G_{\mu,K} = \sum_a \mathbb{E}\, \|q^\star_a(X) - P_K q^\star_a(X)\|^2_{\mathsf{H}} \le \frac{D_Y^2}{2K}. \tag{57}$$

Write $D_X = \operatorname{diam}(\mathcal{X})$, $D_Z = \operatorname{diam}(\mathcal{Z})$, and $\Gamma_{0,K,n}$ for (48) with both regularization weights equal to zero. Under the fixed-overlap condition of Corollary 5.2, the proof below establishes the finite-sample bound

$$\begin{aligned}\mathcal{E}_\mu(\widehat{\theta}) \le G_{\mu,K} + \frac{1}{\eta}\big\{ &\mathcal{A}_K + 2\Gamma_{0,K,n}(\delta) + 2\overline{\Xi}_M(\delta_M) \\ &+ \lambda_D(D_Z + b_{\mathrm{OT}}) + 2\lambda_R D_X + \epsilon_{\mathrm{opt}} \big\}.\end{aligned} \tag{58}$$

Proof of Corollary 5.2. All expectations below evaluate a fitted function on a new unit from the target population. The randomness in the fitted function comes from the unit-level training sample, repeated draws, and any training or selection randomness. The simultaneous deviation event permits any measurable returned model, with its actual global empirical objective gap.

For this proof, write

$$S_K(\theta) = \sum_a \mathbb{E}[\big\| f_{\theta,a}(X) - q^\star_{a,K}(X) \big\|^2_{\mathsf{H}} \mid A = a], \qquad c_K = v_{0,K} + v_{1,K}.$$

Thus $\inf_\theta S_K(\theta) = \mathcal{A}_K$ and c_K is the irreducible projected factual variance from (31). The factual part of the event in Appendix F, combined with (34), gives, uniformly over θ,

$$\big|\widehat{L}_K(\theta) - S_K(\theta) - c_K\big| \le \Gamma_{0,K,n}(\delta) + \overline{\Xi}_M(\delta_M).$$

This step holds for arbitrary bounded response variance.

For brevity in this proof put $t_n = \Gamma_{0,K,n}(\delta) + \overline{\Xi}_M(\delta_M)$ and $u_n = \lambda_D(D_Z + b_{\mathrm{OT}}) + 2\lambda_R D_X$. Because all encoded observations lie in $\mathcal{Z}$, the exact empirical discrepancy is at most D_Z. Its certified approximation is at most $D_Z + b_{\mathrm{OT}}$ by (20). Each of the two arm-normalized reconstruction averages is at most D_X. Nonnegativity of the penalties and the definition of ϵ_{opt} therefore imply, for every comparison model $\theta \in \Theta_K$,

$$\begin{aligned} S_K(\widehat{\theta}) + c_K &\le \widehat{L}_K(\widehat{\theta}) + t_n \\ &\le \widetilde{J}_{\lambda,K}(\widehat{\theta}) + t_n \\ &\le \widetilde{J}_{\lambda,K}(\theta) + \epsilon_{\mathrm{opt}} + t_n \\ &\le \widehat{L}_K(\theta) + u_n + \epsilon_{\mathrm{opt}} + t_n \\ &\le S_K(\theta) + c_K + u_n + \epsilon_{\mathrm{opt}} + 2t_n. \end{aligned}$$

Cancel c_K and take the infimum over the same candidate class to obtain

$$S_K(\widehat{\theta}) \le \mathcal{A}_K + 2\Gamma_{0,K,n}(\delta) + 2\overline{\Xi}_M(\delta_M) + \lambda_D(D_Z + b_{\mathrm{OT}}) + 2\lambda_R D_X + \epsilon_{\mathrm{opt}}. \tag{59}$$

An approximating sequence justifies taking the infimum. The transport approximation is certified on the comparison class as in Theorem 5.1, and every numerical solve used in that comparison must satisfy the stated certificate.

For a nonnegative measurable function g, the fixed overlap lower bound gives

$$\eta \mathbb{E}[g(X)] \le \mathbb{E}[\mathbb{P}(A = a \mid X) g(X)] = p_a \mathbb{E}[g(X) \mid A = a].$$

Apply this pointwise in the fitted function with $g(X) = \left\| f_{\widehat{\theta},a}(X) - q^{\star}_{a,K}(X) \right\|_{\mathsf{H}}^2$. Orthogonality of P_K and the fact that $f_{\widehat{\theta},a}$ is binwise constant imply

$$\begin{aligned}\mathcal{E}_{\mu}(\widehat{\theta}) &= G_{\mu,K} + \sum_a \mathbb{E}\left\| f_{\widehat{\theta},a}(X) - q^{\star}_{a,K}(X) \right\|_{\mathsf{H}}^2 \\ &\le G_{\mu,K} + \eta^{-1} \sum_a p_a \mathbb{E}[\left\| f_{\widehat{\theta},a}(X) - q^{\star}_{a,K}(X) \right\|_{\mathsf{H}}^2 \mid A = a] \\ &\le G_{\mu,K} + \eta^{-1} S_K(\widehat{\theta}),\end{aligned}$$

where the last inequality uses $p_a \le 1$. Applying (33) separately to the two conditional quantiles gives $G_{\mu,K} \le D_Y^2/(2K)$. Substituting (59) proves (58).

For the asymptotic statement, keep the latent unit law and η fixed, and allow measurement designs, network classes, grids, and weights to follow the specified sequence. With independent units and $p_0, p_1 > 0$, the probability of an empty arm is $p_0^n + p_1^n \to 0$. The unconditional probability that the bound fails on a nonempty-arm sample is at most $\delta_n + \delta_{M,n} \to 0$. Its right-hand side converges to zero in probability by the stated assumptions and $K \to \infty$. Hence $\mathcal{E}_{\mu}(\widehat{\theta}) \to 0$ in probability. Equation (8) then gives the same conclusion for $\mathcal{E}_{\Delta}(\widehat{\theta})$. □

The sufficient conditions are a fixed positive overlap lower bound, vanishing penalty contributions, small approximation error, controlled network complexity, and a vanishing global optimization gap. They permit noninjective encoders and allow D_θ, R_θ, and the regularization weights to follow any sequence satisfying (58). When overlap deteriorates with sample size, the factor $1/\eta$ must be tracked explicitly.

J Implementation Specification

This section specifies a reproducible optimization procedure for the analyzed objective. Because neural optimization is nonconvex, the oracle comparison retains the global objective gap ϵ_{opt}. The uniform bound applies to every returned feasible network.

J.1 Empirical quantile bins and isotonic outputs

Sort the measurements in unit i as $Z_{i(1)} \le \cdots \le Z_{i(M_i)}$. The empirical quantile equals $Z_{i(\ell)}$ on $C_{i\ell} = ((\ell-1)/M_i, \ell/M_i]$. Therefore compute its bin-average target exactly by

$$[y_i^{(K)}]_j = K \sum_{\ell=1}^{M_i} Z_{i(\ell)} |B_j \cap C_{i\ell}|. \tag{60}$$

A two-pointer traversal of the sorted bin endpoints evaluates this expression in $O(M_i + K)$ time after sorting, with $O(M_i \log M_i)$ sorting cost. Units have equal weight within their treatment arm, irrespective of their measurement counts. Weighting every measurement equally instead would change the population regression problem when counts are informative.

To project a raw head vector, use the pool-adjacent-violators algorithm for ordinary least-squares isotonic regression (Barlow et al., 1972). Start with one block per coordinate. Each block stores its size and raw-output sum. Whenever two consecutive block means are in decreasing order, merge those blocks and replace their means by the mean of the combined block. Recheck the preceding block and continue until the block means are ordered. Assign each coordinate its block mean, then clip all coordinates to I.

The standard isotonic solution is characterized by the least-squares variational inequality over nondecreasing vectors. Clipping it preserves the constrained projection as follows. Let y be the unbounded isotonic projection of a raw vector r, and let t be its clipping to I. Every constant block of y has zero sum of residuals $r - y$, so $\langle r - y, y - t\rangle = 0$. For any nondecreasing $v \in I^K$,

$$\langle r - t, v - t\rangle = \langle r - y, v - y\rangle + \langle r - y, y - t\rangle + \langle y - t, v - t\rangle \le 0.$$

The first term is nonpositive by isotonic optimality. The second is zero, and the third is nonpositive coordinatewise by clipping. The projection variational inequality therefore identifies $t = \Pi_{\mathcal{C}_K} r$.

On a neighborhood where the block partition and clipping statuses remain unchanged, the derivative within a free block B is $\partial t_i / \partial r_j = 1/|B|$ for $i, j \in B$, and zero across different blocks. A clipped block has zero derivative. At block-merging or clipping boundaries the map can be nondifferentiable. A consistent limiting derivative may be used by the optimizer. The statistical analysis uses only the well-defined projected output and nonexpansiveness, not differentiability of this layer.

J.2 Networks, constraints, and optimization

A default predictive architecture has encoder widths $d \to 64 \to 64 \to r$, raw head widths $r \to 64 \to K$ for each arm, and decoder widths $r \to 64 \to d$. Apply ReLU at hidden layers only, then the output operations in Section 5. A default specification is $r = 4$, $K = 128$, and matrix and bias bounds $s = b = 3$ in Appendix F. These are starting choices for the prescribed experiments. To enforce the spectral constraint after each update, compute each weight matrix's singular value decomposition and cap its singular values at s. Project each bias vector onto its radius-b Euclidean ball. Exact singular-value capping supplies the certificate for the hard matrix-norm constraint.

Use full-batch Adam updates (Kingma & Ba, 2014) with a default learning rate 10^{-3}, at most 3000 updates, and validation checks every 50 updates. Retain the checkpoint with lowest unit-held-out factual loss. Stop after ten consecutive checks without improvement. The global optimization gap ϵ_{opt} remains unobserved under gradient-based training and validation stopping. The theorem's training sample consists of the units used in these gradient updates. Validation and test units remain outside that sample.

A default entropic transport specification is $\epsilon = 0.01$ and $\tau_{\mathrm{OT}} = 10^{-4}$ using the Euclidean metric on $\mathcal{Z} = [-1, 1]^r$. Use log-sum-exp computations for the updates in Appendix G and check rounded primal and dual values every 20 iterations. A solve is accepted only when the certificate reaches the prescribed gap tolerance within 10000 inner iterations. Otherwise the fit is recorded as failed. The inner-iteration budget and numerical tolerance were fixed before the reported benchmark. Every run records $\epsilon \log \min(n_0, n_1) + \tau_{\mathrm{OT}}$, so the transport approximation entering the analysis remains explicit.

At a gradient step, the computed plan supplies the derivative of the ideal entropic cost with respect to the cost matrix when the optimum is exact. With an inexact plan it supplies an approximate derivative. The full objective value is evaluated separately using its feasible primal value. More precisely, if the plan has gap at most τ_{OT}, for another cost matrix c' it satisfies

$$F_\epsilon(c') \le F_\epsilon(c) + \langle \overline{\pi}, c' - c \rangle + \tau_{\mathrm{OT}},$$

by evaluating the c' objective at the same feasible plan. Thus it is a controlled approximate supergradient of the concave cost-matrix value function. Backpropagate this plan through the Euclidean costs, with a valid subgradient at coincident latent points. This first-order procedure targets the specified transport objective. The gap between the returned checkpoint's actual objective and the global infimum therefore remains in (24).

For each method and synthetic outcome family, use 20 complete candidate fits on an independent factual-only pilot dataset, then freeze the selected setting across every reported repetition. Checkpoint selection within each reported fit still uses unit-held-out factual loss. The main-method catalogue uses $r \in \{2, 4, 8, 16\}$ and learning rates in $[3 \cdot 10^{-4}, 10^{-3}]$ on a log scale. Positive discrepancy coefficients are sampled on a log scale from $[10^{-4}, 3 \cdot 10^{-2}]$, while positive reconstruction and cycle coefficients are sampled from $[10^{-4}, 10^{-2}]$. Fixed anchor configurations at small coefficients are included. The Shared + discrepancy and Shared + discrepancy + reconstruction catalogues also include an explicit unregularized candidate. The cycle-balanced baseline keeps its defining balance, reconstruction, and cycle coefficients strictly positive. All random candidate catalogues are generated from the fixed experiment seed before any test contrast-risk evaluation. Hold $K = 128$ fixed in the primary comparison and vary it only in the designated quantile discretization experiment.

For the reported four-way component diagnostic, fix $r = 4$ and use a common frozen setting across the four cells. We reuse the positive coefficient pair $(\lambda_D, \lambda_R) = (0.002310, 0.005472)$ and learning rate 4.536×10^{-4} from the completed factual-only Shared + discrepancy + reconstruction pilot for the Smooth family, while fixing $r = 4$ for the controlled diagnostic. The same pair, learning

Algorithm 1 Shared-representation estimation of conditional barycenter contrasts

Require: Unit-level training and validation sets and a predeclared finite list of network settings, K, (λ_D, λ_R), and numerical tolerances.

for each candidate setting **do**
  Compute exact bin-average targets by (60), initialize all four networks, and enforce their norm constraints.
  for each full-batch update **do**
    Encode all training covariates, project both head outputs onto $\mathcal{C}_K$, and decode the representations.
    Compute the factual and reconstruction losses with the arm-normalized weights in (17)–(18).
    Compute a feasible transport plan and its primal-dual certificate as in Appendix G, and reject an uncertified numerical exit.
    Take an Adam step using the loss derivatives and the transport plan derivative, then reimpose matrix and bias constraints.
    At each validation check, evaluate factual validation loss, retain the best checkpoint, and apply the specified stopping rule.
  end for
  Record the retained model's training objective, numerical certificates, and validation loss.
end for
Select the candidate with smallest factual validation loss, with a deterministic tie-breaking rule.
return both quantile predictors $x \mapsto h_a(\Phi(x))$, $a = 0, 1$, and their contrast, then discard the decoder.

rate, backbone, and predictive initialization are used in the Shared, Shared + discrepancy, Shared + reconstruction, and Shared + discrepancy + reconstruction cells, with only the corresponding terms switched on or off. The same frozen settings are used at every κ and γ value. The main-method settings reported in Appendix J.3 were already frozen by the completed independent pilot and are not retuned in the reported benchmark.

J.3 Frozen Main-Method Pilot Settings

The table below records the settings selected by the independent pilot using factual validation loss only. These values were frozen before the reported contrast-risk benchmark and are reused without retuning in the primary comparison. The corresponding Smooth settings are also reused in the WDBC semi-synthetic experiment. Here lr denotes the learning rate and λ_{cyc} is the cycle-consistency coefficient used only by Cycle-balanced. The final column reports factual validation loss on the independent pilot. The r column is the shared representation dimension and is not applicable to Separate or DR learner.

Table 2: Frozen hyperparameters from the independent main-method pilot.

Family	Method	r	lr	λ_D	λ_R	λ_{cyc}	Val. loss
Smooth	Separate	–	0.000502	0	0	0	0.001674
Smooth	Shared	16	0.000887	0	0	0	0.001386
Smooth	Shared + discrepancy	16	0.000427	0.000147	0	0	0.001373
Smooth	Shared + discrepancy + reconstruction	16	0.000454	0.00231	0.00547	0	0.0014
Smooth	DR learner	–	0.000957	0	0	0	0.001923
Smooth	Cycle-balanced	16	0.000474	0.00577	0.000953	0.00723	0.001409
Equal mean	Separate	–	0.001	0	0	0	0.001678
Equal mean	Shared	8	0.000884	0	0	0	0.001386
Equal mean	Shared + discrepancy	4	0.001	0.0003	0	0	0.001402
Equal mean	Shared + discrepancy + reconstruction	4	0.001	0	0	0	0.001405
Equal mean	DR learner	–	0.001	0	0	0	0.001928
Equal mean	Cycle-balanced	4	0.001	0.001	0.0001	0.0003	0.001399
Two-point	Separate	–	0.000483	0	0	0	0.005095
Two-point	Shared	16	0.000441	0	0	0	0.004745
Two-point	Shared + discrepancy	4	0.000952	0.00174	0	0	0.004803
Two-point	Shared + discrepancy + reconstruction	8	0.000545	0.00108	0.000974	0	0.0048
Two-point	DR learner	–	0.000882	0	0	0	0.005401
Two-point	Cycle-balanced	16	0.000473	0.0015	0.000149	0.00546	0.00476

K Experimental Details and Additional Results

This appendix gives the data-generating processes, estimator definitions, experimental design, and supporting results. The main-method settings were frozen by the independent pilot using factual validation loss only in Appendix J.3 before the reported contrast-risk evaluations.

K.1 A common distribution-valued causal data model

Generate $X \sim \mathrm{Uniform}([0,1]^{10})$ and put $t_j = 2X_j - 1$. Independently generate unit-level random variables $\xi_0, \xi_1 \sim \mathrm{Uniform}([-1,1])$. With $\sigma(v) = (1+e^{-v})^{-1}$, treatment assignment is

$$\mathbb{P}(A = 1 \mid X) = 0.02 + 0.96\sigma\{\kappa(0.8t_1 + 0.8t_2 + \gamma t_3)\}. \tag{61}$$

The default is $\kappa = 1$, $\gamma = 0$. Set

$$b_0(x) = 0.15\sin(\pi t_1), \qquad b_1(x) = b_0(x) + 0.15t_2,$$
$$s_0(x) = 0.15 + 0.05t_2, \qquad s_1(x) = s_0(x) + 0.025(1+t_1).$$

For the smooth setting, define the potential random measure by its quantile

$$Q_{Y(a)}(u) = b_a(X) + 0.04\xi_a + s_a(X)(2u-1). \tag{62}$$

It is monotone because $s_a \geq 0.1$, and its values lie in $[-1,1]$. The unit-level random shift makes the potential outcome a genuinely random measure given X. The barycenter target has the same formula without $0.04\xi_a$, since this shift has conditional mean zero. Thus

$$\Delta^\star(x,u) = 0.15t_2 + 0.025(1+t_1)(2u-1).$$

Generate repeated measurements by evaluating the factual potential quantile at independent uniform ranks. Only the factual measurements are supplied to the estimator. Independently stored analytic targets are used only for test evaluation and diagnostics that explicitly use population quantities.

For an equal-mean/different-shape setting, retain b_0 but set $b_1 = b_0$, $s_0 = 0.15$, and $s_1 = 0.15 + 0.06(1+t_1)$. Its mean treatment effect is zero, whereas $\Delta^\star(x,u) = 0.06(1+t_1)(2u-1)$ can be nonzero. This isolates what the full contrast curve adds beyond mean effect estimation.

For a discontinuous-quantile setting, retain the original b_a, s_a, set $u_0^\circ = 0.37$, $u_1^\circ = 0.61$, and use

$$Q_{Y(a)}(u) = b_a(X) + 0.04\xi_a + 2s_a(X)\{\mathbf{1}(u > u_a^\circ) - (1-u_a^\circ)\}. \tag{63}$$

These are two-point distributions with fixed arm-specific masses and covariate-dependent support points. They remain within $[-1,1]$. Their barycenter quantiles again omit the mean-zero unit shift. Fixed jump ranks ensure Lipschitz dependence on covariates in L^2.

To create a setting where one covariate affects treatment assignment but neither conditional barycenter, set $\gamma = 2$ in (61) while leaving the potential outcomes unchanged. The variable X_3 then affects treatment but neither arm's conditional barycenter. Covariates $X_4, \ldots, X_{10}$ are irrelevant to both. In all cases, assignment is generated from X independently of ξ_0, ξ_1 and of all measurement ranks. Thus exchangeability and consistency hold by construction. The lower bound 0.02 is maintained in these finite-sample benchmarks. The analytic example in Proposition 4.1 examines a lower bound tending to zero.

K.2 Comparisons and information available to each method

The main comparison has six estimators: (i) Separate, which fits treatment-specific quantile-output regressions independently, (ii) Shared, which uses a common encoder with treatment-specific heads, (iii) Shared + discrepancy, which adds representation-discrepancy regularization, (iv) Shared + discrepancy + reconstruction, which adds both discrepancy and covariate-reconstruction penalties, (v) a cross-fitted quantile-function doubly robust learner, and (vi) Cycle-balanced, a quantile-output adaptation of cycle-balanced representation learning (Zhou et al., 2022). Shared + reconstruction is used only in the four-way component ablation.

For the doubly robust baseline, fit a propensity model and treatment-specific regressions for the K-dimensional bin-average quantile target on complementary unit folds, following the two-stage

DR-learner construction and cross-fitting principle (Kennedy, 2023; Chernozhukov et al., 2018). Let $\widehat{e}(x)$ be the fitted treated propensity, set $\widehat{e}_1 = \widehat{e}$ and $\widehat{e}_0 = 1 - \widehat{e}$, and let $\widehat{m}_a(x) \in \mathbb{R}^K$ be the fitted treatment-specific mean vector. Form held-out pseudo-outcomes

$$\widehat{\Gamma}_{a,i} = \widehat{m}_a(X_i) + \frac{\mathbf{1}(A_i = a)}{\widehat{e}_a(X_i)}\{y_i^{(K)} - \widehat{m}_a(X_i)\}. \tag{64}$$

The effect regression fits $\widehat{\Gamma}_{1,i} - \widehat{\Gamma}_{0,i}$ on X_i with the $K^{-1}\|\cdot\|_2^2$ loss. A contrast curve is unrestricted in rank order, so neither the pseudo-outcome nor the fitted contrast is projected onto $\mathcal{C}_K$. Final predictions are clipped coordinatewise to $[-D_Y, D_Y]$, while the training pseudo-outcome remains unclipped. We use three-fold unit cross-fitting. The propensity model is logistic regression on the raw covariates. Its fitted probabilities are clipped to $[0.02, 0.98]$ before forming the inverse-propensity factors. With finite M, the augmented regression targets the conditional empirical-quantile mean and therefore inherits the latent-target measurement bias. For hyperparameter and checkpoint selection of the effect regression, we construct the fixed nuisance offset $\widehat{r}(x) = \{1-\widehat{e}(x)\}\widehat{m}_0(x)+\widehat{e}(x)\widehat{m}_1(x)$ from training-fold nuisances and minimize held-out factual loss of $\widehat{r}(X) + \{A - \widehat{e}(X)\}\widehat{\Delta}(X)$. This validation criterion uses only observed factual outcomes.

For the Cycle-balanced baseline, we retain the published representation-balancing and cycle-reconstruction objectives (Zhou et al., 2022). We replace the scalar factual response loss by the same normalized squared loss on the K-dimensional bin-average quantile targets used by the other neural estimators. The Separate baseline is an unregularized neural T-learner that fits the two treatment regressions independently (Künzel et al., 2019). Each arm can therefore learn its own covariate representation.

To isolate parameter sharing from predictive-capacity differences, Separate and Shared use closely matched predictive parameter budgets and the same factual-only tuning budget. Their selected predictive parameter counts are reported in Appendix K.4. The four component ablations use exactly the same encoder and prediction heads, and the training-only decoder is removed at inference. All methods receive the same raw covariates, factual measurements, unit splits, and hyperparameter-trial budget.

K.3 Executed experimental groups

Conditional estimation and distribution-valued effects. The three primary outcome families use $n_{\text{train}} = 2000$, $n_{\text{val}} = 1000$, $n_{\text{test}} = 5000$, $M = 256$, and $K = 128$. Test covariates are independent of training and validation. The complete data-generating process is regenerated ten times with paired data seeds across methods. Table 1 uses all ten scheduled paired repetitions for every method and synthetic family combination.

Controlled regularization stress test. The reported diagnostic fixes $r = 4$ and uses four matched penalty configurations with the common positive coefficient pair described in Appendix J. Five independent data-generating repetitions are run for each of $(\kappa, \gamma) \in \{(0,0), (1,0), (3,0), (1,2)\}$. The $\gamma = 2$ condition changes treatment assignment through X_3 while leaving both treatment-specific conditional barycenters unchanged. We report contrast risk, empirical representation discrepancy, and empirical reconstruction for the four configurations.

Measurement and quantile discretization diagnostics. The main benchmark evaluates finite-sample estimator performance. The measurement and quantile discretization statements are evaluated separately on the quantities to which they directly apply. For $M \in \{4, 16, 64, 256\}$, fixed data-independent predictors $f_0 \equiv -0.25$ and $f_1 \equiv 0.25$ are evaluated with the continuous squared-L^2 quantile loss against latent and empirical measures. The predictors are unchanged across M, so the comparison isolates repeated-measurement approximation. A second diagnostic changes the fraction of units with $M_i = 4$ while the remaining units have $M_i = 1024$. For $K \in \{16, 32, 64, 128, 256\}$, G_K is computed analytically from the known conditional barycenter contrast. These diagnostics isolate the measurement and quantile discretization terms from the complete learned estimator.

Semi-synthetic real covariates. The Wisconsin Diagnostic Breast Cancer data (Wolberg et al., 1993) are used only as a finite pool of baseline covariates, and the diagnosis label is discarded.

We retain the ten "mean" features and transform coordinate j as $x_j = \min\{1, \max\{0, (x_j^{\mathrm{raw}} - x_j^{\mathrm{lo}})/(x_j^{\mathrm{hi}} - x_j^{\mathrm{lo}})\}\}$ using fixed vectors $x^{\mathrm{lo}} = (5, 5, 30, 100, .04, 0, 0, 0, .10, .04)$ and $x^{\mathrm{hi}} = (30, 40, 200, 2500, .17, .40, .50, .25, .35, .10)$. These constants are fixed before the reported experiments. A permanent pool of 142 of the 569 records is reserved for pilot construction and never enters reported fits. Each of the five reported repetitions permutes the remaining 427 records and uses 256 for training, 85 for validation, and 86 for testing. Treatment assignments and distribution-valued outcomes are newly simulated from the Smooth formulas. No WDBC-specific hyperparameter tuning is performed, and each method reuses its frozen Smooth configuration. The reported uncertainty is conditional on this finite covariate pool.

K.4 Additional quantitative results

Separate vs. Shared capacity and arm-specific barycenter errors. The Separate architecture uses two independent quantile networks. For each Separate pilot candidate, its hidden width is chosen to match the predictive parameter count of the corresponding shared-architecture candidate at the same candidate value of r. Separate and Shared are selected independently by factual validation, so their final selected dimensions can differ. The selected predictive parameter counts are 22,208 versus 24,720 (Smooth), 23,224 versus 23,176 (Equal mean), and 24,778 versus 24,720 (Two-point) for Separate versus Shared, respectively. The two architectures therefore have similar predictive parameter counts at the scale of this benchmark, and each estimator uses the same factual-only tuning budget.

Treatment-specific conditional barycenter risks are reported in Panel B of Table 1. Shared reduces both arm-specific risks relative to Separate in every synthetic family.

Controlled discrepancy and reconstruction diagnostic. Panels (a) and (b) of Figure 1 plot target risk against empirical representation discrepancy or empirical reconstruction error. The arrows highlight the comparisons used in the main text: in these settings, the penalized quantity decreases while contrast risk worsens. This is consistent with Theorem 4.3, which controls target error jointly through factual prediction, representation discrepancy, and reconstruction.

For the condition $(\kappa, \gamma) = (3, 0)$, the mean triples $(10^4 \mathcal{E}_\Delta, \widetilde{D}_\theta, \widehat{R}_\theta)$ are $(2.16, 0.952, 1.802)$ for Shared, $(5.57, 0.416, 1.803)$ for Shared + discrepancy, $(3.32, 0.894, 1.392)$ for Shared + reconstruction, and $(8.27, 0.452, 1.383)$ for Shared + discrepancy + reconstruction. For the condition $(\kappa, \gamma) = (1, 2)$, the corresponding triples are $(1.33, 0.419, 1.804)$, $(2.31, 0.287, 1.805)$, $(2.11, 0.395, 1.381)$, and $(3.12, 0.296, 1.376)$. These are descriptive means over five independent data-generating repetitions. Paired comparisons are computed within the same regenerated data sets.

Repeated-measurement diagnostics. For the Smooth family, the fixed-predictor absolute loss perturbation is 0.03419 ± 0.00041, 0.01678 ± 0.00019, 0.00840 ± 0.00011, and 0.00409 ± 0.00005 for $M = 4, 16, 64, 256$, respectively. The fitted log-log slope of the four means is -0.509. The Two-point family gives 0.04720 ± 0.00064, 0.02252 ± 0.00036, 0.01131 ± 0.00011, and 0.00565 ± 0.00012, with slope -0.509. Here the uncertainty is Monte Carlo SE over five regenerated measurement experiments. In the heterogeneous-count Smooth diagnostic, increasing the fraction of $M_i = 4$ units from 1% to 10%, 50%, and 100% changes the arm-average $M_i^{-1/2}$ sum from 0.0717 to 0.156, 0.531, and 1.000, while the corresponding perturbation rises from 0.00234 to 0.00528, 0.01818, and 0.03419.

Exact quantile discretization. For the Smooth family, $K^2 G_K = 2.78 \times 10^{-4}$ for every $K \in \{16, 32, 64, 128, 256\}$. For the Equal mean family, $K^2 G_K = 1.60 \times 10^{-3}$ throughout. For the Two-point family, $K G_K$ ranges from 0.0252 to 0.0442 over the same values of K, reflecting the jump/bin-boundary alignment described in Appendix D. Thus the exact projection calculation distinguishes the K^{-2} behavior for quantiles that are Lipschitz in u from the general K^{-1} guarantee for bounded monotone quantiles. The calculation contains no neural optimization error.

K.5 METRICS AND UNCERTAINTY

The principal metric is the test average of (5) against the analytic conditional barycenter contrast. This differs from the contrast between two realized unit-specific potential measures, which contains unit-level randomness. Armwise prediction error is the test mean of $\|f_a - q_a^\star\|_{\mathsf{H}}^2$, equivalently squared W_2 distance to the conditional barycenter, evaluated over the same target covariate population for each treatment. These armwise errors correspond to the summands of $\mathcal{E}_\mu$ in (7) and are summarized in Panel B of Table 1.

Settings are selected from factual validation data with unit-level splits. Paired means and Monte Carlo standard errors use ten independently regenerated data sets for the primary synthetic comparison and five for each controlled diagnostic. Each regenerated data set uses one neural initialization. WDBC repetitions condition on the observed covariate pool and vary the split, treatment assignment, and simulated outcomes.