POLICY LEARNING UNDER STABLE RANK DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Policy learning aims to personalize treatment decisions to improve population welfare. Yet population welfare maximization does not account for individual-level benefit and harm. Such quantities depend on the joint distribution of potential outcomes, while observed data generally identify only their conditional marginals. Existing approaches either optimize over separate worst-case couplings across covariate values or impose parametric restrictions on the unidentified dependence between potential outcomes, leading to conservativeness or strong structural assumptions. In this paper, we study policy learning when rank dependence is stable across covariates, allowing the conditional marginals to vary freely while leaving the common copula otherwise unrestricted. This restriction forces the least-favorable copula to be chosen jointly over the covariate distribution, yielding a sharper maximin policy value than rectangular partial identification. We derive the sharp value, characterize exactly when copula invariance changes the optimal policy, and develop a finite-sample learner that directly optimizes the resulting objective without estimating the unidentified copula. We establish uniform policy-value and regret guarantees and quantify robustness to approximate violations of copula invariance. Experiments confirm the predicted policy-switch behavior and show that the learner remains effective under moderate violations of copula invariance.

1 INTRODUCTION

Policy learning seeks to optimize treatment assignment rules based on observed covariates (Qian & Murphy, 2011; Kitagawa & Tetenov, 2018; Kallus, 2018). A standard objective is to maximize the policy value, defined as the expected outcome under the learned assignment rule (Manski, 2004; Swaminathan & Joachims, 2015; Athey & Wager, 2021). While policy value provides a natural criterion for population welfare, it does not capture how individuals fare under the assigned treatment relative to the alternative (Linn et al., 2017; Wang et al., 2018; Qi et al., 2023). In medicine, for example, a policy may improve outcomes on average while assigning treatment to patients who would have fared better under control (Shen et al., 2013; Ben-Michael et al., 2024). This motivates accounting for individual treatment comparisons in policy learning (Zhang et al., 2013; Kallus, 2022; Ben-Michael et al., 2024).

Individual treatment comparisons, however, are generally not point identified from observational data. For each individual, we observe only the potential outcome under the received treatment, so the two potential outcomes are never jointly observed (Rubin, 1974; Holland, 1986). Under standard causal assumptions, the observed data identify only the conditional *marginal* distributions of the two potential outcomes, leaving their conditional *joint* distribution unidentified (Zhang et al., 2013; Fan et al., 2017), which is particularly important for characterizing the probabilities of individual-level benefit and harm (Gadbury et al., 2004; Zhang et al., 2013; Fan et al., 2017; Kallus, 2022; Li et al., 2023; Ben-Michael et al., 2024; Cui & Han, 2025). Consequently, policy learning frameworks that account for individual-level benefit and harm must address the unidentified dependence between the two potential outcomes.

Existing approaches address this identification problem in two main ways. One approach leaves the dependence between the two potential outcomes unrestricted at each covariate value and evaluates policies under the corresponding worst-case joint distribution (Heckman et al., 1997; Fan & Park,

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.

2010; Fan et al., 2017; Kallus, 2022). This requires no structural assumptions on the dependence, but can be conservative when the marginal outcome distributions vary with covariates while the rank dependence remains similar across covariate values (Gijbels et al., 2015; Portier & Segers, 2018). In this setting, allowing a different least-favorable dependence at each covariate value fails to exploit the dependence shared across covariates. A second approach imposes additional restrictions on the joint distribution, for example through correlation models or parametric copulas (Gadbury et al., 2001; Shen et al., 2013; Zhang et al., 2013; Cai et al., 2022; Wu et al., 2024; He & Li, 2026). These restrictions can sharpen identification, but they specify features of the dependence between the two potential outcomes that are not identified from the observed data (Zhang et al., 2013). Together, these studies leave open how to exploit structure in the dependence between the two potential outcomes without imposing restrictive assumptions, which motivates the following question:

> *Can we exploit dependence shared across covariate values to learn less conservative policies, while leaving the dependence structure itself completely unspecified?*

To study this question, we formalize stable dependence through a common conditional copula across covariate values, while leaving the copula itself otherwise unrestricted. The conditional marginal distributions may vary freely with covariates, while the same rank dependence applies across them. This restriction requires the least-favorable dependence structure to be chosen jointly over the covariate distribution, instead of separately at each covariate value. We derive the resulting sharp maximin policy value, characterize when this restriction changes the optimal policy, and develop a finite-sample learner that directly optimizes the resulting objective without estimating the unidentified copula.

Our work makes the following **contributions**: (1) We formulate policy learning under copula invariance across covariates, allowing the common copula to remain unrestricted. (2) We derive the sharp maximin policy value and characterize when copula invariance changes the optimal policy relative to allowing the copula to vary freely across covariate values. (3) We develop a finite-sample policy learner with regret guarantees that does not require estimation of the unidentified copula. Experimental results demonstrate the practical benefits of exploiting copula invariance across covariates.

2 Related Work

Standard policy learning. Standard policy learning aims to assign treatments based on observed characteristics to maximize policy value, namely the average outcome induced by the treatment rule (Manski, 2004; Qian & Murphy, 2011; Kitagawa & Tetenov, 2018; Athey & Wager, 2021). Common approaches include outcome regression (OR), inverse propensity weighting (IPW), and doubly robust (DR) methods (Qian & Murphy, 2011; Swaminathan & Joachims, 2015; Kallus, 2018; Dudík et al., 2011; Athey & Wager, 2021), as well as empirical welfare maximization with finite-sample regret guarantees (Manski, 2004; Kitagawa & Tetenov, 2018). However, these methods compare policies through population-average outcomes. Their objectives depend on the marginal distributions of the potential outcomes, but not on how these outcomes are paired within individuals.

Policy learning for rank dependence. A growing literature instead studies policy objectives based on within-individual treatment comparisons, which depend on the unidentified joint distribution of potential outcomes. Partial identification methods derive sharp bounds over couplings compatible with the identified marginals (Heckman et al., 1997; Gadbury et al., 2004; Fan & Park, 2010; Fan et al., 2017; Kallus, 2022). Closest to our work, Ben-Michael et al. (2024) and Cui & Han (2025) use such bounds to learn minimax policies for counterfactual utility and distributional welfare. However, allowing the least-favorable coupling to vary freely across covariate values can be overly conservative. An alternative imposes structural restrictions on rank dependence, often through sensitivity analysis (Gadbury et al., 2001; Shen et al., 2013; Zhang et al., 2013; Cai et al., 2022; Wu et al., 2024; He & Li, 2026). These restrictions can sharpen inference but rely on unverifiable dependence specifications, and most focus on fixed treatment comparisons.

3 PROBLEM SETUP

Data and identification. We observe N independent and identically distributed observations $\mathcal{D}_N = \{O_i = (X_i, A_i, Y_i)\}_{i=1}^N$, where $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ denotes pretreatment covariates, $A \in \{0,1\}$ indicates whether treatment is received, and $Y \in \mathcal{Y} \subseteq \mathbb{R}$ is the observed outcome. We orient the outcome so that larger values are preferable. Let $Y(1)$ and $Y(0)$ denote the potential outcomes with and without treatment.

Assumption 1 (Standard causal assumptions)**.** *The observed data satisfy the following standard causal assumptions: (1)* ***Consistency:*** *$Y = AY(1) + (1-A)Y(0)$; (2)* ***Conditional ignorability:*** *$\{Y(0), Y(1)\} \perp A \mid X$; and (3)* ***Overlap:*** *there exists $\eta \in (0, 1/2)$ such that $\eta \le e(X) := \Pr(A = 1 \mid X) \le 1 - \eta$ almost surely.*

Under Assumption 1, the conditional marginal distributions of the potential outcomes are identified from the observed-data law as $F_a(y \mid x) := \Pr\{Y(a) \le y \mid X = x\} = \Pr(Y \le y \mid A = a, X = x)$ for $a \in \{0,1\}$. We further assume that $F_a(\cdot \mid x)$ is continuous and write $Q_a(u \mid x) := \inf\{y : F_a(y \mid x) \ge u\}$, $u \in (0,1)$, for the corresponding conditional quantile functions.

Assumption 2 (Bounded quantiles)**.** *Let $\widehat{Q}_a(u \mid x)$ denote an estimator of the conditional quantile function $Q_a(u \mid x)$, obtained from the nuisance-training sample. There exists a constant $B < \infty$ such that $|Q_a(u \mid x)| \le B$ and $|\widehat{Q}_a(u \mid x)| \le B$ for all $a \in \{0,1\}$, $u \in (0,1)$, and $x \in \mathcal{X}$. Estimated conditional quantiles may be clipped to this range.*

Common unrestricted copula. The conditional marginal distributions of the potential outcomes are identified, while their joint distribution remains unidentified. We define the conditional ranks $U_a = F_a\{Y(a) \mid X\}$ for $a \in \{0,1\}$. By continuity of the conditional marginal distributions, $U_a \mid X = x \sim \mathrm{Unif}(0,1)$. Let $C_x(u,v) = \Pr(U_0 \le u, U_1 \le v \mid X = x)$ denote the conditional copula of the two potential-outcome ranks. The copula C_x characterizes the conditional dependence between the ranks of $Y(0)$ and $Y(1)$. We say that this rank dependence is stable across covariate values when $C_x = C$ for P_X-almost every x. The common copula C is otherwise unrestricted. Let $\gamma \in \Pi(\lambda, \lambda)$ denote the copula measure induced by C, where λ is the uniform distribution on $[0,1]$. Then the potential outcomes admit the representation $Y(a) = Q_a(U_a \mid X)$ with $(U_0, U_1) \perp X$. Thus, the conditional marginals may vary with x, while rank dependence remains invariant without imposing further assumptions on the copula.

Policies and counterfactual utility. Let $\pi : \mathcal{X} \to [0,1]$ denote a treatment policy, where $\pi(x)$ is the probability of assigning treatment to an individual with covariates x. Fix a clinically meaningful improvement threshold $\delta \ge 0$ and a transition width $\kappa > 0$. We use the continuous improvement score $s_{\delta,\kappa}(t)$, which equals 0 when $t \le \delta$, increases linearly as $(t-\delta)/\kappa$ when $\delta < t < \delta + \kappa$, and equals 1 when $t \ge \delta + \kappa$. This score is a continuous surrogate for the indicator that the selected treatment option improves the outcome by more than δ. For covariates x and potential ranks (u,v), define the policy utility

$$c_{\pi,x}(u,v) = \pi(x) s_{\delta,\kappa}\{Q_1(v \mid x) - Q_0(u \mid x)\} + \{1 - \pi(x)\} s_{\delta,\kappa}\{Q_0(u \mid x) - Q_1(v \mid x)\}. \quad (1)$$

The first term rewards assigning treatment when receiving treatment yields a clinically meaningful improvement relative to not receiving treatment. The second rewards withholding treatment when not receiving treatment yields the clinically preferable outcome.

Objective. Because the common copula γ is not identified from the observed data, we evaluate policies using a maximin criterion. For a given policy π, we take the worst case over all dependence structures compatible with the shared-copula assumption and define

$$\underline{V}_{\mathrm{cc}}(\pi) := \inf_{\gamma \in \Pi(\lambda,\lambda)} \mathbb{E}_X \left[\int_{[0,1]^2} c_{\pi,X}(u,v)\, \mathrm{d}\gamma(u,v) \right], \quad (2)$$

The learned policy is then defined as $\hat{\pi} \in \arg\max_{\pi \in \mathcal{P}} \underline{V}_{\mathrm{cc}}(\pi)$.

4 SHARP IDENTIFICATION UNDER A SHARED COPULA

In this section, we study sharp policy identification under a shared unrestricted counterfactual copula. First, we characterize the sharp maximin value and its implications for policy choice. We then

develop a finite-sample learning procedure for the resulting objective. Finally, we study robustness to approximate copula invariance.

4.1 Shared-Copula Gain and Decision Relevance

Proposition 1 (Sharp identification under a shared copula)**.** *For any fixed policy π, under the shared-copula model, the identified set of its counterfactual value is $\mathcal{I}_{\mathrm{cc}}(\pi) = \big[\underline{V}_{\mathrm{cc}}(\pi), \overline{V}_{\mathrm{cc}}(\pi)\big]$, where*

$$\underline{V}_{\mathrm{cc}}(\pi) = \inf_{\gamma\in\Pi(\lambda,\lambda)} \int_{[0,1]^2} \bar{c}_\pi(u,v)\,\mathrm{d}\gamma(u,v), \quad \overline{V}_{\mathrm{cc}}(\pi) = \sup_{\gamma\in\Pi(\lambda,\lambda)} \int_{[0,1]^2} \bar{c}_\pi(u,v)\,\mathrm{d}\gamma(u,v), \tag{3}$$

and $\bar{c}_\pi(u,v) = \mathbb{E}_X\big[c_{\pi,X}(u,v)\big]$. Both endpoints are attained, and every value between them is attainable. Moreover, $\underline{V}_{\mathrm{cc}}(\pi) \geq \underline{V}_{\mathrm{rect}}(\pi)$. Here, $\underline{V}_{\mathrm{rect}}(\pi)$ is the rectangular worst-case value obtained by selecting the least-favorable copula separately for each covariate value x. The proof is deferred to Appendix A.1.

Proposition 1 shows that imposing $C_x = C$ for P_X-almost every x sharpens the worst-case policy value relative to the rectangular benchmark while leaving the common dependence structure otherwise unrestricted. However, a tighter value for a fixed policy does not by itself imply a different policy choice. To isolate the policy-specific effect of the restriction, we introduce the following quantity.

Definition 1 (Shared-copula gain)**.** *Let $X \sim P_X$. For a policy π and covariate value x, define $f_{\pi,x}(\gamma) = \int_{[0,1]^2} c_{\pi,x}(u,v)\,\mathrm{d}\gamma(u,v)$ for $\gamma \in \Pi(\lambda,\lambda)$, and let $m_\pi(x) = \inf_{\gamma\in\Pi(\lambda,\lambda)} f_{\pi,x}(\gamma)$ denote the pointwise rectangular worst-case value, so that $\underline{V}_{\mathrm{rect}}(\pi) = \mathbb{E}_X[m_\pi(X)]$. The shared-copula gain of policy π is defined as*

$$H(\pi) = \inf_{\gamma\in\Pi(\lambda,\lambda)} \mathbb{E}_X[f_{\pi,X}(\gamma) - m_\pi(X)]\,. \tag{4}$$

Intuitively, $H(\pi)$ measures the excess over pointwise worst-case values from requiring a common copula across covariates, capturing the policy-specific benefit of the shared-copula restriction. We next give a finite rank-block representation that makes this gain directly computable.

Proposition 2 (Finite rank-block characterization)**.** *Fix a policy π. Suppose that, for P_X-almost every x, the cost $c_{\pi,x}$ is constant on the cells of a common $K \times K$ equal-mass partition of the rank space $[0,1]^2$. Let $L_{\pi,x} \in \mathbb{R}^{K\times K}$ be the resulting cost matrix, let $\mathcal{B}_K$ denote the scaled Birkhoff polytope with row and column sums $1/K$, and let P_σ denote its permutation vertex associated with $\sigma \in S_K$. Then*

$$m_\pi(x) = \min_{\sigma\in S_K} \langle P_\sigma, L_{\pi,x}\rangle, \quad H(\pi) = \min_{\sigma\in S_K} \mathbb{E}_X[\langle P_\sigma, L_{\pi,X}\rangle - m_\pi(X)]\,, \quad P_X\text{-a.e. } x. \tag{5}$$

The proof is deferred to Appendix A.2.

Proposition 2 reduces the infinite-dimensional copula optimization to a finite assignment problem and makes the source of the shared-copula gain explicit: while the rectangular benchmark may select a different least-favorable permutation at each covariate value, the shared-copula model must use a single common permutation for P_X-almost every x. Thus, the gain arises precisely from incompatibility among the pointwise least-favorable dependence structures across covariate values. We next connect this policy-specific gain to policy choice and characterize when it is large enough to change the optimal policy.

Lemma 1 (Characterization of the shared-copula gain)**.** *Under the measurability and attainment conditions of Proposition 1, for any policy π, the shared-copula value satisfies $\underline{V}_{\mathrm{cc}}(\pi) = \underline{V}_{\mathrm{rect}}(\pi) + H(\pi)$, and hence $H(\pi) \geq 0$. Moreover, $H(\pi) = 0$ if and only if there exists $\gamma^\star_\pi \in \Pi(\lambda,\lambda)$ such that $\gamma^\star_\pi \in \arg\min_{\gamma\in\Pi(\lambda,\lambda)} f_{\pi,x}(\gamma)$ for P_X-almost every x. The proof is deferred to Appendix A.3.*

Lemma 1 shows that $H(\pi)$ is exactly the gap between the shared-copula and rectangular worst-case values. Together, Definition 1 and Lemma 1 reduce the decision question to how these policy-specific gains alter the rectangular ranking, which leads to the following theorem.

Theorem 1 (Policy-switch condition)**.** *Let $\pi_R \in \arg\max_{\pi\in\mathcal{P}} \underline{V}_{\mathrm{rect}}(\pi)$ be a rectangularly optimal policy. The shared-copula restriction induces a strict policy switch away from π_R if and only if there exists a policy $\pi \in \mathcal{P}$ such that*

$$H(\pi) - H(\pi_R) > \underline{V}_{\mathrm{rect}}(\pi_R) - \underline{V}_{\mathrm{rect}}(\pi). \tag{6}$$

The proof is deferred to Appendix A.4.

Theorem 1 characterizes when the shared-copula restriction changes policy choice. While $H(\pi)$ weakly raises each policy's worst-case value, a switch occurs only when gains differ sufficiently across policies. A competitor overturns the rectangular ranking when its gain advantage exceeds its rectangular value deficit, so decision relevance depends on relative gains.

4.2 Finite-Sample Learning under a Shared Copula

With the shared-copula objective characterized at the population level, we now turn to its finite-sample implementation. The main challenge is to learn the policy from observed data while the conditional outcome distributions must be estimated and the common copula remains unidentified. To this end, we develop the empirical procedure summarized in Algorithm 1, which directly targets the shared-copula maximin objective without estimating the copula itself.

Algorithm 1 Sharp maximin policy learning under a shared unrestricted copula

Input: Data $\mathcal{D} = \{(X_i, A_i, Y_i)\}_{i=1}^{N}$; policy class $\mathcal{P}$; rank resolution M; utility parameters (δ, κ); entropic regularization ε.

1: Split $\mathcal{D}$ into an independent nuisance sample $\mathcal{D}_Q$ of size n_q and evaluation sample $\mathcal{D}_E$ of size n, where $N = n_q + n$
2: Fit conditional quantile function $\widehat{Q}_a(u \mid x)$ on $\mathcal{D}_Q$
3: Set midpoint ranks $u_j \leftarrow (j - \frac{1}{2})/M$;
4: Set the scaled Birkhoff polytope $\mathcal{B}_M$ and uniform reference matrix U_M;
5: Initialize a parameterized policy $\pi_\theta \in \mathcal{P}$.
6: **for** $t = 0$ to $T - 1$ **do**
7: Compute $\widehat{C}_\pi(j, k)$ using Equation (1);
8: Compute $\widehat{V}_{M,\varepsilon}(\pi)$ using Equation (10);
9: Update the policy to increase $\widehat{V}_{M,\varepsilon}$
10: **end for**

Output: Policy $\widehat{\pi}$ and robust value $\widehat{V}_{M,\varepsilon}(\widehat{\pi})$.

We first split the data into a nuisance-training sample $\mathcal{D}_Q$ of size n_q and an evaluation sample $\mathcal{D}_E$ of size n, where $N = n_q + n$, and use $\mathcal{D}_Q$ to estimate the treatment-specific conditional quantile functions with a flexible learner. We approximate the common copula on the midpoint grid $u_j = (j - \frac{1}{2})/M$, $j = 1, \ldots, M$. Let P_{jk} denote the probability mass assigned to the rank cell indexed by (u_j, u_k). Since both potential-outcome ranks are uniformly distributed, P must have uniform row and column marginals, so $P \in \mathcal{B}_M := \left\{P \in \mathbb{R}_+^{M \times M} : P\mathbf{1} = \frac{1}{M}\mathbf{1},\ P^\top \mathbf{1} = \frac{1}{M}\mathbf{1}\right\}$. Thus, the optimization over an unrestricted common copula is reduced to an optimal-transport problem over $\mathcal{B}_M$. For a policy π, we evaluate the estimated utility on the same rank grid and average over the evaluation sample to obtain the empirical cost matrix $\widehat{C}_\pi$. The empirical shared-copula value is then obtained by optimizing a single coupling $P \in \mathcal{B}_M$ against $\widehat{C}_\pi$. Because both the nuisance estimates and the policy objective are evaluated on this common grid, the relevant nuisance error is the average conditional-quantile error over the grid points and evaluation covariates.

Definition 2 (Conditional-quantile estimation error)**.** *Let $u_j = (j - \frac{1}{2})/M$, $j = 1, \ldots, M$, denote the midpoint rank grid, and let $\widehat{Q}_a(u \mid x)$ be the estimated conditional quantile function for treatment a. We define the realized average conditional-quantile estimation error on the evaluation sample as*

$$\widehat{r}_{n,M} = \sum_{a=0}^{1} \frac{1}{M} \sum_{j=1}^{M} \frac{1}{n} \sum_{i=1}^{n} \left| \widehat{Q}_a(u_j \mid X_i) - Q_a(u_j \mid X_i) \right|. \tag{7}$$

For each evaluation covariate and rank pair, the algorithm evaluates the estimated policy utility by replacing the conditional quantiles in Equation (1) with their estimates. Averaging these utilities over the evaluation sample forms the empirical cost matrix $\widehat{C}_\pi$. The following lemma establishes stability of the objective value under cost-matrix perturbations.

Lemma 2 (Lipschitz continuity of the OT value)**.** *For any two cost matrices $C, C' \in \mathbb{R}^{M \times M}$,*

$$\left| \min_{P \in \mathcal{B}_M} \langle P, C \rangle - \min_{P \in \mathcal{B}_M} \langle P, C' \rangle \right| \le \|C - C'\|_\infty. \tag{8}$$

The proof is deferred to Appendix A.5.

Lemma 2 reduces the analysis of the empirical OT value to controlling perturbations of its cost matrix. We first control the sampling error from averaging the rank-indexed costs over the evaluation sample.

Lemma 3 (Rank-grid sampling complexity). *Let $\mathcal{F}_M = \{x \mapsto c_{\pi,x}(u_j, u_k) : \pi \in \mathcal{P},\ j,k \in [M]\}$. For each fixed rank pair (j,k), write $c_{\pi,x}(u_j,u_k) = b_{jk}(x) + \pi(x) d_{jk}(x)$, where $b_{jk}(x) \in [0,1]$ and $|d_{jk}(x)| \le 1$. Then there exists a universal constant $C > 0$ such that, with probability at least $1-\alpha$,*

$$\sup_{\pi\in\mathcal{P}} \sup_{j,k\in[M]} |(\mathbb{P}_n - P) c_{\pi,X}(u_j,u_k)| \le C \left[\mathfrak{R}_n(\mathcal{P}) + \sqrt{\frac{\log(M/\alpha)}{n}} \right]. \tag{9}$$

In particular, the additional term $\sqrt{\log M/n}$ accounts for uniform control over the M^2 rank pairs. The proof is deferred to Appendix A.6.

Under the shared-copula restriction, the same dependence structure must apply across all evaluation covariates. Accordingly, we first aggregate the estimated rank-level utilities over the evaluation sample into the empirical cost matrix $\widehat{C}_\pi$, and then optimize a single coupling against this matrix. With entropic regularization, the empirical robust value is

$$\widehat{V}_{M,\varepsilon}(\pi) = \min_{P\in\mathcal{B}_M} \left\{ \langle P, \widehat{C}_\pi \rangle + \varepsilon \,\mathrm{KL}(P \| U_M) \right\}, \tag{10}$$

where $U_M = M^{-2}\mathbf{1}\mathbf{1}^\top$ is the uniform product coupling. The impact of nuisance estimation is thus transmitted entirely through the estimated cost matrix $\widehat{C}_\pi$. Lemma 4 bounds this perturbation in terms of the average conditional-quantile estimation error.

Lemma 4 (Average quantile error). *Let $\widetilde{C}_\pi$ denote the empirical cost matrix constructed using the true conditional quantiles, and let $\widehat{C}_\pi$ denote its counterpart constructed using the estimated conditional quantiles. Then, for any policy $\pi \in \mathcal{P}$ and any $P \in \mathcal{B}_M$,*

$$\left| \left\langle P, \widehat{C}_\pi - \widetilde{C}_\pi \right\rangle \right| \le \frac{1}{\kappa} \sum_{a=0}^{1} \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_n \left| \widehat{Q}_a(u_j \mid X) - Q_a(u_j \mid X) \right|. \tag{11}$$

Consequently,

$$\sup_{\pi\in\mathcal{P}} \left| \widehat{V}_M(\pi) - \widetilde{V}_M(\pi) \right| \le \frac{\widehat{r}_{n,M}}{\kappa}, \tag{12}$$

where $\widehat{V}_M(\pi)$ and $\widetilde{V}_M(\pi)$ are the corresponding unregularized empirical OT values. The proof is deferred to Appendix A.7.

Having controlled the sampling and conditional-quantile estimation errors on the fixed rank grid, it remains to quantify the approximation error introduced by discretizing the continuous rank space.

Lemma 5 (Rank discretization error). *Under Assumption 2, let $C_{\pi,M}(j,k) = \mathbb{E}_X[c_{\pi,X}(u_j,u_k)]$ and define the population grid objective $V_M(\pi) = \min_{P\in\mathcal{B}_M} \langle P, C_{\pi,M}\rangle$. Then*

$$\sup_{\pi\in\mathcal{P}} |V_M(\pi) - \underline{V}_{\mathrm{cc}}(\pi)| \le \frac{2B}{\kappa M}. \tag{13}$$

In particular, this bound requires no Lipschitz continuity of the conditional quantile functions with respect to the rank variable. The proof is deferred to Appendix A.8.

Lemma 5 shows that the finite-grid objective approximates the continuous shared-copula value at rate $O(M^{-1})$, without requiring rank smoothness of the conditional quantile functions. The algorithm, however, solves an entropically regularized transport problem on this grid, so it remains to quantify the additional approximation error introduced by the regularization.

Lemma 6 (Entropic approximation error). *Let $U_M = M^{-2}\mathbf{1}\mathbf{1}^\top$ denote the uniform product coupling. For any cost matrix $C \in \mathbb{R}^{M\times M}$,*

$$0 \le \min_{P\in\mathcal{B}_M} \{\langle P, C\rangle + \varepsilon \mathrm{KL}(P\|U_M)\} - \min_{P\in\mathcal{B}_M} \langle P, C\rangle \le \varepsilon \log M. \tag{14}$$

The proof is deferred to Appendix A.9.

This lemma shows that the computational gain from entropic regularization comes at a uniformly controlled approximation cost of at most $\varepsilon \log M$. Finally, combining the preceding error controls yields a uniform finite-sample guarantee for the empirical shared-copula objective and the resulting learned policy.

Theorem 2 (Uniform finite-sample policy-value bound)**.** *Under Assumption 2, combining Lemmas 2–6 yields, with probability at least* $1-\alpha$*,*

$$\Delta_n := \sup_{\pi\in\mathcal{P}} \left|\widehat{V}_{M,\varepsilon}(\pi) - \underline{V}_{\mathrm{cc}}(\pi)\right| \le C\left[\mathfrak{R}_n(\mathcal{P}) + \sqrt{\frac{\log(M/\alpha)}{n}} + \frac{\widehat{r}_{n,M}}{\kappa} + \frac{2B}{\kappa M}\right] + \varepsilon \log M, \quad (15)$$

where $C>0$ *is a universal constant. Moreover, suppose that the learned policy* $\widehat{\pi}$ *satisfies* $\widehat{V}_{M,\varepsilon}(\widehat{\pi}) \ge \sup_{\pi\in\mathcal{P}} \widehat{V}_{M,\varepsilon}(\pi) - \xi_n$*. Then, for* $\pi^\star_{\mathrm{cc}} \in \arg\max_{\pi\in\mathcal{P}} \underline{V}_{\mathrm{cc}}(\pi)$*, the shared-copula policy regret satisfies* $\underline{V}_{\mathrm{cc}}(\pi^\star_{\mathrm{cc}}) - \underline{V}_{\mathrm{cc}}(\widehat{\pi}) \le 2\Delta_n + \xi_n$*. Under the additional rate conditions in Assumption 3, the policy regret further satisfies*

$$\underline{V}_{\mathrm{cc}}(\pi^\star_{\mathrm{cc}}) - \underline{V}_{\mathrm{cc}}(\widehat{\pi}) = O_p\left(\sqrt{\frac{\log|\mathcal{P}|}{n}} + \sqrt{\frac{\log n}{n}} + \frac{r_{n_q}}{\kappa}\right). \quad (16)$$

The proof is deferred to Appendix A.10.

Theorem 2 has two main implications. (1) It establishes that the sharp shared-copula maximin policy is learnable from finite data without estimating the unidentified copula itself, thereby connecting the population identification results to an implementable policy-learning procedure. (2) The conditional-quantile estimation enters only through the average rank error rather than a worst-case error over all M^2 rank pairs.

4.3 Sensitivity to Approximate Copula Invariance

The shared-copula model assumes exact stability of rank dependence across covariates. Since this structural restriction is not testable from standard observational data, we study robustness to approximate violations by allowing conditional copulas to deviate from their closest common copula in average Wasserstein distance. This leads to the following relaxation of exact copula invariance.

Definition 3 (Wasserstein-stable rank dependence)**.** *Let* $\gamma_x \in \Pi(\lambda,\lambda)$ *denote the conditional copula measure of the potential-outcome ranks given* $X=x$*. For* $\rho\ge 0$*, define the class of* ρ*-stable copula families*

$$\Gamma_\rho = \left\{\{\gamma_x\}_{x\in\mathcal{X}} : \inf_{\bar\gamma\in\Pi(\lambda,\lambda)} \mathbb{E}_X[W_1(\gamma_X,\bar\gamma)] \le \rho\right\},$$

where W_1 *denotes the 1-Wasserstein distance. The corresponding robust lower policy value is* $\underline{V}_\rho(\pi) = \inf_{\{\gamma_x\}\in\Gamma_\rho} \mathbb{E}_X\left[\int_{[0,1]^2} c_{\pi,X}(u,v)\,\mathrm{d}\gamma_X(u,v)\right]$*. When* $\rho=0$*,* $\underline{V}_0(\pi) = \underline{V}_{\mathrm{cc}}(\pi)$*.*

Corollary 1 (Sensitivity to approximate copula invariance)**.** *Under Assumption 4, for every policy* $\pi\in\mathcal{P}$*,*

$$0 \le \underline{V}_{\mathrm{cc}}(\pi) - \underline{V}_\rho(\pi) \le \frac{L_Q}{\kappa}\rho.$$

Moreover, let $\pi^\star_\rho \in \arg\max_{\pi\in\mathcal{P}} \underline{V}_\rho(\pi)$*. If* $\widehat{\pi}$ *is learned under the shared-copula objective,* $\underline{V}_\rho(\pi^\star_\rho) - \underline{V}_\rho(\widehat{\pi}) \le (L_Q/\kappa)\rho + 2\Delta_n + \xi_n$*. The proof is deferred to Appendix A.11.*

Corollary 1 quantifies the cost of relaxing exact copula invariance. The robust policy value deteriorates at most linearly with the deviation radius ρ, and the policy learned under the shared-copula model remains near-optimal relative to the corresponding ρ-stable robust oracle. Thus, moderate violations of exact invariance introduce only a controlled additional regret term.

5 Experiments

Datasets. We evaluate our method on both semi-synthetic and synthetic datasets to study its performance under controlled data-generating processes and realistic covariate distributions. For the semi-synthetic experiments, we use the ACIC 2016 benchmark (Dorie et al., 2019), developed for the 2016 Atlantic Causal Inference Conference. It contains real-world covariates from 4,802 individuals with 58 pretreatment variables. We retain these covariates and generate treatments and potential outcomes using our own data-generating process. For the fully synthetic experiments, all variables are generated from a known observational model with correlated Gaussian covariates, nonlinear treatment assignment, and controlled counterfactual structure. Further details on data generation and preprocessing are provided in Appendix B.1.

Experimental Details. We compare our method with the rectangular robust baseline (Cui & Han, 2025), which allows the least-favorable copula to vary across covariate values. We also consider standard policy-learning baselines based on mean outcomes. These include outcome regression (OR), inverse propensity weighting (IPW), and doubly robust (DR) learning. Unless otherwise specified, we use a rank resolution of $M = 50$, set $\delta = 0.25$ and $\kappa = 0.25$ for the improvement score, and apply entropic regularization with $\varepsilon = 0.02$.

5.1 Policy Learning under Copula Invariance

We use the ACIC 2016 covariates to examine the decision value of copula invariance across three scenarios indexed by $g \in \{1, 3, 5\}$. The parameter g controls how unevenly treatment outcomes are distributed within a covariate-defined subgroup: larger g makes poor outcomes worse and favorable outcomes better while keeping conditional average treatment effects fixed. This lets us test whether exploiting copula invariance improves decisions when average treatment effects alone cannot distinguish the scenarios. The true copula is the independence copula and remains common across covariates in every scenario. We evaluate a learned policy $\hat{\pi}$ using two metrics. Shared-copula robust regret is defined as $R^{\mathrm{cc}}(\hat{\pi}) = \underline{V}^{\mathrm{cc}}(\pi^{\star}_{\mathrm{cc}}) - \underline{V}^{\mathrm{cc}}(\hat{\pi})$, where $\underline{V}^{\mathrm{cc}}$ denotes the shared-copula worst-case utility and $\pi^{\star}_{\mathrm{cc}}$ is the corresponding population oracle policy; lower values are better. True policy utility is defined as $V^{\mathrm{true}}(\hat{\pi}) = \mathbb{E}[\hat{\pi}(X)s_{\delta,\kappa}\{Y(1) - Y(0)\} + \{1 - \hat{\pi}(X)\}s_{\delta,\kappa}\{Y(0) - Y(1)\}]$. For evaluation, we use the true conditional quantile functions to isolate policy performance from conditional-quantile estimation error. Together, they assess whether our learner both approaches the shared-copula robust optimum and improves actual decision utility beyond its own worst-case objective.

Table 1: Results for R^{cc} and V^{true} ($\times 10^{-2}$) in the semi-synthetic experiment. Results are means $\pm$ standard errors over 10 repetitions. Lower R^{cc} and higher V^{true} are better; the best result in each column is shown in bold. Improvement is computed as the best baseline minus ours for R^{cc}, and ours minus the best baseline for V^{true}, using unrounded values.

	$g = 1.0$		$g = 3.0$		$g = 5.0$	
	$R^{\mathrm{cc}} \downarrow$	$V^{\mathrm{true}} \uparrow$	$R^{\mathrm{cc}} \downarrow$	$V^{\mathrm{true}} \uparrow$	$R^{\mathrm{cc}} \downarrow$	$V^{\mathrm{true}} \uparrow$
OR	30.79 ± 0.10	46.14 ± 0.03	30.63 ± 0.10	44.94 ± 0.03	30.80 ± 0.10	43.81 ± 0.03
IPW	22.56 ± 0.71	45.96 ± 0.16	21.57 ± 0.63	45.66 ± 0.13	20.59 ± 0.88	45.60 ± 0.15
DR	23.73 ± 0.67	46.02 ± 0.11	21.59 ± 0.75	45.74 ± 0.10	20.46 ± 0.99	45.42 ± 0.13
Rectangular	27.25 ± 0.10	46.63 ± 0.03	26.02 ± 0.12	45.98 ± 0.03	22.73 ± 0.16	46.29 ± 0.05
Shared (ours)	$\mathbf{1.66 \pm 0.09}$	$\mathbf{50.78 \pm 0.03}$	$\mathbf{1.57 \pm 0.09}$	$\mathbf{52.14 \pm 0.03}$	$\mathbf{1.61 \pm 0.07}$	$\mathbf{53.47 \pm 0.03}$
Improvement	$+20.90$	$+4.16$	$+20.00$	$+6.16$	$+18.85$	$+7.18$

Table 1 shows that the shared-copula learner consistently outperforms all competing methods under both evaluation metrics. Its substantially lower shared-copula robust regret indicates that the proposed learner more closely approximates the oracle policy for the shared-copula objective. More importantly, it also achieves higher true policy utility, which is evaluated using the actual joint distribution of the potential outcomes. This shows that the advantage extends beyond the proposed robust objective and translates into better treatment decisions in terms of the underlying counterfactual utility of interest.

5.2 Finite-Sample Performance with Estimated Quantiles

We next examine how conditional quantile estimation affects finite-sample policy learning. We fix the data-generating process and vary the total training sample size N. For each N, half the observations are used to estimate $Q_a(u \mid x)$ and the other half for policy learning. We compare our shared-copula learner with the rectangular robust and doubly robust baselines, evaluating all methods on the same fixed test distribution using shared-copula robust regret.

Figure 1 evaluates our method in finite samples when the conditional quantile functions must be estimated from data. To additionally assess robustness to the choice of quantile estimator, we consider three different estimators, including non-crossing quantile regression (Zhou et al., 2020), simultaneous quantile regression (Tagasovska & Lopez-Paz, 2019), and neural spline flow (Durkan et al., 2019). Their implementation details are provided in Appendix B.2. As the training sample

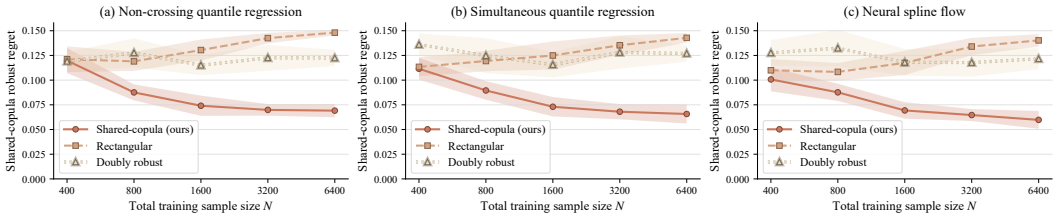

Figure 1: Finite-sample performance with estimated conditional quantiles. Lower shared-copula robust regret is better.

grows, the robust regret of the shared-copula learner decreases steadily, indicating that improved conditional-quantile estimation translates into better policy decisions. By contrast, the rectangular and doubly robust learners do not approach the shared-copula oracle with more data. As estimation error decreases, their learned policies increasingly reflect objectives that differ from $\underline{V}_{\mathrm{cc}}$, leaving a persistent shared-copula regret. Consequently, the advantage of our method becomes progressively clearer as more data are available.

5.3 Sensitivity to Copula Violations

To assess sensitivity to departures from exact copula invariance, we vary the Wasserstein radius ρ in the ambiguity set $\Gamma_\rho = \big\{\{\gamma_x\}_{x\in\mathcal{X}} : \inf_{\bar{\gamma}\in\Pi(\lambda,\lambda)} \mathbb{E}_X[W_1(\gamma_X, \bar{\gamma})] \le \rho\big\}$, where γ_x denotes the conditional copula at covariate value x. Thus, $\rho = 0$ recovers exact copula invariance, while larger ρ permits greater average deviation from a common copula. For each ρ, we evaluate the learned policies using the ρ-stable robust regret $\underline{V}_\rho(\pi^\star_\rho) - \underline{V}_\rho(\widehat{\pi})$. This directly examines how policy performance changes as the shared-copula restriction is progressively relaxed.

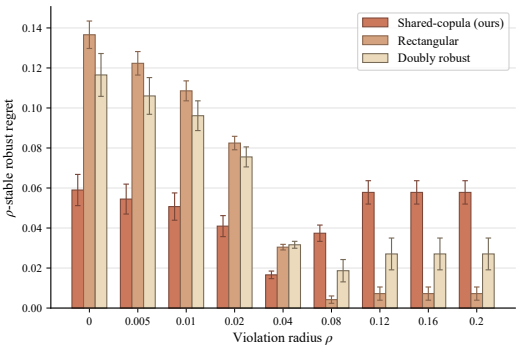

Figure 2: Sensitivity to violations of copula invariance. Lower ρ-stable robust regret is better.

Figure 2 shows that our method retains the lowest robust regret both under exact copula invariance and under moderate violations thereof. Its performance degrades only gradually as ρ increases, consistent with the controlled sensitivity predicted by Corollary 1. As larger values of ρ admit substantially more copula heterogeneity, the advantage of the shared-copula restriction diminishes and the rectangular learner becomes preferable. These results support the theoretical prediction that the proposed policy remains effective under moderate departures from exact copula invariance.

6 Conclusion

We study policy learning under copula invariance across covariates, allowing the common copula to remain otherwise unrestricted. This restriction yields a sharper maximin policy value than rectangular partial identification and can change the optimal policy when the resulting shared-copula gain is sufficiently large. We develop a finite-sample learner that directly optimizes the shared-copula objective without estimating the unidentified copula, and establish policy-value and regret guarantees. We further quantify robustness to approximate violations of copula invariance through a Wasserstein relaxation. Across semi-synthetic and synthetic experiments, the proposed method improves policy learning when rank dependence is informative and remains effective under moderate departures from exact invariance. Our current formulation is limited to binary treatment decisions, where counterfactual dependence can be represented by a bivariate copula and the robust objective admits a tractable OT formulation. Extending the framework to multiple or continuous treatments would require handling richer dependence structures and developing scalable multi-marginal learning procedures.

AI USE STATEMENT

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

REFERENCES

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

Eli Ben-Michael, Kosuke Imai, and Zhichao Jiang. Policy learning with asymmetric counterfactual utilities. *Journal of the American Statistical Association*, 119(548):3045–3058, 2024.

Mingyang Cai, Stef van Buuren, and Gerko Vink. How to relate potential outcomes: Estimating individual treatment effects under a given specified partial correlation, 2022.

Yifan Cui and Sukjin Han. Policy learning with distributional welfare. *Journal of the American Statistical Association*, pp. 1–12, 2025.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019. doi: 10.1214/18-STS667.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

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

Yanqin Fan, Emmanuel Guerre, and Dongming Zhu. Partial identification of functionals of the joint distribution of potential outcomes. *Journal of Econometrics*, 197(1):42–59, 2017.

Gary L. Gadbury, Hari K. Iyer, and David B. Allison. Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics*, 11(4):313–333, 2001.

Gary L. Gadbury, Hari K. Iyer, and Jeffrey M. Albert. Individual treatment effects in randomized trials with binary outcomes. *Journal of Statistical Planning and Inference*, 121(2):163–174, 2004.

Irène Gijbels, Marek Omelka, and Noël Veraverbeke. Estimation of a copula when a covariate affects only marginal distributions. *Scandinavian Journal of Statistics*, 42(4):1109–1126, 2015.

Peiyu He and Fan Li. Causal inference with ordinal outcomes: Copula-based identification, estimation and sensitivity analysis, 2026.

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

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396):945–960, 1986.

Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8895–8906. Curran Associates, Inc., 2018.

Nathan Kallus. What's the harm? sharp bounds on the fraction negatively affected by treatment. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15996–16009. Curran Associates, Inc., 2022.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu, and Peng Wu. Trustworthy policy learning under the counterfactual no-harm criterion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20575–20598. PMLR, 2023.

Kristin A. Linn, Eric B. Laber, and Leonard A. Stefanski. Interactive Q-learning for quantiles. *Journal of the American Statistical Association*, 112(518):638–649, 2017.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4): 1221–1246, 2004.

François Portier and Johan Segers. On the weak convergence of the empirical conditional copula under a simplifying assumption. *Journal of Multivariate Analysis*, 166:160–181, 2018.

Zhengling Qi, Jong-Shi Pang, and Yufeng Liu. On robustness of individualized decision rules. *Journal of the American Statistical Association*, 118(543):2143–2157, 2023.

Min Qian and Susan A. Murphy. Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210, 2011.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Changyu Shen, Jaesik Jeong, Xiaochun Li, Peng-Sheng Chen, and Alfred Buxton. Treatment benefit and treatment harm rate to characterize heterogeneity in treatment effect. *Biometrics*, 69(3): 724–731, 2013.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 814–823. PMLR, 2015.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *Advances in neural information processing systems*, 32, 2019.

Yuanjia Wang, Haoda Fu, and Donglin Zeng. Learning optimal personalized treatment rules in consideration of benefit and risk: With an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association*, 113(521):1–13, 2018.

Peng Wu, Peng Ding, Zhi Geng, and Yue Liu. Quantifying individual risk for binary outcome, 2024.

Zhiwei Zhang, Chenguang Wang, Lei Nie, and Guoxing Soon. Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(5):687–704, 2013.

Fan Zhou, Jianing Wang, and Xingdong Feng. Non-crossing quantile regression for distributional reinforcement learning. *Advances in neural information processing systems*, 33:15909–15919, 2020.

A Detailed Theoretical Proofs

A.1 Proof of Proposition 1

Proof. Fix a policy π, and define, for $\gamma \in \Pi(\lambda, \lambda)$,

$$\Phi_\pi(\gamma) := \mathbb{E}_X \left[\int_{[0,1]^2} c_{\pi,X}(u,v)\, \mathrm{d}\gamma(u,v) \right].$$

Since $0 \le c_{\pi,x} \le 1$, Fubini's theorem gives

$$\Phi_\pi(\gamma) = \int_{[0,1]^2} \bar{c}_\pi(u,v)\, \mathrm{d}\gamma(u,v), \qquad \bar{c}_\pi(u,v) = \mathbb{E}_X[c_{\pi,X}(u,v)].$$

We first show that the set of values compatible with the observed-data law and the shared-copula restriction is exactly

$$\{\Phi_\pi(\gamma) : \gamma \in \Pi(\lambda, \lambda)\}.$$

Consider any full-data law compatible with the maintained model. By continuity of the conditional marginals,

$$U_a = F_a\{Y(a) \mid X\} \mid X = x \sim \mathrm{Unif}(0,1), \qquad a \in \{0,1\},$$

and the shared-copula restriction implies that the conditional law of (U_0, U_1) given $X = x$ equals the same $\gamma \in \Pi(\lambda, \lambda)$ for P_X-almost every x. Hence $(U_0, U_1) \perp X$. Moreover, the usual quantile representation gives

$$Y(a) = Q_a(U_a \mid X) \qquad \text{a.s.},$$

so the policy value under this full-data law is

$$\mathbb{E}[c_{\pi,X}(U_0, U_1)] = \Phi_\pi(\gamma).$$

Thus every admissible full-data law induces a value of the form $\Phi_\pi(\gamma)$ for some $\gamma \in \Pi(\lambda, \lambda)$.

Conversely, fix an arbitrary $\gamma \in \Pi(\lambda, \lambda)$. Draw

$$X \sim P_X, \qquad (U_0, U_1) \sim \gamma, \qquad (U_0, U_1) \perp X,$$

and define

$$Y(a) = Q_a(U_a \mid X), \qquad a \in \{0,1\}.$$

Because each marginal of γ is uniform, $Y(a) \mid X = x$ has conditional distribution $F_a(\cdot \mid x)$. Continuity of $F_a(\cdot \mid x)$ further implies $F_a\{Y(a) \mid X\} = U_a$ almost surely, up to rank values of Lebesgue measure zero. Therefore the conditional copula of the potential-outcome ranks is precisely γ for P_X-almost every x.

To complete the construction, generate

$$A \mid X \sim \mathrm{Bernoulli}\{e(X)\}$$

independently of (U_0, U_1) conditional on X, and set

$$Y = AY(1) + (1 - A)Y(0).$$

Then conditional ignorability holds by construction, while

$$Y \mid (A = a, X = x) \sim F_a(\cdot \mid x).$$

Hence the resulting law of (X, A, Y) coincides with the observed-data law identified under Assumption 1. Thus every $\gamma \in \Pi(\lambda, \lambda)$ is observationally compatible with the maintained shared-copula model and attains the value $\Phi_\pi(\gamma)$. It follows that

$$\mathcal{I}_{\mathrm{cc}}(\pi) = \{\Phi_\pi(\gamma) : \gamma \in \Pi(\lambda, \lambda)\}.$$

We next establish that the extrema over $\Pi(\lambda, \lambda)$ are attained. The set $\Pi(\lambda, \lambda)$ is weakly compact: the underlying space $[0,1]^2$ is compact, and the requirement that both marginals equal λ defines a weakly closed subset of the space of probability measures on $[0,1]^2$.

It remains to verify continuity of Φ_π under weak convergence. Suppose $\gamma_m \Rightarrow \gamma$ with $\gamma_m, \gamma \in \Pi(\lambda, \lambda)$. For a fixed x, each map $u \mapsto Q_0(u \mid x)$ and $v \mapsto Q_1(v \mid x)$ is monotone and therefore has at most countably many discontinuities. Let $D_{0,x}$ and $D_{1,x}$ denote the corresponding discontinuity sets. Since $s_{\delta,\kappa}$ is continuous, the discontinuity set of $c_{\pi,x}$ is contained in

$$\big(D_{0,x} \times [0,1]\big) \cup \big([0,1] \times D_{1,x}\big).$$

Every $\gamma \in \Pi(\lambda, \lambda)$ has uniform marginals, and hence

$$\gamma(D_{0,x} \times [0,1]) = \lambda(D_{0,x}) = 0, \qquad \gamma([0,1] \times D_{1,x}) = \lambda(D_{1,x}) = 0.$$

Thus $c_{\pi,x}$ is γ-almost everywhere continuous. Because it is also bounded, weak convergence implies

$$\int c_{\pi,x}\,\mathrm{d}\gamma_m \longrightarrow \int c_{\pi,x}\,\mathrm{d}\gamma$$

for P_X-almost every x. Since all these integrals lie in $[0,1]$, dominated convergence yields

$$\Phi_\pi(\gamma_m) \longrightarrow \Phi_\pi(\gamma).$$

Therefore Φ_π is continuous on the weakly compact set $\Pi(\lambda,\lambda)$, so there exist $\gamma_\pi^-, \gamma_\pi^+ \in \Pi(\lambda,\lambda)$ such that

$$\Phi_\pi(\gamma_\pi^-) = \min_{\gamma\in\Pi(\lambda,\lambda)} \Phi_\pi(\gamma) = \underline{V}_{\mathrm{cc}}(\pi),$$

and

$$\Phi_\pi(\gamma_\pi^+) = \max_{\gamma\in\Pi(\lambda,\lambda)} \Phi_\pi(\gamma) = \overline{V}_{\mathrm{cc}}(\pi).$$

This proves attainment of both endpoints.

Finally, $\Pi(\lambda,\lambda)$ is convex and Φ_π is affine in γ. Hence, for any $t \in [0,1]$,

$$\gamma_t = (1-t)\gamma_\pi^- + t\gamma_\pi^+ \in \Pi(\lambda,\lambda)$$

and

$$\Phi_\pi(\gamma_t) = (1-t)\underline{V}_{\mathrm{cc}}(\pi) + t\overline{V}_{\mathrm{cc}}(\pi).$$

Thus every value between the two endpoints is attainable, and therefore

$$\mathcal{I}_{\mathrm{cc}}(\pi) = \left[\underline{V}_{\mathrm{cc}}(\pi), \overline{V}_{\mathrm{cc}}(\pi)\right].$$

It remains to compare the shared-copula lower value with the rectangular benchmark. For

$$f_{\pi,x}(\gamma) = \int_{[0,1]^2} c_{\pi,x}(u,v)\,\mathrm{d}\gamma(u,v),$$

we have

$$\begin{aligned}\underline{V}_{\mathrm{cc}}(\pi) &= \inf_{\gamma\in\Pi(\lambda,\lambda)} \mathbb{E}_X[f_{\pi,X}(\gamma)] \\ &\geq \mathbb{E}_X\left[\inf_{\gamma\in\Pi(\lambda,\lambda)} f_{\pi,X}(\gamma)\right] \\ &= \underline{V}_{\mathrm{rect}}(\pi),\end{aligned}$$

where the inequality follows because, for every fixed $\gamma \in \Pi(\lambda,\lambda)$,

$$f_{\pi,x}(\gamma) \geq \inf_{\widetilde{\gamma}\in\Pi(\lambda,\lambda)} f_{\pi,x}(\widetilde{\gamma}) \qquad P_X\text{-a.e. } x.$$

The rectangular model may choose a different least-favorable copula at each covariate value, whereas the shared-copula model must use the same γ throughout. This proves the claimed comparison and completes the proof. □

A.2 Proof of Proposition 2

Proof. Fix a policy π. Let $I_1, \ldots, I_K$ denote the common equal-mass partition of $[0,1]$, so that $\lambda(I_r) = 1/K$ for every $r \in [K]$, and write $L_{\pi,x}(r,s)$ for the constant value of $c_{\pi,x}$ on $I_r \times I_s$.

We first show that optimization over copulas is exactly equivalent to optimization over the scaled Birkhoff polytope. For any $\gamma \in \Pi(\lambda,\lambda)$, define its block-mass matrix $P^\gamma \in \mathbb{R}^{K\times K}$ by

$$P^\gamma_{rs} = \gamma(I_r \times I_s), \qquad r,s \in [K].$$

Because both marginals of γ equal λ,

$$\sum_{s=1}^{K} P^{\gamma}_{rs} = \gamma(I_r \times [0,1]) = \lambda(I_r) = \frac{1}{K},$$

and similarly

$$\sum_{r=1}^{K} P^{\gamma}_{rs} = \gamma([0,1] \times I_s) = \lambda(I_s) = \frac{1}{K}.$$

Hence $P^{\gamma} \in \mathcal{B}_K$.

Conversely, every $P \in \mathcal{B}_K$ can be induced by a copula. In particular, define a probability measure γ_P on $[0,1]^2$ by distributing the mass P_{rs} uniformly over each cell $I_r \times I_s$, that is,

$$\gamma_P(\,\mathrm{d}u,\,\mathrm{d}v) = \sum_{r=1}^{K}\sum_{s=1}^{K} K^2 P_{rs}\mathbf{1}\{u \in I_r,\ v \in I_s\}\,\mathrm{d}u\,\mathrm{d}v.$$

For $u \in I_r$, its first marginal density is

$$\sum_{s=1}^{K} K^2 P_{rs}\lambda(I_s) = K\sum_{s=1}^{K} P_{rs} = 1,$$

and the same calculation gives a unit density for the second marginal. Thus $\gamma_P \in \Pi(\lambda,\lambda)$, and by construction

$$\gamma_P(I_r \times I_s) = P_{rs}.$$

Therefore the set of block-mass matrices induced by copulas is exactly $\mathcal{B}_K$.

Since $c_{\pi,x}$ is constant on each rank cell, for any $\gamma \in \Pi(\lambda,\lambda)$,

$$\begin{aligned} f_{\pi,x}(\gamma) &= \int_{[0,1]^2} c_{\pi,x}(u,v)\,\mathrm{d}\gamma(u,v) \\ &= \sum_{r=1}^{K}\sum_{s=1}^{K} L_{\pi,x}(r,s)\gamma(I_r \times I_s) \\ &= \langle P^{\gamma}, L_{\pi,x}\rangle\,. \end{aligned}$$

Consequently, for P_X-almost every x,

$$m_\pi(x) = \inf_{\gamma\in\Pi(\lambda,\lambda)} f_{\pi,x}(\gamma) = \min_{P\in\mathcal{B}_K} \langle P, L_{\pi,x}\rangle.$$

The scaled Birkhoff polytope is the convex hull of its scaled permutation vertices $\{P_\sigma : \sigma \in S_K\}$. Since the objective is linear in P, the Birkhoff–von Neumann theorem therefore gives

$$m_\pi(x) = \min_{\sigma\in S_K} \langle P_\sigma, L_{\pi,x}\rangle, \qquad P_X\text{-a.e. } x.$$

It remains to characterize the shared-copula gain. By Definition 1 and the exact correspondence established above,

$$\begin{aligned} H(\pi) &= \inf_{\gamma\in\Pi(\lambda,\lambda)} \mathbb{E}_X\left[f_{\pi,X}(\gamma) - m_\pi(X)\right] \\ &= \min_{P\in\mathcal{B}_K} \mathbb{E}_X\left[\langle P, L_{\pi,X}\rangle - m_\pi(X)\right]. \end{aligned}$$

Here the same matrix P must be used for every covariate value because it is induced by a single common copula. Moreover,

$$P \longmapsto \mathbb{E}_X\left[\langle P, L_{\pi,X}\rangle - m_\pi(X)\right]$$

is linear in P. Hence its minimum over $\mathcal{B}_K$ is again attained at a permutation vertex. Therefore

$$H(\pi) = \min_{\sigma\in S_K} \mathbb{E}_X\left[\langle P_\sigma, L_{\pi,X}\rangle - m_\pi(X)\right].$$

□

A.3 Proof of Lemma 1

Proof. Fix a policy π. Recall that

$$f_{\pi,x}(\gamma) = \int_{[0,1]^2} c_{\pi,x}(u,v)\,\mathrm{d}\gamma(u,v), \qquad m_\pi(x) = \inf_{\widetilde{\gamma}\in\Pi(\lambda,\lambda)} f_{\pi,x}(\widetilde{\gamma}),$$

so that

$$\underline{V}_{\mathrm{rect}}(\pi) = \mathbb{E}_X[m_\pi(X)].$$

By definition of the shared-copula value,

$$\underline{V}_{\mathrm{cc}}(\pi) = \inf_{\gamma\in\Pi(\lambda,\lambda)} \mathbb{E}_X[f_{\pi,X}(\gamma)].$$

Since $m_\pi(X)$ does not depend on the common copula γ, we may subtract and add its expectation inside the optimization to obtain

$$\begin{aligned}\underline{V}_{\mathrm{cc}}(\pi) &= \inf_{\gamma\in\Pi(\lambda,\lambda)} \big\{\mathbb{E}_X[m_\pi(X)] + \mathbb{E}_X\big[f_{\pi,X}(\gamma) - m_\pi(X)\big]\big\} \\ &= \underline{V}_{\mathrm{rect}}(\pi) + \inf_{\gamma\in\Pi(\lambda,\lambda)} \mathbb{E}_X\big[f_{\pi,X}(\gamma) - m_\pi(X)\big] \\ &= \underline{V}_{\mathrm{rect}}(\pi) + H(\pi).\end{aligned}$$

This proves the claimed decomposition.

For every x and every $\gamma \in \Pi(\lambda,\lambda)$, the definition of $m_\pi(x)$ gives

$$f_{\pi,x}(\gamma) - m_\pi(x) \geq 0.$$

Therefore every feasible value in the optimization defining $H(\pi)$ is nonnegative, and hence

$$H(\pi) \geq 0.$$

It remains to characterize the zero-gain case. First suppose that there exists a copula $\gamma_\pi^\star \in \Pi(\lambda,\lambda)$ such that

$$\gamma_\pi^\star \in \operatorname*{arg\,min}_{\gamma\in\Pi(\lambda,\lambda)} f_{\pi,x}(\gamma) \qquad \text{for } P_X\text{-almost every } x.$$

Then

$$f_{\pi,x}(\gamma_\pi^\star) = m_\pi(x) \qquad \text{for } P_X\text{-almost every } x,$$

and consequently

$$\mathbb{E}_X\big[f_{\pi,X}(\gamma_\pi^\star) - m_\pi(X)\big] = 0.$$

Since $H(\pi) \geq 0$, evaluating its defining infimum at $\gamma_\pi^\star$ yields $H(\pi) = 0$.

Conversely, suppose that $H(\pi) = 0$. Under the attainment conditions stated in Proposition 1, the shared-copula infimum is attained by some $\gamma_\pi^\star \in \Pi(\lambda,\lambda)$. Because the term $\mathbb{E}_X[m_\pi(X)]$ is constant in γ, the same copula also attains the infimum in the definition of $H(\pi)$. Thus

$$0 = H(\pi) = \mathbb{E}_X\big[f_{\pi,X}(\gamma_\pi^\star) - m_\pi(X)\big].$$

The integrand is measurable and nonnegative almost surely. A nonnegative random variable with expectation zero must vanish almost surely, so

$$f_{\pi,x}(\gamma_\pi^\star) = m_\pi(x) \qquad \text{for } P_X\text{-almost every } x.$$

Equivalently,

$$\gamma_\pi^\star \in \operatorname*{arg\,min}_{\gamma\in\Pi(\lambda,\lambda)} f_{\pi,x}(\gamma) \qquad \text{for } P_X\text{-almost every } x.$$

This proves the equivalence and completes the proof. □

A.4 Proof of Theorem 1

Proof. Let

$$\pi_R \in \arg\max_{\pi \in \mathcal{P}} \underline{V}_{\mathrm{rect}}(\pi)$$

be a rectangularly optimal policy. By Lemma 1, for every policy $\pi \in \mathcal{P}$,

$$\underline{V}_{\mathrm{cc}}(\pi) = \underline{V}_{\mathrm{rect}}(\pi) + H(\pi).$$

The shared-copula restriction induces a strict policy switch away from π_R if and only if π_R is not optimal under the shared-copula objective. Equivalently, this occurs if and only if there exists some $\pi \in \mathcal{P}$ such that

$$\underline{V}_{\mathrm{cc}}(\pi) > \underline{V}_{\mathrm{cc}}(\pi_R).$$

Using the decomposition above, this inequality is equivalent to

$$\underline{V}_{\mathrm{rect}}(\pi) + H(\pi) > \underline{V}_{\mathrm{rect}}(\pi_R) + H(\pi_R).$$

Rearranging terms gives

$$H(\pi) - H(\pi_R) > \underline{V}_{\mathrm{rect}}(\pi_R) - \underline{V}_{\mathrm{rect}}(\pi),$$

which is exactly Equation (6).

Conversely, suppose there exists a policy $\pi \in \mathcal{P}$ satisfying

$$H(\pi) - H(\pi_R) > \underline{V}_{\mathrm{rect}}(\pi_R) - \underline{V}_{\mathrm{rect}}(\pi).$$

Rearranging this inequality yields

$$\underline{V}_{\mathrm{rect}}(\pi) + H(\pi) > \underline{V}_{\mathrm{rect}}(\pi_R) + H(\pi_R).$$

Applying Lemma 1 once more gives

$$\underline{V}_{\mathrm{cc}}(\pi) > \underline{V}_{\mathrm{cc}}(\pi_R).$$

Hence π_R cannot be optimal under the shared-copula objective, and the shared-copula restriction induces a strict policy switch away from π_R.

Therefore, a strict policy switch occurs if and only if there exists $\pi \in \mathcal{P}$ such that

$$H(\pi) - H(\pi_R) > \underline{V}_{\mathrm{rect}}(\pi_R) - \underline{V}_{\mathrm{rect}}(\pi),$$

as claimed. □

A.5 Proof of Lemma 2

Proof. Define

$$V(C) := \min_{P \in \mathcal{B}_M} \langle P, C \rangle, \qquad V(C') := \min_{P \in \mathcal{B}_M} \langle P, C' \rangle.$$

Because $\mathcal{B}_M$ is a nonempty compact polytope and the objective is linear in P, both minima are attained.

Let

$$P' \in \arg\min_{P \in \mathcal{B}_M} \langle P, C' \rangle.$$

Using P' as a feasible candidate for the optimization defining $V(C)$, we obtain

$$V(C) \le \langle P', C \rangle.$$

Therefore,

$$\begin{aligned} V(C) - V(C') &\le \langle P', C \rangle - \langle P', C' \rangle \\ &= \langle P', C - C' \rangle. \end{aligned}$$

Since $P' \in \mathcal{B}_M$, its entries are nonnegative and its row sums are equal to $1/M$. Hence

$$\sum_{j=1}^{M} \sum_{k=1}^{M} P'_{jk} = \sum_{j=1}^{M} \frac{1}{M} = 1.$$

It follows that

$$\begin{aligned}\langle P', C - C' \rangle &\leq |\langle P', C - C' \rangle| \\ &\leq \sum_{j=1}^{M} \sum_{k=1}^{M} P'_{jk} \, |C_{jk} - C'_{jk}| \\ &\leq \|C - C'\|_\infty \sum_{j=1}^{M} \sum_{k=1}^{M} P'_{jk} \\ &= \|C - C'\|_\infty.\end{aligned}$$

Thus

$$V(C) - V(C') \leq \|C - C'\|_\infty.$$

Repeating the same argument with the roles of C and C' reversed gives

$$V(C') - V(C) \leq \|C - C'\|_\infty.$$

Combining the two inequalities yields

$$|V(C) - V(C')| \leq \|C - C'\|_\infty.$$

Equivalently,

$$\left| \min_{P \in \mathcal{B}_M} \langle P, C \rangle - \min_{P \in \mathcal{B}_M} \langle P, C' \rangle \right| \leq \|C - C'\|_\infty,$$

□

A.6 Proof of Lemma 3

Proof. For each fixed rank pair $(j, k) \in [M]^2$, define

$$\mathcal{F}_{jk} = \{x \mapsto c_{\pi,x}(u_j, u_k) : \pi \in \mathcal{P}\}.$$

By the definition of the policy utility in Equation (1), we can write

$$c_{\pi,x}(u_j, u_k) = b_{jk}(x) + \pi(x) d_{jk}(x),$$

where

$$b_{jk}(x) = s_{\delta,\kappa}\{Q_0(u_j \mid x) - Q_1(u_k \mid x)\}$$

and

$$d_{jk}(x) = s_{\delta,\kappa}\{Q_1(u_k \mid x) - Q_0(u_j \mid x)\} - s_{\delta,\kappa}\{Q_0(u_j \mid x) - Q_1(u_k \mid x)\}.$$

Since $s_{\delta,\kappa}$ takes values in $[0, 1]$, we have

$$0 \leq b_{jk}(x) \leq 1, \qquad |d_{jk}(x)| \leq 1.$$

In particular, every function in $\mathcal{F}_{jk}$ is bounded in $[0, 1]$.

We first control the complexity of $\mathcal{F}_{jk}$ for a fixed (j, k). Conditional on the observed covariates $X_1, \ldots, X_n$, its empirical Rademacher complexity is

$$\begin{aligned}\widehat{\mathfrak{R}}_n(\mathcal{F}_{jk}) &= \mathbb{E}_\sigma \left[\sup_{\pi \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \{b_{jk}(X_i) + \pi(X_i) d_{jk}(X_i)\} \right] \\ &= \mathbb{E}_\sigma \left[\sup_{\pi \in \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \pi(X_i) d_{jk}(X_i) \right].\end{aligned}$$

Indeed, the term involving b_{jk} does not depend on π and has zero expectation with respect to the Rademacher variables.

For each i, the map

$$t \longmapsto d_{jk}(X_i) t$$

is 1-Lipschitz and vanishes at zero because $|d_{jk}(X_i)| \leq 1$. Therefore, by the Rademacher contraction inequality,

$$\widehat{\mathfrak{R}}_n(\mathcal{F}_{jk}) \leq \widehat{\mathfrak{R}}_n(\mathcal{P}).$$

Taking expectation over the sample gives

$$\mathfrak{R}_n(\mathcal{F}_{jk}) \le \mathfrak{R}_n(\mathcal{P}).$$

Since the functions in $\mathcal{F}_{jk}$ are uniformly bounded in $[0,1]$, the standard Rademacher uniform-deviation inequality implies that, for any $\alpha_{jk} \in (0,1)$, with probability at least $1-\alpha_{jk}$,

$$\sup_{\pi\in\mathcal{P}} |(\mathbb{P}_n - P)c_{\pi,X}(u_j,u_k)| \le C_0 \left[\mathfrak{R}_n(\mathcal{P}) + \sqrt{\frac{\log(1/\alpha_{jk})}{n}}\right],$$

where $C_0 > 0$ is a universal constant.

It remains to obtain simultaneous control over the rank grid. There are M^2 possible rank pairs (j,k). Applying the preceding bound with

$$\alpha_{jk} = \frac{\alpha}{M^2}$$

and taking a union bound over all $(j,k) \in [M]^2$ gives, with probability at least $1-\alpha$,

$$\sup_{\pi\in\mathcal{P}} \sup_{j,k\in[M]} |(\mathbb{P}_n - P)c_{\pi,X}(u_j,u_k)| \le C_0 \left[\mathfrak{R}_n(\mathcal{P}) + \sqrt{\frac{\log(M^2/\alpha)}{n}}\right].$$

Since $\alpha \in (0,1)$ and $M \ge 1$,

$$\log(M^2/\alpha) = 2\log M + \log(1/\alpha) \le 2\log(M/\alpha).$$

Absorbing the resulting numerical factor into a universal constant $C > 0$, we obtain

$$\sup_{\pi\in\mathcal{P}} \sup_{j,k\in[M]} |(\mathbb{P}_n - P)c_{\pi,X}(u_j,u_k)| \le C \left[\mathfrak{R}_n(\mathcal{P}) + \sqrt{\frac{\log(M/\alpha)}{n}}\right].$$

This proves Equation (9).

The additional $\sqrt{\log M/n}$ term arises precisely from requiring uniform control over the M^2 rank pairs in the discretized rank space. □

A.7 Proof of Lemma 4

Proof. Fix a policy $\pi \in \mathcal{P}$. For notational convenience, define

$$\Delta_{a,j}(x) := \left|\widehat{Q}_a(u_j \mid x) - Q_a(u_j \mid x)\right|, \qquad a \in \{0,1\},\ j \in [M].$$

Recall that $s_{\delta,\kappa}$ is $1/\kappa$-Lipschitz. Hence, for every $x \in \mathcal{X}$ and every rank pair (u_j,u_k),

$$\left|s_{\delta,\kappa}\left\{\widehat{Q}_1(u_k \mid x) - \widehat{Q}_0(u_j \mid x)\right\} - s_{\delta,\kappa}\left\{Q_1(u_k \mid x) - Q_0(u_j \mid x)\right\}\right| \le \frac{1}{\kappa}\left\{\Delta_{0,j}(x) + \Delta_{1,k}(x)\right\}.$$

The same inequality holds for the control utility,

$$s_{\delta,\kappa}\left\{Q_0(u_j \mid x) - Q_1(u_k \mid x)\right\}.$$

Since $\pi(x) \in [0,1]$ and the policy utility is a convex combination of the treatment and control utilities, it follows that

$$|\widehat{c}_{\pi,x}(u_j,u_k) - c_{\pi,x}(u_j,u_k)| \le \frac{1}{\kappa}\left\{\Delta_{0,j}(x) + \Delta_{1,k}(x)\right\}.$$

Averaging over the evaluation sample gives

$$\left|\widehat{C}_\pi(j,k) - \widetilde{C}_\pi(j,k)\right| \le \frac{1}{\kappa}\left\{\mathbb{P}_n\Delta_{0,j}(X) + \mathbb{P}_n\Delta_{1,k}(X)\right\}.$$

Now fix any $P \in \mathcal{B}_M$. Using nonnegativity of the entries of P,

$$\begin{aligned}\left|\left\langle P, \widehat{C}_\pi - \widetilde{C}_\pi \right\rangle\right| &\le \sum_{j=1}^{M}\sum_{k=1}^{M} P_{jk}\left|\widehat{C}_\pi(j,k) - \widetilde{C}_\pi(j,k)\right| \\ &\le \frac{1}{\kappa}\sum_{j=1}^{M}\sum_{k=1}^{M} P_{jk}\left\{\mathbb{P}_n\Delta_{0,j}(X) + \mathbb{P}_n\Delta_{1,k}(X)\right\}.\end{aligned}$$

Because $P \in \mathcal{B}_M$, its row and column sums satisfy

$$\sum_{k=1}^{M} P_{jk} = \frac{1}{M}, \qquad \sum_{j=1}^{M} P_{jk} = \frac{1}{M}.$$

Therefore,

$$\begin{aligned}\left|\left\langle P, \widehat{C}_\pi - \widetilde{C}_\pi \right\rangle\right| &\le \frac{1}{\kappa}\left[\sum_{j=1}^{M}\left(\sum_{k=1}^{M} P_{jk}\right)\mathbb{P}_n\Delta_{0,j}(X)\right. \\ &\qquad \left. + \sum_{k=1}^{M}\left(\sum_{j=1}^{M} P_{jk}\right)\mathbb{P}_n\Delta_{1,k}(X)\right] \\ &= \frac{1}{\kappa}\sum_{a=0}^{1}\frac{1}{M}\sum_{j=1}^{M}\mathbb{P}_n\left|\widehat{Q}_a(u_j \mid X) - Q_a(u_j \mid X)\right|.\end{aligned}$$

This proves Equation (11).

It remains to translate the cost perturbation bound into a bound on the optimal transport values. Define

$$\widehat{V}_M(\pi) = \min_{P\in\mathcal{B}_M}\left\langle P, \widehat{C}_\pi\right\rangle, \qquad \widetilde{V}_M(\pi) = \min_{P\in\mathcal{B}_M}\left\langle P, \widetilde{C}_\pi\right\rangle.$$

Let

$$\widetilde{P}_\pi \in \arg\min_{P\in\mathcal{B}_M}\left\langle P, \widetilde{C}_\pi\right\rangle.$$

Using $\widetilde{P}_\pi$ as a feasible candidate in the optimization defining $\widehat{V}_M(\pi)$ gives

$$\begin{aligned}\widehat{V}_M(\pi) - \widetilde{V}_M(\pi) &\le \left\langle \widetilde{P}_\pi, \widehat{C}_\pi\right\rangle - \left\langle \widetilde{P}_\pi, \widetilde{C}_\pi\right\rangle \\ &= \left\langle \widetilde{P}_\pi, \widehat{C}_\pi - \widetilde{C}_\pi\right\rangle \\ &\le \frac{\widehat{r}_{n,M}}{\kappa}.\end{aligned}$$

Similarly, if

$$\widehat{P}_\pi \in \arg\min_{P\in\mathcal{B}_M}\left\langle P, \widehat{C}_\pi\right\rangle,$$

then

$$\begin{aligned}\widetilde{V}_M(\pi) - \widehat{V}_M(\pi) &\le \left\langle \widehat{P}_\pi, \widetilde{C}_\pi - \widehat{C}_\pi\right\rangle \\ &\le \frac{\widehat{r}_{n,M}}{\kappa}.\end{aligned}$$

Combining the two inequalities yields

$$\left|\widehat{V}_M(\pi) - \widetilde{V}_M(\pi)\right| \le \frac{\widehat{r}_{n,M}}{\kappa}.$$

Since the right-hand side does not depend on π, taking the supremum over $\pi \in \mathcal{P}$ gives

$$\sup_{\pi\in\mathcal{P}}\left|\widehat{V}_M(\pi) - \widetilde{V}_M(\pi)\right| \le \frac{\widehat{r}_{n,M}}{\kappa},$$

which is Equation (12) and completes the proof. □

A.8 Proof of Lemma 5

Proof. Fix a policy $\pi \in \mathcal{P}$. Recall that

$$u_j = \frac{j - \frac{1}{2}}{M}, \qquad j = 1, \dots, M,$$

are the midpoint ranks. Let

$$I_j = \left(\frac{j-1}{M}, \frac{j}{M} \right], \qquad j = 1, \dots, M,$$

and define the midpoint projection

$$q_M(u) = u_j \qquad \text{for } u \in I_j.$$

We first establish a deterministic approximation bound for any bounded monotone quantile function.

For notational convenience, extend each $Q_a(\cdot \mid x)$ to $[0,1]$ by its one-sided endpoint limits. This preserves monotonicity and the bound $|Q_a(u \mid x)| \le B$.

Let $Q : [0,1] \to [-B, B]$ be any nondecreasing function. On the cell

$$I_j = [a_j, b_j], \qquad a_j = \frac{j-1}{M}, \quad b_j = \frac{j}{M},$$

with midpoint $u_j = (a_j + b_j)/2$, monotonicity implies

$$Q(a_j) \le Q(u) \le Q(u_j) \qquad \text{for } u \in [a_j, u_j],$$

and

$$Q(u_j) \le Q(u) \le Q(b_j) \qquad \text{for } u \in [u_j, b_j].$$

Therefore,

$$\begin{aligned} \int_{a_j}^{b_j} |Q\{q_M(u)\} - Q(u)| \, \mathrm{d}u &= \int_{a_j}^{u_j} \{Q(u_j) - Q(u)\} \, \mathrm{d}u \\ &\quad + \int_{u_j}^{b_j} \{Q(u) - Q(u_j)\} \, \mathrm{d}u \\ &\le \frac{1}{2M} \{Q(b_j) - Q(a_j)\}. \end{aligned}$$

Summing over $j = 1, \dots, M$ gives

$$\begin{aligned} \int_0^1 |Q\{q_M(u)\} - Q(u)| \, \mathrm{d}u &\le \frac{1}{2M} \sum_{j=1}^{M} \{Q(b_j) - Q(a_j)\} \\ &= \frac{Q(1) - Q(0)}{2M} \\ &\le \frac{B}{M}, \end{aligned}$$

where the last inequality follows from $Q(1) - Q(0) \le 2B$. Consequently, uniformly over $a \in \{0,1\}$ and $x \in \mathcal{X}$,

$$\int_0^1 |Q_a\{q_M(u) \mid x\} - Q_a(u \mid x)| \, \mathrm{d}u \le \frac{B}{M}. \tag{17}$$

We next translate this one-dimensional approximation bound into a bound for the rank-space utility. Since $s_{\delta,\kappa}$ is $1/\kappa$-Lipschitz, for every $(u, v) \in [0,1]^2$ and every $x \in \mathcal{X}$,

$$\begin{aligned} &|c_{\pi,x}\{q_M(u), q_M(v)\} - c_{\pi,x}(u, v)| \\ &\quad \le \frac{1}{\kappa} \left[|Q_0\{q_M(u) \mid x\} - Q_0(u \mid x)| + |Q_1\{q_M(v) \mid x\} - Q_1(v \mid x)| \right]. \end{aligned}$$

Indeed, the same bound holds separately for the treatment and control utility terms, and $\pi(x) \in [0,1]$ forms a convex combination of the two.

Now let $\gamma \in \Pi(\lambda, \lambda)$ be arbitrary. Since both marginals of γ are uniform on $[0,1]$, integrating the preceding inequality with respect to γ yields

$$\begin{aligned}\left| \int_{[0,1]^2} c_{\pi,x}\{q_M(u), q_M(v)\} \, \mathrm{d}\gamma(u,v) - \int_{[0,1]^2} c_{\pi,x}(u,v) \, \mathrm{d}\gamma(u,v) \right| \\ \leq \frac{1}{\kappa} \int_0^1 |Q_0\{q_M(u) \mid x\} - Q_0(u \mid x)| \, \mathrm{d}u \\ + \frac{1}{\kappa} \int_0^1 |Q_1\{q_M(v) \mid x\} - Q_1(v \mid x)| \, \mathrm{d}v \\ \leq \frac{2B}{\kappa M},\end{aligned}$$

where the last step follows from Equation (17). Averaging over X therefore gives, uniformly over $\pi \in \mathcal{P}$ and $\gamma \in \Pi(\lambda, \lambda)$,

$$\left| \mathbb{E}_X \left[\int c_{\pi,X}\{q_M(u), q_M(v)\} \, \mathrm{d}\gamma(u,v) \right] - \mathbb{E}_X \left[\int c_{\pi,X}(u,v) \, \mathrm{d}\gamma(u,v) \right] \right| \leq \frac{2B}{\kappa M}. \tag{18}$$

We now compare the continuous copula problem with the discrete Birkhoff problem.

First, fix any $\gamma \in \Pi(\lambda, \lambda)$ and define its cell-mass matrix $P^\gamma \in \mathbb{R}_+^{M \times M}$ by

$$P^\gamma_{jk} = \gamma(I_j \times I_k).$$

Because both marginals of γ are uniform,

$$\sum_{k=1}^M P^\gamma_{jk} = \gamma(I_j \times [0,1]) = \frac{1}{M},$$

and similarly

$$\sum_{j=1}^M P^\gamma_{jk} = \frac{1}{M}.$$

Hence $P^\gamma \in \mathcal{B}_M$. Moreover,

$$\begin{aligned}\langle P^\gamma, C_{\pi,M} \rangle &= \sum_{j=1}^M \sum_{k=1}^M P^\gamma_{jk} \mathbb{E}_X[c_{\pi,X}(u_j, u_k)] \\ &= \mathbb{E}_X \left[\int_{[0,1]^2} c_{\pi,X}\{q_M(u), q_M(v)\} \, \mathrm{d}\gamma(u,v) \right].\end{aligned}$$

By Equation (18),

$$\langle P^\gamma, C_{\pi,M} \rangle \leq \mathbb{E}_X \left[\int c_{\pi,X}(u,v) \, \mathrm{d}\gamma(u,v) \right] + \frac{2B}{\kappa M}.$$

Since $V_M(\pi)$ is the minimum over all $P \in \mathcal{B}_M$,

$$V_M(\pi) \leq \mathbb{E}_X \left[\int c_{\pi,X}(u,v) \, \mathrm{d}\gamma(u,v) \right] + \frac{2B}{\kappa M}.$$

Taking the infimum over $\gamma \in \Pi(\lambda, \lambda)$ gives

$$V_M(\pi) \leq \underline{V}_{\mathrm{cc}}(\pi) + \frac{2B}{\kappa M}. \tag{19}$$

For the reverse direction, fix any $P \in \mathcal{B}_M$. Construct a probability measure γ_P on $[0,1]^2$ by distributing the mass P_{jk} uniformly over the cell $I_j \times I_k$:

$$\gamma_P(\, \mathrm{d}u, \, \mathrm{d}v) = \sum_{j=1}^M \sum_{k=1}^M M^2 P_{jk} \mathbf{1}\{u \in I_j, \; v \in I_k\} \, \mathrm{d}u \, \mathrm{d}v.$$

The row and column constraints defining $\mathcal{B}_M$ imply that both marginals of γ_P are uniform, and hence $\gamma_P \in \Pi(\lambda, \lambda)$.

By construction,

$$\mathbb{E}_X \left[\int c_{\pi,X}\{q_M(u), q_M(v)\} \, \mathrm{d}\gamma_P(u,v) \right] = \langle P, C_{\pi,M} \rangle.$$

Applying Equation (18) to γ_P gives

$$\mathbb{E}_X \left[\int c_{\pi,X}(u,v) \, \mathrm{d}\gamma_P(u,v) \right] \leq \langle P, C_{\pi,M} \rangle + \frac{2B}{\kappa M}.$$

Since γ_P is feasible for the continuous shared-copula optimization,

$$\underline{V}_{\mathrm{cc}}(\pi) \leq \langle P, C_{\pi,M} \rangle + \frac{2B}{\kappa M}.$$

Taking the minimum over $P \in \mathcal{B}_M$ yields

$$\underline{V}_{\mathrm{cc}}(\pi) \leq V_M(\pi) + \frac{2B}{\kappa M}. \tag{20}$$

Combining Equation (19) and Equation (20), we obtain

$$|V_M(\pi) - \underline{V}_{\mathrm{cc}}(\pi)| \leq \frac{2B}{\kappa M}.$$

The bound is independent of π, so taking the supremum over $\pi \in \mathcal{P}$ gives

$$\sup_{\pi \in \mathcal{P}} |V_M(\pi) - \underline{V}_{\mathrm{cc}}(\pi)| \leq \frac{2B}{\kappa M},$$

which proves Equation (13). □

A.9 Proof of Lemma 6

Proof. Fix an arbitrary cost matrix $C \in \mathbb{R}^{M \times M}$, and define the unregularized and entropically regularized optimal transport values by

$$V_M(C) := \min_{P \in \mathcal{B}_M} \langle P, C \rangle$$

and

$$V_{M,\varepsilon}(C) := \min_{P \in \mathcal{B}_M} \left\{ \langle P, C \rangle + \varepsilon \mathrm{KL}(P \| U_M) \right\},$$

respectively, where

$$U_M = M^{-2} \mathbf{1}\mathbf{1}^\top.$$

We first prove the lower bound. Since the Kullback–Leibler divergence is nonnegative,

$$\mathrm{KL}(P \| U_M) \geq 0 \qquad \text{for every } P \in \mathcal{B}_M.$$

Therefore,

$$\langle P, C \rangle + \varepsilon \mathrm{KL}(P \| U_M) \geq \langle P, C \rangle$$

for every feasible P. Taking the minimum over $P \in \mathcal{B}_M$ on both sides gives

$$V_{M,\varepsilon}(C) \geq V_M(C),$$

and hence

$$V_{M,\varepsilon}(C) - V_M(C) \geq 0.$$

It remains to establish the upper bound. Because $\mathcal{B}_M$ is a nonempty compact polytope and the objective is continuous, there exists an unregularized optimizer

$$P^\star \in \arg\min_{P \in \mathcal{B}_M} \langle P, C \rangle.$$

Using $P^\star$ as a feasible candidate in the regularized problem gives

$$\begin{aligned} V_{M,\varepsilon}(C) &\le \langle P^\star, C\rangle + \varepsilon \mathrm{KL}(P^\star \| U_M) \\ &= V_M(C) + \varepsilon \mathrm{KL}(P^\star \| U_M). \end{aligned}$$

Thus it suffices to show that

$$\mathrm{KL}(P\|U_M) \le \log M \qquad \text{for every } P \in \mathcal{B}_M.$$

Indeed, every $P \in \mathcal{B}_M$ has nonnegative entries and row sums equal to $1/M$. Consequently,

$$0 \le P_{jk} \le \frac{1}{M} \qquad \text{for all } j,k \in [M].$$

Since $(U_M)_{jk} = 1/M^2$, for every entry with $P_{jk} > 0$,

$$\frac{P_{jk}}{(U_M)_{jk}} = M^2 P_{jk} \le M.$$

Using the convention $0 \log 0 = 0$, we therefore obtain

$$\begin{aligned} \mathrm{KL}(P\|U_M) &= \sum_{j=1}^{M}\sum_{k=1}^{M} P_{jk} \log\left(\frac{P_{jk}}{1/M^2}\right) \\ &\le \sum_{j=1}^{M}\sum_{k=1}^{M} P_{jk} \log M. \end{aligned}$$

Moreover, because the row sums of P are $1/M$,

$$\sum_{j=1}^{M}\sum_{k=1}^{M} P_{jk} = \sum_{j=1}^{M} \frac{1}{M} = 1.$$

Hence

$$\mathrm{KL}(P\|U_M) \le \log M.$$

Applying this bound to $P^\star$ yields

$$V_{M,\varepsilon}(C) \le V_M(C) + \varepsilon \log M.$$

Combining the upper and lower bounds gives

$$0 \le V_{M,\varepsilon}(C) - V_M(C) \le \varepsilon \log M,$$

or equivalently,

$$0 \le \min_{P\in\mathcal{B}_M} \{\langle P, C\rangle + \varepsilon \mathrm{KL}(P\|U_M)\} - \min_{P\in\mathcal{B}_M} \langle P, C\rangle \le \varepsilon \log M.$$

This proves Equation (14). □

A.10 Proof of Theorem 2

Assumption 3 (End-to-end rate conditions)**.** *The policy class $\mathcal{P}$ is finite. The conditional-quantile estimators are trained on an independent sample of size $n_q \asymp n$ and satisfy*

$$\sum_{a=0}^{1} \mathbb{E}_X \int_0^1 \left|\widehat{Q}_a(u \mid X) - Q_a(u \mid X)\right| \, \mathrm{d}u = O_p(r_{n_q}).$$

Moreover,

$$M \asymp \sqrt{\frac{n}{\log n}},$$

and the entropic-regularization and outer-optimization errors are of no larger order than the statistical terms. The estimated conditional quantile curves are taken to be nondecreasing in the rank argument, as is ensured by a non-crossing quantile estimator or by monotone rearrangement.

Proof. For a fixed policy $\pi \in \mathcal{P}$, recall the following three cost matrices:

$$\widehat{C}_\pi(j,k) = \mathbb{P}_n \widehat{c}_{\pi,X}(u_j, u_k),$$
$$\widetilde{C}_\pi(j,k) = \mathbb{P}_n c_{\pi,X}(u_j, u_k),$$

and

$$C_{\pi,M}(j,k) = \mathbb{E}_X \left[c_{\pi,X}(u_j, u_k) \right], \qquad j,k \in [M].$$

The first matrix uses the estimated conditional quantiles, the second uses the true conditional quantiles on the evaluation sample, and the third is the corresponding population cost matrix.

Define the associated unregularized optimal-transport values by

$$\widehat{V}_M(\pi) = \min_{P \in \mathcal{B}_M} \left\langle P, \widehat{C}_\pi \right\rangle,$$
$$\widetilde{V}_M(\pi) = \min_{P \in \mathcal{B}_M} \left\langle P, \widetilde{C}_\pi \right\rangle,$$

and

$$V_M(\pi) = \min_{P \in \mathcal{B}_M} \langle P, C_{\pi,M} \rangle .$$

Step 1: Uniform approximation of the shared-copula value. By Lemma 3, there exists a universal constant $C_0 > 0$ such that, with probability at least $1 - \alpha$,

$$\sup_{\pi \in \mathcal{P}} \left\| \widetilde{C}_\pi - C_{\pi,M} \right\|_\infty \le C_0 \left[\mathfrak{R}_n(\mathcal{P}) + \sqrt{\frac{\log(M/\alpha)}{n}} \right]. \tag{21}$$

On this event, Lemma 2 implies

$$\begin{aligned} \sup_{\pi \in \mathcal{P}} \left| \widetilde{V}_M(\pi) - V_M(\pi) \right| &\le \sup_{\pi \in \mathcal{P}} \left\| \widetilde{C}_\pi - C_{\pi,M} \right\|_\infty \\ &\le C_0 \left[\mathfrak{R}_n(\mathcal{P}) + \sqrt{\frac{\log(M/\alpha)}{n}} \right]. \end{aligned} \tag{22}$$

For every $\pi \in \mathcal{P}$, insert the two intermediate unregularized grid objectives to obtain

$$\begin{aligned} &\left| \widehat{V}_{M,\varepsilon}(\pi) - \underline{V}_{\mathrm{cc}}(\pi) \right| \\ &\quad \le \left| \widehat{V}_{M,\varepsilon}(\pi) - \widehat{V}_M(\pi) \right| + \left| \widehat{V}_M(\pi) - \widetilde{V}_M(\pi) \right| \\ &\qquad + \left| \widetilde{V}_M(\pi) - V_M(\pi) \right| + \left| V_M(\pi) - \underline{V}_{\mathrm{cc}}(\pi) \right|. \end{aligned} \tag{23}$$

We control the four terms in Equation (23) separately.

First, Lemma 6 gives

$$\sup_{\pi \in \mathcal{P}} \left| \widehat{V}_{M,\varepsilon}(\pi) - \widehat{V}_M(\pi) \right| \le \varepsilon \log M. \tag{24}$$

Second, by Lemma 4,

$$\sup_{\pi \in \mathcal{P}} \left| \widehat{V}_M(\pi) - \widetilde{V}_M(\pi) \right| \le \frac{\widehat{r}_{n,M}}{\kappa}. \tag{25}$$

Third, the sampling term is bounded by Equation (22). Finally, Lemma 5 gives

$$\sup_{\pi \in \mathcal{P}} \left| V_M(\pi) - \underline{V}_{\mathrm{cc}}(\pi) \right| \le \frac{2B}{\kappa M}. \tag{26}$$

Combining Equations (22)–(26) and taking the supremum over $\pi \in \mathcal{P}$ yields, on the event in Equation (21),

$$\begin{aligned} \Delta_n &:= \sup_{\pi \in \mathcal{P}} \left| \widehat{V}_{M,\varepsilon}(\pi) - \underline{V}_{\mathrm{cc}}(\pi) \right| \\ &\le C_0 \left[\mathfrak{R}_n(\mathcal{P}) + \sqrt{\frac{\log(M/\alpha)}{n}} \right] + \frac{\widehat{r}_{n,M}}{\kappa} + \frac{2B}{\kappa M} + \varepsilon \log M. \end{aligned}$$

After enlarging the universal constant so that $C \geq \max\{C_0, 1\}$, we obtain

$$\Delta_n \leq C \left[\mathfrak{R}_n(\mathcal{P}) + \sqrt{\frac{\log(M/\alpha)}{n}} + \frac{\widehat{r}_{n,M}}{\kappa} + \frac{2B}{\kappa M} \right] + \varepsilon \log M,$$

which proves Equation (15).

Step 2: Shared-copula policy regret. Let

$$\pi^\star_{\mathrm{cc}} \in \arg\max_{\pi \in \mathcal{P}} \underline{V}_{\mathrm{cc}}(\pi).$$

By the definition of Δ_n,

$$\underline{V}_{\mathrm{cc}}(\pi^\star_{\mathrm{cc}}) - \widehat{V}_{M,\varepsilon}(\pi^\star_{\mathrm{cc}}) \leq \Delta_n$$

and

$$\widehat{V}_{M,\varepsilon}(\widehat{\pi}) - \underline{V}_{\mathrm{cc}}(\widehat{\pi}) \leq \Delta_n.$$

Moreover, the assumed outer-optimization guarantee implies

$$\widehat{V}_{M,\varepsilon}(\pi^\star_{\mathrm{cc}}) - \widehat{V}_{M,\varepsilon}(\widehat{\pi}) \leq \xi_n.$$

Therefore,

$$\begin{aligned} &\underline{V}_{\mathrm{cc}}(\pi^\star_{\mathrm{cc}}) - \underline{V}_{\mathrm{cc}}(\widehat{\pi}) \\ &= \left[\underline{V}_{\mathrm{cc}}(\pi^\star_{\mathrm{cc}}) - \widehat{V}_{M,\varepsilon}(\pi^\star_{\mathrm{cc}}) \right] \\ &\quad + \left[\widehat{V}_{M,\varepsilon}(\pi^\star_{\mathrm{cc}}) - \widehat{V}_{M,\varepsilon}(\widehat{\pi}) \right] \\ &\quad + \left[\widehat{V}_{M,\varepsilon}(\widehat{\pi}) - \underline{V}_{\mathrm{cc}}(\widehat{\pi}) \right] \\ &\leq 2\Delta_n + \xi_n. \end{aligned} \tag{27}$$

Step 3: End-to-end regret rate. We now impose Assumption 3. Because $\mathcal{P}$ is a finite class of $[0,1]$-valued policies, Massart's finite-class bound gives

$$\mathfrak{R}_n(\mathcal{P}) \lesssim \sqrt{\frac{\log|\mathcal{P}|}{n}}. \tag{28}$$

It remains to relate the realized grid error $\widehat{r}_{n,M}$ to the integrated conditional-quantile error in Assumption 3. For $a \in \{0,1\}$ and $x \in \mathcal{X}$, define

$$h_{a,x}(u) = \left| \widehat{Q}_a(u \mid x) - Q_a(u \mid x) \right|.$$

Both $\widehat{Q}_a(\cdot \mid x)$ and $Q_a(\cdot \mid x)$ are nondecreasing and take values in $[-B, B]$. Hence

$$\mathrm{TV}_{[0,1]} \left(\widehat{Q}_a(\cdot \mid x) \right) \leq 2B$$

and

$$\mathrm{TV}_{[0,1]} \left(Q_a(\cdot \mid x) \right) \leq 2B.$$

Since the absolute-value map is 1-Lipschitz,

$$\mathrm{TV}_{[0,1]}(h_{a,x}) \leq 4B. \tag{29}$$

Let

$$I_j = \left(\frac{j-1}{M}, \frac{j}{M} \right], \qquad u_j = \frac{j - \frac{1}{2}}{M}.$$

For any function h of bounded variation,

$$\begin{aligned}\left|\frac{1}{M}\sum_{j=1}^{M}h(u_j)-\int_0^1 h(u)\,\mathrm{d}u\right| &= \left|\sum_{j=1}^{M}\int_{I_j}\{h(u_j)-h(u)\}\,\mathrm{d}u\right| \\ &\le \sum_{j=1}^{M}\int_{I_j}|h(u_j)-h(u)|\,\mathrm{d}u \\ &\le \frac{1}{M}\sum_{j=1}^{M}\mathrm{TV}_{I_j}(h) \\ &\le \frac{\mathrm{TV}_{[0,1]}(h)}{M}.\end{aligned}$$

Applying this inequality to $h_{a,x}$ and using Equation (29) gives

$$\frac{1}{M}\sum_{j=1}^{M}\left|\widehat{Q}_a(u_j\mid x)-Q_a(u_j\mid x)\right| \le \int_0^1\left|\widehat{Q}_a(u\mid x)-Q_a(u\mid x)\right|\,\mathrm{d}u+\frac{4B}{M}. \tag{30}$$

Define

$$g_M(x)=\sum_{a=0}^{1}\frac{1}{M}\sum_{j=1}^{M}\left|\widehat{Q}_a(u_j\mid x)-Q_a(u_j\mid x)\right|.$$

Then

$$\widehat{r}_{n,M}=\mathbb{P}_n g_M.$$

Averaging Equation (30) over X and summing over $a\in\{0,1\}$ yields

$$\mathbb{E}_X[g_M(X)]\le\sum_{a=0}^{1}\mathbb{E}_X\int_0^1\left|\widehat{Q}_a(u\mid X)-Q_a(u\mid X)\right|\,\mathrm{d}u+\frac{8B}{M}.$$

Therefore, by Assumption 3,

$$\mathbb{E}_X[g_M(X)]=O_p\left(r_{n_q}+\frac{1}{M}\right). \tag{31}$$

Because the nuisance-training sample is independent of the evaluation sample, we may condition on the fitted quantile functions. Under Assumption 2,

$$0\le g_M(x)\le 4B.$$

Conditional Hoeffding concentration therefore gives

$$\mathbb{P}_n g_M-\mathbb{E}_X[g_M(X)]=O_p(n^{-1/2}).$$

Combining this relation with Equation (31) shows that

$$\widehat{r}_{n,M}=O_p\left(r_{n_q}+\frac{1}{M}+\frac{1}{\sqrt{n}}\right). \tag{32}$$

Now take

$$M\asymp\sqrt{\frac{n}{\log n}}.$$

Then

$$\frac{1}{M}=O\left(\sqrt{\frac{\log n}{n}}\right)$$

and

$$\sqrt{\frac{\log(M/\alpha)}{n}}=O\left(\sqrt{\frac{\log n}{n}}\right)$$

for any fixed confidence level $\alpha \in (0,1)$. Equivalently, one may take $\alpha = n^{-2}$, in which case the same rate holds on an event whose probability converges to one.

Combining Equations (28) and (32) with Equation (15), and treating the fixed constants B and κ as independent of n, gives

$$\Delta_n = O_p\left(\sqrt{\frac{\log|\mathcal{P}|}{n}} + \sqrt{\frac{\log n}{n}} + \frac{r_{n_q}}{\kappa} + \varepsilon \log M\right).$$

By the final condition in Assumption 3, the entropic regularization error $\varepsilon \log M$ and the outer-optimization error ξ_n are of no larger order than the displayed statistical terms. Substituting into Equation (27) therefore yields

$$\underline{V}_{\mathrm{cc}}(\pi^\star_{\mathrm{cc}}) - \underline{V}_{\mathrm{cc}}(\widehat{\pi}) = O_p\left(\sqrt{\frac{\log|\mathcal{P}|}{n}} + \sqrt{\frac{\log n}{n}} + \frac{r_{n_q}}{\kappa}\right),$$

which proves Equation (16) and completes the proof. □

A.11 Proof of Corollary 1

Assumption 4 (Rank smoothness for sensitivity analysis)**.** *There exists a constant $L_Q < \infty$ such that, uniformly over $a \in \{0,1\}$, $x \in \mathcal{X}$, and $u, u' \in (0,1)$,*

$$|Q_a(u \mid x) - Q_a(u' \mid x)| \le L_Q |u - u'|.$$

Proof. We first show that Assumption 4 implies uniform Lipschitz continuity of the rank-space utility. For any $(u,v),(u',v') \in [0,1]^2$, the $1/\kappa$-Lipschitz property of $s_{\delta,\kappa}$ and Assumption 4 give

$$|s_{\delta,\kappa}\{Q_1(v \mid x) - Q_0(u \mid x)\} - s_{\delta,\kappa}\{Q_1(v' \mid x) - Q_0(u' \mid x)\}| \le \frac{L_Q}{\kappa}\left(|u-u'| + |v-v'|\right).$$

The same bound holds for the control utility. Since $\pi(x) \in [0,1]$, their convex combination therefore satisfies

$$|c_{\pi,x}(u,v) - c_{\pi,x}(u',v')| \le \frac{L_Q}{\kappa}\left(|u-u'| + |v-v'|\right),$$

uniformly over $\pi \in \mathcal{P}$ and $x \in \mathcal{X}$. Hence $c_{\pi,x}$ is L_Q/κ-Lipschitz under the ℓ_1 metric on $[0,1]^2$.

We now compare the approximately invariant value $\underline{V}_\rho(\pi)$ with the shared-copula value $\underline{V}_{\mathrm{cc}}(\pi)$. Every constant family $\gamma_x \equiv \gamma$, $\gamma \in \Pi(\lambda,\lambda)$, belongs to Γ_ρ. Therefore,

$$\underline{V}_\rho(\pi) \le \underline{V}_{\mathrm{cc}}(\pi),$$

which gives the lower bound

$$0 \le \underline{V}_{\mathrm{cc}}(\pi) - \underline{V}_\rho(\pi).$$

For the reverse inequality, fix any $\{\gamma_x\}_{x\in\mathcal{X}} \in \Gamma_\rho$ and any $\zeta > 0$. By the definition of Γ_ρ, there exists a common copula $\bar{\gamma} \in \Pi(\lambda,\lambda)$ such that

$$\mathbb{E}_X[W_1(\gamma_X, \bar{\gamma})] \le \rho + \zeta.$$

By Kantorovich–Rubinstein duality and the uniform L_Q/κ-Lipschitz property established above,

$$\int c_{\pi,x}\,\mathrm{d}\gamma_x \ge \int c_{\pi,x}\,\mathrm{d}\bar{\gamma} - \frac{L_Q}{\kappa} W_1(\gamma_x, \bar{\gamma}).$$

Averaging over X yields

$$\begin{aligned}\mathbb{E}_X\left[\int c_{\pi,X}\,\mathrm{d}\gamma_X\right] &\ge \mathbb{E}_X\left[\int c_{\pi,X}\,\mathrm{d}\bar{\gamma}\right] - \frac{L_Q}{\kappa}\mathbb{E}_X[W_1(\gamma_X,\bar{\gamma})] \\ &\ge \underline{V}_{\mathrm{cc}}(\pi) - \frac{L_Q}{\kappa}(\rho+\zeta).\end{aligned}$$

Since this holds for every admissible family $\{\gamma_x\} \in \Gamma_\rho$, taking the infimum over Γ_ρ and then letting $\zeta \downarrow 0$ gives

$$\underline{V}_\rho(\pi) \ge \underline{V}_{\mathrm{cc}}(\pi) - \frac{L_Q}{\kappa}\rho.$$

Combining the two inequalities proves, uniformly over $\pi \in \mathcal{P}$,

$$0 \le \underline{V}_{\mathrm{cc}}(\pi) - \underline{V}_{\rho}(\pi) \le \frac{L_Q}{\kappa}\rho.$$

It remains to establish the robust-oracle regret bound. Let

$$\pi^{\star}_{\rho} \in \arg\max_{\pi \in \mathcal{P}} \underline{V}_{\rho}(\pi).$$

The population sensitivity bound implies

$$\underline{V}_{\rho}(\pi^{\star}_{\rho}) \le \underline{V}_{\mathrm{cc}}(\pi^{\star}_{\rho}), \qquad \underline{V}_{\rho}(\widehat{\pi}) \ge \underline{V}_{\mathrm{cc}}(\widehat{\pi}) - \frac{L_Q}{\kappa}\rho.$$

Therefore,

$$\underline{V}_{\rho}(\pi^{\star}_{\rho}) - \underline{V}_{\rho}(\widehat{\pi}) \le \underline{V}_{\mathrm{cc}}(\pi^{\star}_{\rho}) - \underline{V}_{\mathrm{cc}}(\widehat{\pi}) + \frac{L_Q}{\kappa}\rho.$$

On the event in Theorem 2,

$$\sup_{\pi \in \mathcal{P}} \left| \widehat{V}_{M,\varepsilon}(\pi) - \underline{V}_{\mathrm{cc}}(\pi) \right| \le \Delta_n,$$

and the learned policy satisfies

$$\widehat{V}_{M,\varepsilon}(\widehat{\pi}) \ge \sup_{\pi \in \mathcal{P}} \widehat{V}_{M,\varepsilon}(\pi) - \xi_n.$$

Hence

$$\begin{aligned} \underline{V}_{\mathrm{cc}}(\pi^{\star}_{\rho}) - \underline{V}_{\mathrm{cc}}(\widehat{\pi}) &= \left[\underline{V}_{\mathrm{cc}}(\pi^{\star}_{\rho}) - \widehat{V}_{M,\varepsilon}(\pi^{\star}_{\rho})\right] + \left[\widehat{V}_{M,\varepsilon}(\pi^{\star}_{\rho}) - \widehat{V}_{M,\varepsilon}(\widehat{\pi})\right] \\ &\quad + \left[\widehat{V}_{M,\varepsilon}(\widehat{\pi}) - \underline{V}_{\mathrm{cc}}(\widehat{\pi})\right] \\ &\le 2\Delta_n + \xi_n. \end{aligned}$$

Substituting this bound above gives

$$\underline{V}_{\rho}(\pi^{\star}_{\rho}) - \underline{V}_{\rho}(\widehat{\pi}) \le \frac{L_Q}{\kappa}\rho + 2\Delta_n + \xi_n,$$

which proves Corollary 1. □

B Experimental Details

B.1 Datasets

ACIC 2016 semi-synthetic data. We use the covariates from the ACIC 2016 benchmark (Dorie et al., 2019) and generate treatments and outcomes from a controlled observational DGP. The dataset contains 4,802 observations and 58 original pretreatment variables. We retain all rows and variables, median-impute missing numerical entries, standardize nonbinary numerical variables, and one-hot encode categorical variables. The binary variable x_{38} is placed first in the encoded feature vector. Let $X = (X_1, \ldots, X_d)$ denote the resulting covariates. Treatment is assigned according to $A \mid X = x \sim \mathrm{Bernoulli}\{e(x)\}$, where $e(x)$ is the logistic transform of $0.55x_1 - 0.40x_2 + 0.35\sin(x_3) + 0.20x_4x_5$, truncated to $[0.10, 0.90]$. To induce heterogeneous conditional outcome distributions, we construct two nonlinear binary functions of disjoint subsets of observed covariates: $b(x) = \mathbf{1}\{0.45x_2 - 0.40x_3 + 0.30\sin(x_4) + 0.25x_5x_6 - 0.20x_7 + 0.15x_8x_9 \ge -0.30\}$ and $r(x) = \mathbf{1}\{0.40x_{10} + 0.35x_{11} - 0.30x_{12} + 0.25\sin(x_{13}) - 0.20x_{14}x_{15} + 0.15x_{16} - 0.10x_{17} \ge -0.10\}$. Their cross-product defines four DGP strata $S(x) = 2r(x) + b(x)$, whose labels are not provided as additional inputs to the policy learners. Write $z(u) = 2u - 1$ and $t(u) = \max\{(u - 0.75)/0.25, 0\}$. The stratum-specific base conditional quantiles are $q_{0,r,b,\theta}(u) = 2z(u)$ for both values of r, and $q_{1,r,b,\theta}(u) = 0.55 + 0.10b + 2z(u)$ when $r = 0$, whereas $q_{1,r,b,\theta}(u) = 0.20 + 0.10b + 2z(u) + 2\theta\{12t(u)^2 - 1\}$ when $r = 1$. We additionally introduce a smooth covariate-dependent location shift $m(x) = 0.35\tanh(x^{\top}w)$, where $w_j = \{0.30 - 0.50(j-1)/(d-1)\}/\sqrt{d}$, and define $Q_a(u \mid x) = q_{a,r(x),b(x),\theta}(u) + m(x)$. This construction makes both potential-outcome marginals depend on observed covariates while ensuring that θ changes only the shape of the treated marginal for

the $r(x) = 1$ strata. For the policy-switch experiment, the reported scalar control $g \in \{1, 3, 5\}$ is mapped to the shape parameter through $\theta(g) = 0.30 + 0.04g$. The resulting shared-copula gain and policy-switch gap are therefore consequences of the induced marginal shapes. Finally, we draw $U_0, U_1 \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$, independently of X and the treatment-assignment draw, set $Y(a) = Q_a(U_a \mid X)$, and observe $Y = AY(1) + (1 - A)Y(0)$. Hence the true conditional copula is the independence copula and is common across covariate values. In each repetition, the 4,802 observations are randomly partitioned into 2,000 policy-training observations and 2,802 held-out test observations. The same split is reused across DGP scenarios and methods within a repetition.

Fully simulated data. For the finite-sample experiment, we replace the ACIC covariates by a fully known covariate distribution while retaining the same treatment, stratum, and potential-outcome mechanisms described above. Specifically, we generate $X \in \mathbb{R}^{20}$ from a zero-mean Gaussian distribution with AR(1) covariance $\Sigma_{jk} = \rho^{|j-k|}$ and $\rho = 0.35$. Equivalently, $X_1 = \varepsilon_1$ and $X_j = \rho X_{j-1} + \sqrt{1 - \rho^2}\,\varepsilon_j$ for $j \geq 2$, with independent $\varepsilon_j \sim \mathcal{N}(0, 1)$. Given X, we use the same nonlinear functions $b(X)$ and $r(X)$, the same propensity score $e(X)$, and the same conditional quantile model $Q_a(u \mid X)$ as in the semi-synthetic design, fixing the shape parameter at $\theta = 0.54$. Potential-outcome ranks are again generated as independent $U_0, U_1 \sim \text{Unif}(0, 1)$, so $Y(a) = Q_a(U_a \mid X)$ satisfies the common-copula model with a true independence copula, while $A \mid X \sim \text{Bernoulli}\{e(X)\}$ produces a nonconstant observational treatment mechanism with overlap. We vary the total training sample size N and allocate one half to nuisance estimation and one half to policy learning, so $n_q = n = N/2$. Within each repetition, the nuisance and policy samples at increasing values of N are nested prefixes of independently generated maximum-size pools, so larger training sets add observations rather than replacing those used at smaller sample sizes. Evaluation uses a separate fixed sample of 4,000 covariate draws, shared across all sample sizes, repetitions, and methods, and all reported regrets are computed against the exact shared-copula oracle on this fixed test distribution.

B.2 Conditional Quantile Estimation

We estimate the two conditional marginal quantile functions using only the nuisance-training sample $\mathcal{D}_q = \{(X_i, A_i, Y_i)\}_{i=1}^{n_q}$. Each estimator uses a shared covariate encoder $h(x) = \tanh(Wx + b) \in \mathbb{R}^{32}$ and treatment-specific output heads. An observation contributes to the loss through its observed treatment head; both heads are evaluated subsequently at each policy-training covariate. The nuisance estimators receive observed covariates, treatments, and outcomes, without the DGP stratum labels or unobserved potential outcomes. Within each repetition and sample size, the fitted quantiles are shared by all policy learners being compared. All three estimators return $\widehat{Q}_a(u_j \mid x)$ on the same midpoint grid $u_j = (j - \frac{1}{2})/M$, $j = 1, \ldots, M$, with $M = 50$. The following descriptions specify our supervised adaptations of the cited methods and their implementation details.

Non-crossing quantile regression. Zhou et al. (2020) enforce ordered quantile outputs using cumulative softmax weights and a learned location and scale. We use the same principle of constructing ordered outputs, with an unrestricted central value and positive adjacent increments. Let $j_0 = \lfloor M/2 \rfloor + 1$. For each treatment a, an affine head applied to $h(x)$ produces an anchor $b_a(x)$ and $M - 1$ unconstrained gap logits $r_{a,\ell}(x)$. Defining $g_{a,\ell}(x) = \text{softplus}\{r_{a,\ell}(x)\}$, where $\text{softplus}(t) = \log(1 + e^t)$, we set

$$\widehat{Q}_a(u_j \mid x) = \begin{cases} b_a(x) - \displaystyle\sum_{\ell=j}^{j_0-1} g_{a,\ell}(x), & j < j_0, \\ b_a(x), & j = j_0, \\ b_a(x) + \displaystyle\sum_{\ell=j_0}^{j-1} g_{a,\ell}(x), & j > j_0. \end{cases} \tag{33}$$

Thus adjacent grid quantiles are strictly ordered by construction. With $M = 50$, the anchor is at $u_{26} = 0.51$. The head biases are initialized from the treatment-arm empirical quantiles, with initial adjacent gaps floored at 10^{-4} before applying the inverse softplus.

We train this network with the Huber-smoothed quantile loss

$$\ell_{\tau,h}(r) = |\tau - \mathbf{1}\{r < 0\}|\, H_h(r), \qquad H_h(r) = \begin{cases} r^2/(2h), & |r| \le h, \\ |r| - h/2, & |r| > h, \end{cases} \tag{34}$$

where $h = 0.05$ is distinct from the utility transition width κ. The data-fitting objective is

$$\mathcal{L}_{\text{grid}} = \frac{1}{n_q M} \sum_{i=1}^{n_q} \sum_{j=1}^{M} \ell_{u_j,h}\Big\{Y_i - \widehat{Q}_{A_i}(u_j \mid X_i)\Big\}. \tag{35}$$

This is supervised conditional-outcome estimation: our implementation replaces the reinforcement-learning targets in Zhou et al. (2020) with observed outcomes and uses the parameterization in (33).

Simultaneous quantile regression. Tagasovska & Lopez-Paz (2019) learn a quantile function by averaging the pinball loss over freshly sampled uniform quantile levels. We adopt this random-rank training objective, with the smoothed loss in (34) and an explicitly monotone continuous parameterization. This variant is named `monotone-iqn` in the code. Partition $[0, 1]$ into $K = 16$ equal intervals and define

$$\psi_k(\tau) = \min\left\{\max\left(\tau - \frac{k-1}{K}, 0\right), \frac{1}{K}\right\}, \qquad k = 1, \dots, K.$$

The network predicts an unrestricted median anchor $m_a(x)$ and slopes $d_{a,k}(x) = \text{softplus}\{r_{a,k}(x)\} + 10^{-4}$, giving

$$\widehat{Q}_a(\tau \mid x) = m_a(x) + \sum_{k=1}^{K} d_{a,k}(x)\{\psi_k(\tau) - \psi_k(1/2)\}. \tag{36}$$

The resulting curve is continuous, piecewise linear, and strictly increasing in τ. At each update, we independently draw $T = 16$ ranks per observation from $\text{Unif}(10^{-4}, 1 - 10^{-4})$ and minimize the minibatch average of $\ell_{\tau,h}\{Y_i - \widehat{Q}_{A_i}(\tau \mid X_i)\}$. The small endpoint exclusion is used in the implementation. Anchor biases are initialized at the arm-specific empirical medians; slope biases use empirical quantile differences at $K + 1$ equally spaced probabilities from 0.01 to 0.99, divided by $1/K$ and floored at 10^{-3}. Random-rank fitting follows the SQR approach, whereas the positive-slope representation and Huber smoothing are implementation choices in our experiments. The output grid is used only when querying the fitted function.

Neural spline flow. We use the monotone rational-quadratic transformation of Durkan et al. (2019) in a scalar conditional density model. Let μ_a and σ_a be the empirical mean and sample standard deviation of nuisance-training outcomes in arm a, with σ_a floored at 10^{-3}. Our model is

$$Y \mid X = x, A = a \overset{d}{=} \mu_a + \sigma_a f_a(Z; x), \qquad Z \sim \mathcal{N}(0, 1), \tag{37}$$

where $f_a(\cdot; x)$ is a rational-quadratic spline with $J = 12$ bins on $[-L, L]$, $L = 4$, and identity tails. A treatment-specific head predicts $3J - 1$ parameters: J width logits, J height logits, and $J - 1$ interior-derivative logits. Widths and heights are obtained by a softmax transformation, rescaled to sum to $2L$ with a minimum bin size of $\min(10^{-3}, 0.02L/J)$. Interior derivatives are softplus-transformed with a 10^{-3} floor; endpoint derivatives equal one.

Writing $z_i = f_{A_i}^{-1}\{(Y_i - \mu_{A_i})/\sigma_{A_i}; X_i\}$, the negative conditional log-likelihood is

$$\mathcal{L}_{\text{flow}} = \frac{1}{n_q} \sum_{i=1}^{n_q} \left[\frac{z_i^2}{2} + \frac{\log(2\pi)}{2} + \log \sigma_{A_i} + \log f'_{A_i}(z_i; X_i)\right]. \tag{38}$$

Here the inverse is computed analytically within the selected spline bin. If Φ denotes the standard normal CDF, quantiles are obtained directly as

$$\widehat{Q}_a(\tau \mid x) = \mu_a + \sigma_a f_a\{\Phi^{-1}(\tau); x\}. \tag{39}$$

Our implementation uses one conditional scalar spline per treatment head; the covariates determine its parameters. It does not model the joint distribution of the two potential outcomes.

Training settings and use in policy learning. For all three estimators, we use 200 Adam updates with learning rate 0.002, $(\beta_1, \beta_2) = (0.9, 0.999)$, numerical stabilizer 10^{-8}, and minibatch size $\min(256, n_q)$. Each minibatch is sampled without replacement within an update. The grid network applies an ℓ_2 penalty of 10^{-4} to weight matrices, excluding biases, and clips the resulting gradient norm at 5. The continuous quantile network and spline flow clip the data-loss gradient norm at 5 and use Adam weight decay 10^{-4} on all trainable parameters. Training uses the final iterate after the fixed update budget. These settings are held fixed across sample sizes and agree with the saved finite-sample run metadata.

The finite-sample experiment evaluates the three alternatives separately, refitting the nuisance model for every sample size and repetition on the nuisance half of the training data. Shared-copula and rectangular learners use the same estimated quantile arrays to construct their utility matrices. The DR learner uses the midpoint approximation $\widehat{\mu}_a(x) = M^{-1} \sum_{j=1}^{M} \widehat{Q}_a(u_j \mid x)$ for its conditional mean model, together with the separately estimated propensity score. Quantile predictions are checked for finiteness and strict ordering. In particular, the spline interval $[-L, L]$ specifies the nonlinear transformation region, not a bound on the outcome support.