MINIMUM VOLUME OPTIMAL TRANSPORT FOR MULTIVARIATE CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal transport (OT) enables multivariate conformal prediction (CP) to construct flexible prediction sets that reflect the underlying complex distributional geometry while retaining distribution-free coverage guarantees. Existing OT-based methods typically rely on a fixed Euclidean transport cost, which can be suboptimal in terms of the volume of prediction sets. We propose Minimum-Volume OTCP, a framework that learns a Mahalanobis ground cost to directly minimize prediction-region volume over a class of positive-definite metrics. Using a metric Brenier representation, we derive a volume-based objective that depends explicitly on the Mahalanobis cost matrix and optimize it jointly with the transport map. We establish finite-sample marginal coverage of the output prediction sets and optimality of volume metrics. We further develop an individualized extension that learns covariate-dependent transport metrics, adapting the shape and orientation of prediction sets to heterogeneous conditional distributions. For this extension, we also show the asymptotic conditional coverage of our method. Simulations and real-data experiments demonstrate substantial reductions in prediction-set volume.

1 INTRODUCTION

Conformal prediction (CP) constructs finite-sample, distribution-free prediction sets under exchangeability (Vovk et al., 2005). Suppose we have the labeled data $\{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$ and a new covariate $X_{n+1} \in \mathcal{X}$, whose label $Y_{n+1} \in \mathcal{Y}$ is unobserved. The goal is to construct a prediction set $\mathcal{C}(X_{n+1})$ based on the labeled data, satisfying the following coverage guarantee

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha,$$

where $\alpha \in (0, 1)$ is the nominal coverage level.

A central challenge in multivariate CP is to construct prediction sets that adapt to the geometry of the response distribution. Structured conformity scores offer simple and interpretable constructions, but their geometric flexibility is limited by the chosen score class (Johnstone & Cox, 2021; Messoudi et al., 2022; Braun et al., 2025). Capturing more complex multivariate uncertainty therefore calls for flexible distributional representations (Feldman et al., 2023; Fang et al., 2025; Dheur et al., 2025). Optimal transport (OT) offers such a representation by connecting the response distribution to a simple reference distribution through transport maps. These maps provide a basis for defining multivariate ranks and quantiles (Chernozhukov et al., 2017; Hallin et al., 2021), allowing prediction sets to be constructed according to the geometry of the response distribution.

Generally, conformal prediction for a multivariate response is often equipped with a multivariate score function $s(\cdot, \cdot)$. Considering the example $\mathcal{Y} \subseteq \mathbb{R}^d$, given a pretrained predictor $\widehat{\mu}(\cdot) : \mathcal{X} \to \mathbb{R}^d$, we use the multivariate residual score $s(x, y) = y - \widehat{\mu}(x)$. Let $\mathcal{P}$ denote the distribution of the multivariate score $s(X_i, Y_i)$, and let $\mathcal{Q}$ be a radial reference distribution, e.g., the uniform distribution on the unit ball of $\mathbb{R}^d$. Existing OT-based conformal methods typically transport the score distribution $\mathcal{P}$ to $\mathcal{Q}$ under the squared Euclidean ground cost, yielding the transport map

$$T_I \in \underset{T_\# \mathcal{P} = \mathcal{Q}}{\arg\min} \int \|s - T(s)\|^2 \, d\mathcal{P}(s). \tag{1}$$

The OT conformal prediction set takes the form $\mathcal{C}_I(X_{n+1}) = \{y \in \mathcal{Y} : \|T_I(s(X_{n+1}, y))\| \leq \hat{q}_I\}$, where $\hat{q}_I$ is the empirical quantile of $\{\|T_I(s(X_i, Y_i))\|\}_{i=1}^n$. However, although OT can produce

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.

highly flexible prediction sets, the transport geometry induced by the squared Euclidean ground cost need not be well suited to achieving a small-volume prediction set.

To address this limitation, we introduce *Minimum-Volume Optimal Transport Conformal Prediction* (MV-OTCP), which learns the OT ground metric to minimize prediction-set volume. For a positive-definite matrix M, we replace the squared Euclidean ground cost in (1) with $c_M(s,u) = (s-u)^\top M(s-u)$, obtaining a transport map T_M. Applying the same conformal calibration with this map yields the corresponding prediction set, denoted by $\mathcal{C}_M(X_{n+1})$. Our target metric minimizes the population volume objective:

$$M^\star \in \arg\min_M \mathbb{E}\left[\mathsf{Leb}\{\mathcal{C}_M(X_{n+1})\}\right],$$

where Leb denotes the Lebesgue measure. Because M enters the transport cost, learning M changes the transport map itself and hence the geometry of the resulting prediction region. Our contributions are threefold:

- We introduce MV-OTCP for learning the ground metric of OT-based conformal prediction sets. To account for heterogeneous conditional distributions, we develop an individualized extension, IMV-OTCP, which learns a covariate-dependent metric so that prediction-set shape and orientation can adapt locally to the covariate.
- Using a metric Brenier representation, we derive a Jacobian-based expression for prediction-set volume, yielding a tractable objective for metric learning. We establish finite-sample marginal coverage, consistency of the learned metric, and asymptotic agreement between the learned and oracle prediction regions. Under additional high-level conditions, we also establish asymptotic conditional coverage for IMV-OTCP.
- Experiments on synthetic and real data demonstrate improved prediction-set volume efficiency relative to existing conformal baselines while maintaining the target coverage.

2 Related work

We focus here on the two directions most closely related to our work, and refer readers to Vovk et al. (2005); Angelopoulos et al. (2024) for broader reviews of conformal prediction.

Multivariate conformal prediction. A growing literature develops conformal prediction methods for multivariate responses using different scores and representation-based constructions (Johnstone & Cox, 2021; Messoudi et al., 2022; Izbicki et al., 2022; Feldman et al., 2023; Fang et al., 2025; Dheur et al., 2025; Braun et al., 2025). Within the OTCP framework, SOTCP (Thurin et al., 2025; Klein et al., 2025) uses a transport map learned from an independent sample within split conformal calibration, whereas FOTCP (Ndiaye, 2025) incorporates each candidate test score into the transport assignment to obtain a full-conformal construction. Conditional neural OT (Kondratyev et al., 2026) learns covariate-dependent vector-rank maps whose radial norms yield split-conformal pullback regions. Our work retains this OT-based ranking principle but treats the ground geometry as a learnable component rather than fixing the squared Euclidean cost.

Minimum-volume conformal geometry. Conformal efficiency has been improved through score selection or aggregation and through direct adaptation of prediction-set shape (Yang & Kuchibhotla, 2025; Liang et al., 2026; Xu et al., 2026; Bao et al., 2026b). Kiyani et al. (2024) study length-optimal prediction intervals in the univariate setting through a minimax formulation. Most directly related to our multivariate minimum-volume formulation, the MVCS method of Braun et al. (2026) explicitly minimizes volume within a structured norm-based parametric class. Its multi-norm extension provides additional flexibility by allowing different norm-based regions over a prespecified partition. However, the partition must be specified in advance, and the resulting construction remains restricted to norm-based geometries within each partition element. Our prediction sets are inverse images of a radial reference ball under a generally nonlinear optimal transport map, allowing asymmetric and nonconvex geometries without requiring a prespecified partition or restricting the regions to norm-based shapes. Our extension to more general downstream task losses is also related to work that directly optimizes decision losses of conformal prediction sets (Yeh et al., 2024; Bao et al., 2026a).

3 MINIMUM-VOLUME OTCP

3.1 RECAP OF OT-BASED CONFORMAL PREDICTION

Let $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ be a calibration sample independent of the pretrained predictor, and let $U_1, \ldots, U_{n+1}$ be i.i.d. draws from reference distribution $\mathcal{Q}$. For a trial response y at X_{n+1}, write $S_i^y = s(X_i, Y_i)$ for $i \in [n]$ and $S_{n+1}^y = s(X_{n+1}, y)$. At the empirical level, the optimal transport plan (1) between these two discrete measures is a classic assignment problem. FOTCP then solves

$$\widehat{\sigma}^y \in \operatorname*{arg\,min}_{\sigma \in \mathfrak{S}_{n+1}} \sum_{i=1}^{n+1} \|S_i^y - U_{\sigma(i)}\|^2,$$

where $\mathfrak{S}_{n+1}$ is the set of all permutations of $[n+1]$. Let $\widehat{q}_{1-\alpha}(\{a_i\}_{i=1}^n)$ be the $(1-\alpha)$-quantile of the empirical distribution of $\{a_i\}_{i=1}^n$. Then the FOTCP prediction set is

$$\widehat{\mathcal{C}}^{\text{FOTCP}}(X_{n+1}) = \left\{y \in \mathcal{Y} : \|U_{\widehat{\sigma}^y(n+1)}\| \le \widehat{q}_{1-\alpha}\left(\{\|U_i\|\}_{i=1}^{n+1}\right)\right\}.$$

Under exchangeability of $\{(X_i, Y_i)\}_{i=1}^{n+1}$, we have $\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{\text{FOTCP}}(X_{n+1})\} \ge 1 - \alpha$.

3.2 POPULATION VOLUME METRIC OF OTCP

The standard squared Euclidean transport cost treats all directions in the score space equally, which need not be aligned with the criterion used to evaluate the resulting conformal set. We therefore introduce a volume-adaptive geometry into the transport problem. Specifically, we replace the Euclidean cost by the Mahalanobis-type quadratic cost

$$c_M(s, u) = (s-u)^\top M(s-u), \qquad M = M^\top \succ 0, \tag{2}$$

The matrix M allows the transport geometry to reweight and rotate different score directions according to the downstream objective, while preserving the quadratic structure of the transport problem. For each fixed M, let T_M denote the population optimal transport map from the score distribution $\mathcal{P}$ to the reference distribution $\mathcal{Q}$ under the cost c_M , namely,

$$T_M \in \operatorname*{arg\,min}_{T_\# \mathcal{P} = \mathcal{Q}} \int c_M\{s, T(s)\}\, d\mathcal{P}(s), \tag{3}$$

Let $F_Q(t) = \mathcal{Q}\{u : \|u\| \le t\}$, $q_\star = F_Q^{-1}(1-\alpha)$ and consider residual score $s(x, y) = y - \widehat{\mu}(x)$. The population oracle prediction set can be written as $\mathcal{C}_M(x) = \widehat{\mu}(x) + T_M^{-1}(\mathbb{B}_{1-\alpha})$, where $\mathbb{B}_{1-\alpha} = \{u : \|u\| \le q_\star\}$. Since translation does not change Lebesgue measure, the volume of $\mathcal{C}_M(x)$ is determined entirely by the geometry induced by M. Thus we define the population volume target by

$$M^\star \in \operatorname*{arg\,min}_{M \in \mathcal{M}_0} \mathcal{J}(M), \qquad \mathcal{J}(M) = \mathsf{Leb}\{\mathcal{C}_M(x)\} = \mathsf{Leb}\{T_M^{-1}(\mathbb{B}_{1-\alpha})\},$$

where $\mathcal{M}_0 = \{M = M^\top \succ 0 : \lambda_{\min}(M) \ge \underline{\lambda},\ \lambda_{\max}(M) \le \overline{\lambda},\ \det(M) = 1\}$. The constraint $\det(M) = 1$ resolves the scale non-identifiability of M, while the eigenvalue bounds ensure that the resulting parameter space is compact. The dependence of $\mathcal{J}(M)$ on M is implicit through the transport map. The following metric Brenier representation makes the geometry explicit.

Lemma 3.1 (Metric Brenier representation)**.** *Suppose $\mathcal{P}$ and $\mathcal{Q}$ have finite second moments and $\mathcal{P}$ is absolutely continuous. Then* (3) *has a unique solution up to $\mathcal{P}$-null sets, with $M^{1/2} T_M(s) = \nabla\psi(M^{1/2}s)$ for a convex ψ. If $\mathcal{Q}$ is also absolutely continuous, the reverse map G_M exists, is the almost-everywhere inverse of T_M, and has the representation $G_M(u) = M^{-1/2}\nabla g_M(M^{1/2}u)$ for a convex potential g_M.*

Under the condition of Lemma 3.1, we can rewrite $\mathcal{C}_M(x) = \widehat{\mu}(x) + G_M(\mathbb{B}_{1-\alpha})$. Assuming that g_M is twice differentiable, Lemma 3.1 and a change-of-variables argument yield

$$\mathcal{J}(M) = \mathbb{E}_U\left[\{f_Q(U)\}^{-1}\mathbb{1}\{U \in \mathbb{B}_{1-\alpha}\}\det\{\nabla^2 g_M(M^{\frac{1}{2}}U)\}\right]. \tag{4}$$

This representation is useful because it converts the volume loss into an optimization problem over a metric and a convex transport potential. The coverage level is handled by integrating over the

fixed reference region $\mathbb{B}_{1-\alpha}$, while the geometry of the prediction set enters through the Hessian determinant. In computation, the outer expectation over U can be approximated using independent reference samples, while the transport map can be learned using any suitable parametric model, including neural parameterizations of Brenier maps (Makkuva et al., 2020; Chaudhari et al., 2025).

Example 3.1 (Explicit volume-optimal metric)**.** *Let $\mathcal{Q}$ be a radial reference distribution on the unit ball, represented as $U = R\Theta$, where $R \sim \mathrm{Unif}(0,1)$ and Θ is uniformly distributed on the unit sphere. Let $\mathcal{P}$ have density $f_P(s) = |\Sigma|^{-1/2} g(\|\Sigma^{-1/2}(s-m)\|)$, where g is nonincreasing. Writing $r(s) = \|\Sigma^{-1/2}(s-m)\|$ and F_ρ for the distribution of $r(S)$, the map $T_{M^\star}(s) = F_\rho\{r(s)\}\frac{\Sigma^{-1/2}(s-m)}{r(s)}$ induces the prediction set $T_{M^\star}^{-1}(\mathbb{B}_{1-\alpha}) = \{s : r(s) \le F_\rho^{-1}(1-\alpha)\}$. This set is the central density-superlevel ellipsoid and is therefore the minimum-volume measurable set with probability $1-\alpha$. Moreover, by Lemma 3.1, $T_{M^\star}$ is generated by the metric $M^\star_{\mathrm{vol}} \propto \Sigma^{-1/2}$. See Figure 1 for an illustration and Appendix A.1 for derivation details.*

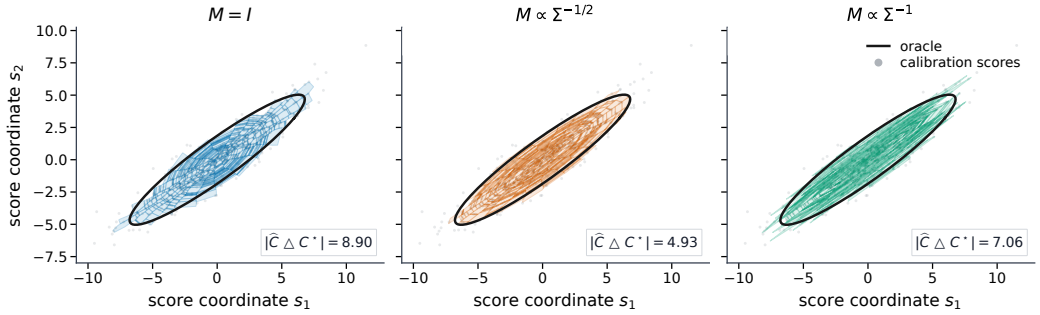


Figure 1: Metric choices on an elliptical score distribution. The accepted assignment regions are finite unions of polyhedral cells (Appendix B.1) under $M = I$, $M \propto \Sigma^{-1/2}$, and $M \propto \Sigma^{-1}$. Among the three choices, $M \propto \Sigma^{-1/2}$ yields the smallest symmetric difference from the oracle region.

General task-aware OTCP. The same transport representation also permits a downstream task to determine which parts of the response space are costly to include. We consider an averaged robust optimization problem. For a decision $z \in \mathcal{Z}$ and an integrable loss density $\ell(z,y)$, define the decision loss over a prediction set $\mathcal{C}(x)$ as $\mathcal{L}(z,\mathcal{C}(x)) = \int_{\mathcal{C}(x)} \ell(z,y)\,dy$. This formulation encompasses the optimization problems considered in Bertsimas & Ma (2023); Cortes-Gomez et al. (2025). The corresponding task-aware target for metric learning is

$$M^\star_{\mathrm{task}} \in \operatorname*{arg\,min}_{M \in \mathcal{M}_0} \mathcal{J}_{\mathrm{task}}(M), \qquad \mathcal{J}_{\mathrm{task}}(M) = \mathbb{E}_X\left[\mathcal{L}\{z^\star_M(X), \mathcal{C}_M(X)\}\right],$$

where $z^\star_M(x) \in \arg\min_{z\in\mathcal{Z}} \mathcal{L}(z, \mathcal{C}_M(x))$. In line (4), the general task target can be rewritten as

$$\mathcal{J}_{\mathrm{task}}(M) = \mathbb{E}_{X,U}\left[\ell\{z^\star_M(X), \widehat{\mu}(X) + G_M(U)\} \cdot \{f_Q(U)\}^{-1} \mathbb{1}\{U \in \mathbb{B}_{1-\alpha}\} \det\{\nabla G_M(U)\}\right]. \tag{5}$$

Volume is recovered as the special case $\ell \equiv 1$, whereas a general task loss additionally accounts for where the accepted volume is placed in the response space; several examples are given in Appendix A.3. The resulting metric is then used to construct the final conformal prediction set. Further implementation details are provided in Appendix A.3.

3.3 Metric estimation and conformal calibration

Let $\mathcal{D}_{\mathrm{est}} = \{(X_i, Y_i)\}_{i=1}^N$ be an independent estimation sample, and let $U_1, \ldots, U_N \sim \mathcal{Q}$ be independent reference draws. For a fixed $M \in \mathcal{M}_0$, let Ψ_θ be a parameterized approximation of the Brenier map $\nabla g_M(\cdot)$. Then the corresponding empirical version of (4) is

$$\widehat{\mathcal{J}}(M,\theta) = \frac{1}{N}\sum_{i=1}^N \{f_Q(U_i)\}^{-1} \mathbb{1}\{U_i \in \mathbb{B}_{1-\alpha}\} \det\{\nabla \Psi_\theta(M^{1/2}U_i) + \delta_N I_d\}, \tag{6}$$

where $\delta_N > 0$ is a small numerical stabilizer. The empirical minimum-volume metric-learning problem is

$$\widehat{M} \in \operatorname*{arg\,min}_{M \in \mathcal{M}_0} \widehat{\mathcal{J}}(M, \widehat{\theta}(M)),$$

where $\widehat{\theta}(M)$ is the output of an OT-fitting subroutine at metric M. Since the fitted transport parameters $\widehat{\theta}(M)$ themselves depend on M, the gradient of the profiled objective must account for this dependence. Therefore we do so via implicit differentiation through the inner OT-fitting problem. Additional implementation details are provided in Appendix A.2.

To form the final conformal set, hold $\widehat{M}$ fixed and use a fresh calibration sample $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$. For a trial response y at a new covariate x, compute the assignment

$$\widehat{\sigma}^{\,y}_{\widehat{M}} \in \operatorname*{arg\,min}_{\sigma \in \mathfrak{S}_{n+1}} \sum_{i=1}^{n+1} (S_i^y - U_{\sigma(i)})^\top \widehat{M} (S_i^y - U_{\sigma(i)}). \tag{7}$$

The resulting Minimum Volume OT Conformal Prediction (MV-OTCP) set is

$$\widehat{\mathcal{C}}^{\text{MV}}(x) = \left\{ y \in \mathcal{Y} : \|U_{\widehat{\sigma}^{\,y}_{\widehat{M}}(n+1)}\| \le \widehat{q}_{1-\alpha}\left(\{\|U_i\|\}_{i=1}^{n+1}\right) \right\}.$$

Theorem 3.1 (Marginal coverage for MV-OTCP)**.** *Assume that $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, that $\widehat{M}$ is learned independently of the final calibration sample and test response, and that ties in* (7) *are resolved permutation-equivariantly. Then $\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{\text{MV}}(X_{n+1})\} \ge 1 - \alpha$.*

Similar to FOTCP (Ndiaye, 2025), for every fixed metric $M = M^\top \succ 0$, the empirical set has a polyhedral representation. See Appendix B.1 for additional details.

3.4 Theoretical results of MV-OTCP

For notational simplicity, define

$$\zeta_M(u) = \{f_Q(U)\}^{-1} \mathbb{1}\{u \in \mathbb{B}_{1-\alpha}\} \det\{\nabla^2 g_M(M^{\frac{1}{2}} u)\}.$$

Then $\mathcal{J}(M) = \mathbb{E}\{\zeta_M(U)\}$. Write $\mathcal{J}_N(M) = N^{-1} \sum_{i=1}^N \zeta_M(U_i)$, and $\widehat{\mathcal{J}}(M) = \widehat{\mathcal{J}}\{M, \widehat{\theta}(M)\}$. Thus $\mathcal{J}_N$ is the infeasible empirical average of the population integrand, while $\widehat{\mathcal{J}}$ is the actual plug-in criterion used for metric learning. For $M_c \in \mathcal{M}_0$, let $\mathcal{M}_0(M_c, \epsilon) = \{M \in \mathcal{M}_0 : \|M - M_c\|_F \le \epsilon\}$.

Assumption 3.1. *(i) There exist $B_\zeta, L_\zeta \in (0, \infty)$ such that $|\zeta_M(U)| \le B_\zeta$ a.s. for all $M \in \mathcal{M}_0$. Moreover, $\sup_{M_c \in \mathcal{M}_0} \mathbb{E}_U \left[\sup_{M \in \mathcal{M}_0(M_c, \epsilon)} |\zeta_M(U) - \zeta_{M_c}(U)|\right] \le L_\zeta \epsilon$, for any $\epsilon > 0$. (ii) $\mathcal{E}_N^{\text{obj}} := \sup_{M \in \mathcal{M}_0} |\widehat{\mathcal{J}}(M) - \mathcal{J}_N(M)|$ satisfies $\mathcal{E}_N^{\text{obj}} = o_{\mathbb{P}}(1)$ as $N \to \infty$. (iii) For every $\epsilon > 0$, $\inf_{M \in \mathcal{M}_0:\, \|M - M^\star\|_F \ge \epsilon} \{\mathcal{J}(M) - \mathcal{J}(M^\star)\} > 0$.*

Assumption 3.1(i) imposes boundedness and local Lipschitz regularity on the population objective integrand. Part (ii) isolates the plug-in transport error induced by $\Psi_{\widehat{\theta}}$ entering metric learning. Part (iii) is a standard separation condition ensuring that metrics bounded away from $M^\star$ have strictly larger population objective values. Appendix F.3 gives primitive sufficient conditions for Assumption 3.1(i).

Theorem 3.2. *Suppose Assumption 3.1(i) holds. For fixed d, there exists a universal constant $C > 0$ such that, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\sup_{M \in \mathcal{M}_0} |\widehat{\mathcal{J}}(M) - \mathcal{J}(M)| \le \mathcal{E}_N^{\text{obj}} + C B_\zeta \sqrt{\frac{d^2 \log\{e + 2\sqrt{d}\,\overline{\lambda} L_\zeta N / B_\zeta\} + \log(2/\delta)}{N}}.$$

In addition, if parts (ii)–(iii) of Assumption 3.1 hold, then $\mathbb{E}\|\widehat{M} - M^\star\|_F \to 0$

Theorem 3.2 decomposes the error of the learned volume objective into the transport approximation error $\mathcal{E}_N^{\text{obj}}$ and a uniform empirical fluctuation term. The latter vanishes at the usual $N^{-1/2}$ rate up to logarithmic factors. Hence, if the fitted transport is uniformly consistent and the population objective is well separated around $M^\star$, the learned metric $\widehat{M}$ converges to the population-optimal metric.

The final conformal geometry depends on both $\widehat{M}$ and the empirical assignment map. Define the oracle set $\mathcal{C}_\star^{\mathrm{MV}}(x) = \mathcal{C}_{M^\star}(x)$. We now investigate the asymptotic agreement between the oracle set $\mathcal{C}_\star^{\mathrm{MV}}(x)$ and empirical set $\widehat{\mathcal{C}}^{\mathrm{MV}}(x)$.

Assumption 3.2. *Let $S \sim \mathcal{P}$ be independent of the calibration and reference data. (i) $\mathcal{Q}$ is absolutely continuous with compact convex support, and F_Q is continuously differentiable near $q_\star$ with $F_Q'(q_\star) > 0$. (ii) $\mathcal{P}$ is absolutely continuous with compact support, and there exists $L_{\mathrm{OT}} < \infty$ such that $\|T_M(s) - T_M(s')\| \le L_{\mathrm{OT}}\|s - s'\|$ for all $M \in \mathcal{M}_0$ and $s, s' \in \mathrm{supp}(\mathcal{P})$. (iii) There exists $L_T < \infty$ such that $\mathbb{E}_S\|T_M(S) - T_{M^\star}(S)\| \le L_T\|M - M^\star\|_F$ for all $M \in \mathcal{M}_0$.*

Assumption 3.2 (i)–(ii) impose standard smoothness and compactness conditions on the reference and score distributions, together with a uniform Lipschitz condition on the population transport map; conditions of this type are used in Deb et al. (2021). Assumption 3.2 (iii) further requires the transport map to vary smoothly with the ground metric, ensuring that perturbations of M induce only proportional changes in the transported scores.

Theorem 3.3. *Suppose Assumptions 3.1 and 3.2 hold. Assume that the observations in $\mathcal{D}_{\mathrm{est}}$, $\mathcal{D}_{\mathrm{cal}}$, and (X_{n+1}, Y_{n+1}) are i.i.d. from the data distribution, and assignment ties are resolved permutation-equivariantly. Then*

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\star^{\mathrm{MV}}(X_{n+1}) \triangle \widehat{\mathcal{C}}^{\mathrm{MV}}(X_{n+1})\} \longrightarrow 0 \qquad \text{as } N, n \to \infty.$$

Theorem 3.3 shows that, as both estimation and calibration samples grow, the probability of the disagreement between the MV-OTCP region and its population oracle vanishes. Thus the learned construction asymptotically recovers the oracle acceptance region in probability.

4 Individualized minimum-volume OTCP

MV-OTCP uses one ground metric for the marginal score law. When the conditional score distribution varies with the covariate, a single geometry can be inefficient even if it remains marginally valid. We therefore first localize the transport map and then learn a local minimum-volume metric. Let $\mathcal{P}_x$ denote the conditional law of $S = s(X, Y)$ given $X = x$. For a fixed localization point x and a candidate metric $M \in \mathcal{M}_0$, define the local population Monge map

$$T_{M,x} \in \operatorname*{arg\,min}_{T_\# \mathcal{P}_x = \mathcal{Q}} \int c_M\{s, T(s)\}\, d\mathcal{P}_x(s).$$

By Lemma 3.1, for a convex potential $g_{M,x}$, $T_{M,x}^{-1}(u) = M^{-1/2}\nabla g_{M,x}(M^{1/2}u)$. The pointwise task-aware metric is defined through the local volume objective

$$\mathcal{J}_x(M) = \mathsf{Leb}\{T_{M,x}^{-1}(\mathbb{B}_{1-\alpha})\} = \mathbb{E}_U\left[\{f_Q(U)\}^{-1}\mathbb{1}\{U \in \mathbb{B}_{1-\alpha}\}\det\{\nabla^2 g_{M,x}(M^{1/2}U)\}\right]. \quad (8)$$

Thus, for each fixed localization point x, the oracle metric is $M_x^\star \in \arg\min_{M \in \mathcal{M}_0} \mathcal{J}_x(M)$. Therefore, define the population oracle set as $\mathcal{C}_\star^{\mathrm{IMV}}(x) = \left\{y \in \mathcal{Y} : \|T_{M_x^\star,x}\{s(x,y)\}\| \le q_\star\right\}$. Since $(T_{M_x^\star,x})_\# \mathcal{P}_x = \mathcal{Q}$, it satisfies $\mathbb{P}\{Y \in \mathcal{C}_\star^{\mathrm{IMV}}(x) \mid X = x\} = \mathcal{Q}\{u : \|u\| \le q_\star\} = 1 - \alpha$.

At the empirical level, for each fixed M and covariate value x, we let $\Psi_{\theta,x}$ denote a parameterized approximation to the Brenier map $\nabla g_{M,x}(\cdot)$ using any suitable parametric optimal transport model, in line with the MV-OTCP construction. Given estimation sample $\mathcal{D}_{\mathrm{est}}$, the empirical counterpart of (8) is

$$\widehat{\mathcal{J}}_x(M, \theta) = \frac{1}{N}\sum_{j=1}^{N}\{f_Q(U_j)\}^{-1}\mathbb{1}\{U_j \in \mathbb{B}_{1-\alpha}\}\det\left\{\nabla\Psi_{\theta,x}(M^{1/2}U_j) + \delta_N I_d\right\}.$$

The pointwise metric estimate is

$$\widehat{M}_x \in \operatorname*{arg\,min}_{M \in \mathcal{M}_0} \widehat{\mathcal{J}}_x\{M, \widehat{\theta}_x(M)\}.$$

After estimating $\widehat{M}_x$, we use the same estimation sample to fit a separate score-to-reference map $\widehat{T}_{\widehat{M}_x,x}$ for split calibration. Implementation details of learning $\widehat{M}_x$ and $\widehat{T}_{\widehat{M}_x,x}$ are given in Appendix A.4. Let $\mathcal{D}_{\mathrm{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ be independent of the estimation sample. For any (x, y),

define $\widehat{R}_N(x,y) = \|\widehat{T}_{\widehat{M}_x,x}\{s(x,y)\}\|$. The resulting Individualized Minimum-Volume OT Conformal Prediction (IMV-OTCP) set is

$$\widehat{\mathcal{C}}^{\mathrm{IMV}}(x) = \left\{ y \in \mathcal{Y} : \widehat{R}_N(x,y) \le \widehat{q}_{1-\alpha}(\{\widehat{R}_N(X_i,Y_i)\}_{i=1}^n \cup \{\infty\}) \right\}. \tag{9}$$

Theorem 4.1 (Marginal coverage for IMV-OTCP)**.** *Assume that $(X_i,Y_i)_{i=1}^{n+1}$ are exchangeable. Suppose all quantities defining the score rule $(x,y) \mapsto \widehat{R}_N(x,y)$ are fitted using data independent of the calibration sample and test point, and that this fixed rule is applied to all calibration and test observations. Then $\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{\mathrm{IMV}}(X_{n+1})\} \ge 1-\alpha$.*

In practice, fitting a separate metric $\widehat{M}_x$ and transport map for every covariate value x can be computationally expensive, so we use anchor-based amortization of the pointwise transport fits; see Appendix A.4. Let $\mathcal{H}_N$ be the σ-algebra generated by $\mathcal{D}_{\mathrm{est}}$, reference draws, and pretrained predictor $\widehat{\mu}$, and define $R^\star(x,y) = \|T_{M_x^\star,x}\{s(x,y)\}\|$.

Assumption 4.1. *At every fixed interior $x \in \mathcal{X}$, $\mathcal{E}_N^{\mathrm{rad}}(x) = o_{\mathbb{P}}(1)$, where $\mathcal{E}_N^{\mathrm{rad}}(x) := \mathbb{E}[|\widehat{R}_N(x,Y) - R^\star(x,Y)| \mid X = x, \mathcal{H}_N]$. Moreover, there exists a nonnegative P_X-integrable function B_{rad} such that $\mathcal{E}_N^{\mathrm{rad}}(x) \le B_{\mathrm{rad}}(x)$ almost surely for all N.*

Theorem 4.2. *Suppose Assumption 4.1 holds, F_Q is continuous and strictly increasing at $q_\star$, and the calibration samples and test observation are i.i.d. and independent of $\mathcal{H}_N$. For any fixed interior x,*

$$\lim_{(N,n)\to\infty} \mathbb{P}\left\{ Y_{n+1} \in \mathcal{C}_\star^{\mathrm{IMV}}(x) \triangle \widehat{\mathcal{C}}^{\mathrm{IMV}}(x) \,\middle|\, X_{n+1} = x \right\} = 0,$$

and therefore $\lim_{(N,n)\to\infty} \mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{\mathrm{IMV}}(x) \mid X_{n+1} = x\} = 1-\alpha$.

5 Experiments

We conduct experiments on synthetic data and five real datasets to compare with SCP (Vovk et al., 2005), MVCS (Braun et al., 2026), FOTCP (Ndiaye, 2025), and RLCP (Hore & Barber, 2025). The nominal coverage is $1-\alpha = 0.9$, using 100 replications with common samples across methods. Predictors and metric parameters are fitted before calibration, and test data are reserved for evaluation. All methods use the residual score $s(x,y) = y - \widehat{\mu}(x)$ with a ridge-regression predictor fitted on the training data. Additional experiments and experimental details appear in Appendices D and E.

Evaluation metrics. For a prediction set $\widehat{C}(X)$ and test observations indexed by $i = 1,\dots,n_{\mathrm{test}}$, let $C_i = \mathbb{1}\{Y_i \in \widehat{C}(X_i)\}$ and $V_i = \mathsf{Leb}\{\widehat{C}(X_i)\}$. The following criteria are used for comparison: (1) *Marginal coverage*: $n_{\mathrm{test}}^{-1}\sum_i C_i$; (2) *Set volume*: $\mathrm{Median}_i\, V_i$ (root volume $\mathrm{Median}_i\, V_i^{1/d}$ for real data); (3) *Symmetric difference*: $\mathrm{Median}_i\, \mathsf{Leb}\{\widehat{C}(X_i) \triangle C^\star(X_i)\}$, where $C^\star(x)$ denotes the oracle $(1-\alpha)$ highest-density region.; (4) *Worst-case coverage*: $\min_B \sum_i \mathbb{1}(X_i \in B) C_i / \sum_i \mathbb{1}(X_i \in B)$, with 100 random balls B on $\mathcal{X}$; (5) $\sqrt{\mathrm{MSCE}}$: $\{\mathbb{E}[\{\mathbb{P}(Y \in \widehat{\mathcal{C}}(X) \mid X) - (1-\alpha)\}^2]\}^{1/2}$, the empirical estimation of this quantity differs between the simulation and real-data experiments; detailed definitions are provided in Appendix E.1. Error bars show mean ± 1.96 standard errors.

5.1 Homogeneous simulations

We generate $Y = m(X) + \varepsilon$ with $X \sim \mathrm{Unif}[-2,2]$ and three bivariate residual laws. Let $e(\theta) = (\cos\theta, \sin\theta)^\top$ and let $R_0 = \{-2\log V\}^{1/2}$ for $V \sim \mathrm{Unif}(0,1)$. We draw Θ with density proportional to $r_j(\theta)^2$ with $j = 1,2,3$ corresponding to the three settings, and generate residuals as $\varepsilon = A_j\{R_0 r_j(\Theta)e(\Theta)\} + \xi$, where $r_j(\theta) = c_j[1 + d_j\cos k_j(\theta - \phi_j)]$, A_j is a rotated anisotropic linear transformation, and $\xi \sim N(0, 0.015^2 I_2)$. Implementation details are provided in Appendix E.2. We consider the following three settings:

(i) *Five-star.* Set $k_1 = 5$ and use the corresponding parameters (c_1, d_1, ϕ_1, A_1).

(ii) *Clover.* Set $k_2 = 4$ with a different choice of (c_2, d_2, ϕ_2, A_2).

(iii) *Exp.* Let $R_3 \sim \mathrm{Gamma}(2,1)$, $\mu_3 = \mathbb{E}\{R_3 r_3(\Theta)e(\Theta)\}$ and $\varepsilon = A_3\{R_3 r_3(\Theta)e(\Theta) - \mu_3\}$.

Figure 2 shows that all methods achieve marginal coverage close to the nominal level, while MV-OTCP consistently attains the smallest mean volume and symmetric difference from the oracle set. To isolate the effect of the learned metric M, we focus on setting (i) and vary the anisotropy through $A_1 = A_1(\rho)$, with larger ρ producing more elongated residual distributions while preserving the overall area scale; see Appendix D.1 for details. Figure 3 shows that MV-OTCP consistently achieves smaller volume than FOTCP, and the advantage grows with anisotropy. These improvements demonstrate the benefit of learning a volume-aware transport geometry.

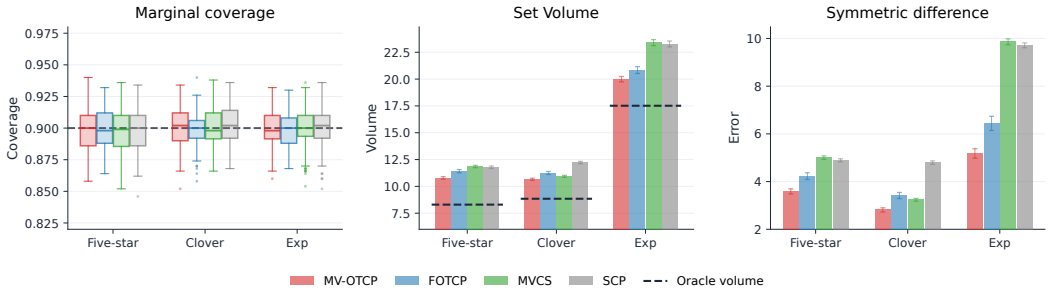


Figure 2: Homogeneous settings. Dashed lines mark nominal coverage and setting-specific oracle minimum volumes.

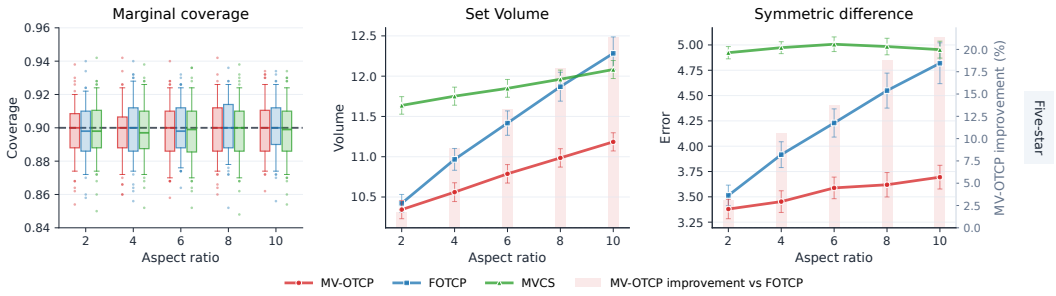


Figure 3: Aspect-ratio ablation for the five-star setting with aspect ratio $\rho \in \{2, 4, 6, 8, 10\}$. Translucent bars report the mean paired improvement of MV-OTCP over FOTCP on the right axis.

5.2 Heterogeneous simulations

We generate $Y = m(X) + \varepsilon_X$ with $X \sim \mathrm{Unif}[-2, 2]$. Let $Z \sim N(0, I_2)$, $U \sim \mathrm{Unif}(0, 1)$, and $\Theta \sim \mathrm{Unif}(0, 2\pi)$ be mutually independent. Implementation details are deferred in Appendix E.3. Consider the following three settings:

(i) *X-mixture.* With $C \mid X = x \sim \mathrm{Categorical}\{p(x)\}$ and $\bar{v}(x) = \sum_{k=1}^{4} p_k(x) v_k$, define $\varepsilon_x = R\{\theta_{\mathrm{m}}(x)\}[a_{\mathrm{m}}(x)\{v_C - \bar{v}(x)\} + \tau(x)Z]$.

(ii) *X-clover.* Let $\varepsilon_x = A_{\mathrm{c}}(x)[\sqrt{U}\, q_x(\Theta) e(\Theta)] + 0.020Z$. The radial boundary q_x varies the lobe strength and phase, while $A_{\mathrm{c}}(x)$ changes orientation, scale, and anisotropy.

(iii) *X-Exp.* For independent $E_1, E_2 \sim \mathrm{Exp}(1)$, define $\varepsilon_X = r(X)R\{\theta_{\mathrm{e}}(X)\}E$ with $E = (E_1, E_2)^\top$. The positive scale $r(x)$ and orientation $\theta_{\mathrm{e}}(x)$ both vary with x.

In Figure 4, all methods maintain near-nominal marginal coverage. IMV-OTCP attains the smallest volume and lowest root MSCE across all settings, with particularly substantial improvements in the *Mixture* and *Exp* settings, while also achieving competitive worst-case coverage. Appendix D.2 reports an ablation study that jointly examines the effect of the learned metric and the number of anchors, by comparing IMV-OTCP with its $M = I$ counterpart. Together, the main and ablation results show that local adaptation supplies the overall heterogeneous fit, while learning the local ground cost yields an additional reduction in volume.

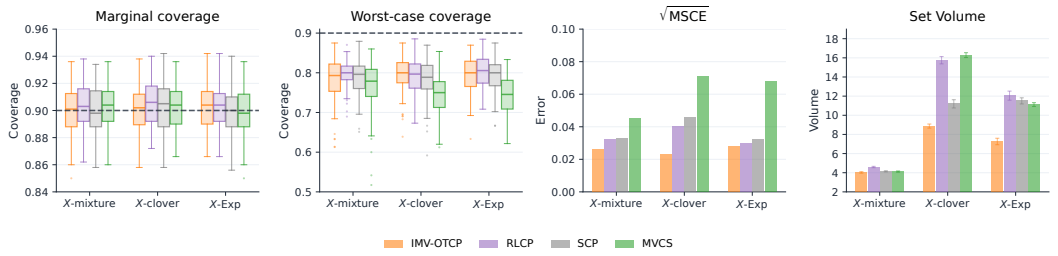


Figure 4: Heterogeneous settings. Boxplots show marginal and worst-case coverage; bars show root MSCE and set volume. Dashed lines mark nominal coverage.

5.3 Real-data regression benchmarks

We use the `Energy`, `Bias`, `Taxi`, `rf1`, and `scm1d` datasets, with response dimensions $2, 2, 2, 8$, and 16, respectively, and follow the preprocessing protocol of Dheur et al. (2025); Braun et al. (2026). We further project the transformed covariates onto a whitened PCA representation with at most eight components, which reduces the cost of localized computation and yields a better-conditioned representation for neighborhood distances. All transformations are fitted only on the training split and then held fixed. Additional implementation details can be found in Appendix E.4.

Table 1: Regression benchmarks over 100 splits. Entries are means (standard errors). Root volume is the median of $V_i^{1/d}$. Bold marks the lowest root volume in each dataset.

Metric	Dataset	SCP	MVCS	RLCP	IMV-OTCP
Marginal coverage	`Energy`	0.909(0.004)	0.910(0.004)	0.912(0.003)	0.917(0.004)
	`Bias`	0.901(0.002)	0.901(0.002)	0.903(0.002)	0.901(0.002)
	`Taxi`	0.901(0.001)	0.902(0.001)	0.903(0.001)	0.900(0.001)
	`rf1`	0.900(0.001)	0.901(0.001)	0.901(0.001)	0.900(0.002)
	`scm1d`	0.898(0.001)	0.899(0.001)	0.901(0.001)	0.897(0.002)
$\sqrt{\text{MSCE}}$	`Energy`	0.115(0.0039)	0.101(0.0028)	0.104(0.0032)	0.091(0.0022)
	`Bias`	0.044(0.0008)	0.042(0.0008)	0.042(0.0008)	0.042(0.0008)
	`Taxi`	0.033(0.0004)	0.029(0.0010)	0.025(0.0004)	0.029(0.0006)
	`rf1`	0.065(0.0007)	0.061(0.0008)	0.051(0.0006)	0.060(0.0007)
	`scm1d`	0.070(0.0005)	0.073(0.0007)	0.054(0.0007)	0.065(0.0006)
Root volume	`Energy`	1.474(0.021)	1.337(0.030)	1.867(0.025)	**1.107(0.019)**
	`Bias`	2.159(0.007)	2.177(0.008)	2.149(0.007)	**1.779(0.007)**
	`Taxi`	3.462(0.010)	3.492(0.090)	3.563(0.009)	**2.703(0.008)**
	`rf1`	1.720(0.006)	**1.286(0.062)**	2.088(0.006)	1.309(0.005)
	`scm1d`	2.025(0.007)	4.266(0.616)	2.878(0.008)	**1.759(0.004)**

Table 1 shows that all methods remain close to the nominal marginal coverage level. IMV-OTCP achieves the smallest volume on all datasets except `rf1`, while also attaining competitive $\sqrt{\text{MSCE}}$ relative to the localized methods. These results indicate that IMV-OTCP improves prediction-set efficiency without sacrificing marginal coverage. Its performance on the conditional diagnostic further suggests that the individualized construction captures meaningful local variation.

6 Discussion

Learning the OT ground cost provides a unified framework for shaping multivariate conformal regions toward either minimum volume or downstream task objectives. MV-OTCP uses a global metric in a full-conformal assignment, while IMV-OTCP fits local transport radii and uses split calibration; both separate conformal calibration from geometric optimization, yielding finite-sample marginal validity and, under additional consistency conditions, oracle-region recovery and asymptotic conditional coverage. Experiments demonstrate the benefits of global metric learning, local adaptation under heterogeneous uncertainty, and task-aware optimization.

7 AI USE STATEMENT

We used AI-assisted tools for literature retrieval and discovery, primarily to identify potentially relevant prior work and related references. The authors independently verified the relevance and accuracy of all retrieved materials and were responsible for the final selection, interpretation, and citation of the literature. AI tools were not relied upon for establishing the technical results or validating the scientific claims presented in this paper.

REFERENCES

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J Zico Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

Yajie Bao, Yang Hu, Haojie Ren, Peng Zhao, and Changliang Zou. Optimal model selection for conformalized robust optimization. *Journal of the American Statistical Association*, 2026a. Accepted.

Yajie Bao, Chuchen Zhang, Zhaojun Wang, Haojie Ren, and Changliang Zou. Shape-adaptive conditional calibration for conformal prediction via minimax optimization. *arXiv preprint arXiv:2603.23374*, 2026b.

Dimitris Bertsimas and Yu Ma. Robust regression over averaged uncertainty. *arXiv preprint arXiv:2311.06960*, 2023.

Sacha Braun, Eugène Berta, Michael I Jordan, and Francis Bach. Multivariate conformal prediction via conformalized Gaussian scoring. *arXiv preprint arXiv:2507.20941*, 2025.

Sacha Braun, Liviu Aolaritei, Michael I Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *Journal of the American Statistical Association*, 2026. doi: 10.1080/01621459.2026.2692078.

Shreyas Chaudhari, Srinivasa Pranav, and José M. F. Moura. GradNetOT: Learning optimal transport maps with GradNets. In *2025 IEEE 10th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2025. doi: 10.1109/CAMSAP66162.2025.11423958.

Abhilash Reddy Chenreddy, Nymisha Bandi, and Erick Delage. Data-driven conditional robust optimization. *Advances in Neural Information Processing Systems*, 35:9525–9537, 2022.

Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256, 2017. doi: 10.1214/16-AOS1450. URL https://doi.org/10.1214/16-AOS1450.

Santiago Cortes-Gomez, Carlos Miguel Patiño, Yewon Byun, Steven Wu, Eric Horvitz, and Bryan Wilder. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=iOMnn1hSBO.

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

Victor Dheur, Matteo Fontana, Yorick Estievenart, Naomi Desobry, and Souhaib Ben Taieb. A unified comparative study with generalized conformity scores for multi-output conformal regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 13444–13485. PMLR, 13–19 Jul 2025.

Zhenhan Fang, Aixin Tan, and Jian Huang. CONTRA: Conformal prediction region via normalizing flow transformation. In *International Conference on Learning Representations*, 2025.

Shai Feldman, Stephen Bates, and Yaniv Romano. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48, 2023.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Marc Hallin, Eustasio del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139 – 1165, 2021. doi: 10.1214/20-AOS1996. URL https://doi.org/10.1214/20-AOS1996.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87 (2):549–578, 2025. doi: 10.1093/jrsssb/qkae103.

Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. CD-split and HPD-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022.

Chancellor Johnstone and Bruce Cox. Conformal uncertainty sets for robust optimization. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pp. 72–90. PMLR, 2021.

Shayan Kiyani, George J Pappas, and Hamed Hassani. Length optimization in conformal prediction. *Advances in Neural Information Processing Systems*, 37:99519–99563, 2024.

Michal Klein, Louis Bethune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport. *arXiv preprint arXiv:2502.03609*, 2025.

Vladimir Kondratyev, Alexander Fishkov, Mahmoud Hegazy, Nikita Kotelevskii, Rémi Flamary, Maxim Panov, and Eric Moulines. Neural optimal transport meets multivariate conformal prediction. In *International Conference on Learning Representations*, pp. 77879–77917, 2026.

Ruiting Liang, Wanrong Zhu, and Rina Foygel Barber. Conformal prediction after data-dependent model selection. *Journal of the American Statistical Association*, 2026. doi: 10.1080/01621459.2026.2663588.

Ashok Vardhan Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason D. Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6672–6681. PMLR, 2020.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Ellipsoidal conformal inference for multi-target regression. In *Proceedings of the Eleventh Symposium on Conformal and Probabilistic Prediction with Applications*, volume 179 of *Proceedings of Machine Learning Research*, pp. 294–306. PMLR, 2022.

Eugene Ndiaye. Beyond uncertainty sets: Leveraging optimal transport to extend conformal predictive distribution to multivariate settings. *arXiv preprint arXiv:2511.15146*, 2025.

Chunlin Sun, Linyu Liu, and Xiaocheng Li. Predict-then-calibrate: A new perspective of robust contextual LP. *Advances in Neural Information Processing Systems*, 36:17713–17741, 2023.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59509–59527. PMLR, 2025.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. doi: 10.1007/b106715.

Congbin Xu, Yue Yu, Zhaojun Wang, Changliang Zou, and Haojie Ren. Aggregating conformal prediction sets via α-allocation. *Biometrika*, 113(3):asag048, 2026. doi: 10.1093/biomet/asag048.

Yachong Yang and Arun Kumar Kuchibhotla. Selection and aggregation of conformal prediction sets. *Journal of the American Statistical Association*, 120(549):435–447, 2025. doi: 10.1080/01621459.2024.2344700.

Christopher Yeh, Nicolas Christianson, Alan Wu, Adam Wierman, and Yisong Yue. End-to-end conformal calibration for optimization under uncertainty. *arXiv preprint arXiv:2409.20534*, 2024.

A IMPLEMENTATION DETAILS

A.1 ELLIPTICAL VOLUME-OPTIMAL GEOMETRY

Suppose that the score density is elliptically symmetric, with $f_P(s) = |\Sigma|^{-1/2} g(\|\Sigma^{-1/2}(s-m)\|)$, where $\Sigma \succ 0$, $m \in \mathbb{R}^d$, and g is nonincreasing, and that the reference distribution $\mathcal{Q}$ is radial, with $U = R\Theta$, $R \sim \mathrm{Unif}(0,1)$, and Θ uniformly distributed on $\{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ independently of R. Let $r(s) = \|\Sigma^{-1/2}(s-m)\|$, $\rho = r(S)$, and let F_ρ be the distribution function of ρ. For a radial reference law with uniform radial component, define

$$T_{M^\star}(s) = F_\rho(r(s)) \frac{\Sigma^{-1/2}(s-m)}{r(s)}$$

for $r(s) > 0$. Then $(T_{M^\star})_\# \mathcal{P} = \mathcal{Q}$, and

$$T_{M^\star}^{-1}(\mathbb{B}_{1-\alpha}) = \{s : r(s) \le F_\rho^{-1}(1-\alpha)\}.$$

This is the central density superlevel ellipsoid and is therefore the minimum-volume measurable set with probability $1-\alpha$. Moreover, by Lemma 3.1, $T_{M^\star}$ is generated by the metric $M^\star_{\mathrm{vol}} \propto \Sigma^{-1/2}$. Indeed, set $A = \Sigma^{-1/4}$, $\tilde{m} = Am$, and $h(r) = \int_0^r F_\rho(t)\,dt$. For $\tilde{s} \in \mathbb{R}^d$, define $\tilde{r}(\tilde{s}) = \|A(\tilde{s} - \tilde{m})\|$ and $\psi(\tilde{s}) = h\{\tilde{r}(\tilde{s})\}$. Since F_ρ is nondecreasing, h is convex and nondecreasing; hence ψ is convex. Differentiating with respect to $\tilde{s}$ and evaluating at $\tilde{s} = As$ gives

$$\nabla\psi(As) = F_\rho(r(s)) \frac{A^3(s-m)}{r(s)} = A T_{M^\star}(s).$$

Since $A = M^{1/2}$ when $M = \Sigma^{-1/2}$, Lemma 3.1 implies that $T_{M^\star}$ is optimal for c_M with $M = \Sigma^{-1/2}$.

A.2 IMPLEMENTATION DETAILS OF MV-OTCP

A direct empirical approach would be to replace the population objective in (4) by its sample analogue and optimize it over M. This is generally infeasible because, apart from the metric M, the Brenier representation only ensures the existence of a convex potential g_M, which is unknown and depends on M implicitly. Thus the objective cannot be evaluated directly unless the associated transport map is estimated. We represent its gradient map directly with GradNet (Chaudhari et al., 2025) and differentiate the empirical task loss through the fitted transport problem.

Let $\mathcal{D}_{\mathrm{est}} = \{(X_i, Y_i)\}_{i=1}^N$ be an independent estimation sample with scores $S_i = Y_i - \widehat{\mu}(X_i)$, and let $U_1, \dots, U_N \sim \mathcal{Q}$ be independent reference draws. For a current metric M, transform the samples by

$$\bar{S}_i = M^{1/2} S_i, \qquad \bar{U}_i = M^{1/2} U_i.$$

Because the empirical score law is discrete, define the KDE of $(M^{1/2})_\# \mathcal{P}$

$$\widehat{f}_{P,M}(z) = \frac{1}{N} \sum_{i=1}^N \kappa_b(z - \bar{S}_i),$$

where κ_b is a score-space kernel. We use an isotropic Gaussian kernel. At each metric iterate, its bandwidth is chosen by the weighted Scott rule

$$b = \max\left\{10^{-3}, \left(\frac{1}{d}\sum_{i=1}^N \widetilde{w}_i \|\bar{S}_i - \bar{S}_w\|^2\right)^{1/2} n_{\mathrm{eff}}^{-1/(d+4)}\right\}, \qquad n_{\mathrm{eff}} = \left(\sum_{i=1}^N \widetilde{w}_i^2\right)^{-1},$$

where $\widetilde{w}_i$ are normalized localization weights and $\bar{S}_w = \sum_i \widetilde{w}_i \bar{S}_i$; global fits use $\widetilde{w}_i = 1/N$. The same rule is used for the source and target KDEs when fitting the score-to-reference map. Let Ψ_θ be a GradNet mapping the transformed reference distribution to the transformed score distribution. Following Chaudhari et al. (2025), we fit Ψ_θ through the Monge–Ampère density identity:

$$\widehat{\mathcal{L}}_{\mathrm{GN}}^{\leftarrow}(M, \theta) = \frac{1}{2N} \sum_{j=1}^N \Big[\log\det\{\nabla\Psi_\theta(M^{1/2}U_j)\} - \log f_Q(U_j) + \log \widehat{f}_{P,M}\{\Psi_\theta(M^{1/2}U_j)\}\Big]^2. \tag{10}$$

Our parameterization adds a positive linear skip connection and a bias to a positive combination of J GradNet modules. Each module has one hidden projection of width H, a weighted-softmax activation, and the tied transpose projection required by the gradient-map construction. The experiment-specific values of (J, H) are reported below. For a fixed M, let $\widehat{\theta}(M)$ minimize (10). We rewrite the empirical volume objective (6) here for convenience:

$$\widehat{\mathcal{J}}(M,\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}\{U_i \in \mathbb{B}_{1-\alpha}\}\det\{\nabla\Psi_\theta(M^{1/2}U_i) + \delta_N I_d\}}{f_Q(U_i)},$$

where $\delta_N > 0$ is a small numerical stabilizer. The empirical metric-learning problem is

$$\widehat{M} \in \operatorname*{arg\,min}_{M\in\mathcal{M}_0} \widehat{\mathcal{J}}\{M, \widehat{\theta}(M)\}.$$

Importantly, this is a profiled objective because the fitted transport map depends on the metric. Its total derivative is

$$\nabla_M \widehat{\mathcal{J}}\{M,\widehat{\theta}(M)\} = \underbrace{\partial_M \widehat{\mathcal{J}}(M,\theta)\Big|_{\theta=\widehat{\theta}(M)}}_{\text{direct metric derivative}} + \underbrace{\{\nabla_M\widehat{\theta}(M)\}^\top\ \nabla_\theta \widehat{\mathcal{J}}(M,\theta)\Big|_{\theta=\widehat{\theta}(M)}}_{\text{transport-response derivative}}.$$

The second term is generally nonzero because the GradNet is fitted by the OT loss rather than the volume loss. If $\widehat{\theta}(M)$ is a locally stationary minimizer of (10) and the parameter Hessian is nonsingular, implicit differentiation gives

$$\nabla_M\widehat{\theta}(M) = -\left[\nabla^2_{\theta\theta}\widehat{\mathcal{L}}^{\leftarrow}_{\mathrm{GN}}\{M,\widehat{\theta}(M)\}\right]^{-1}\nabla^2_{\theta M}\widehat{\mathcal{L}}^{\leftarrow}_{\mathrm{GN}}\{M,\widehat{\theta}(M)\}.$$

In computation, the linear system can be damped and solved by conjugate gradients; differentiating through a finite number of inner updates is an alternative.

To enforce the metric constraints, define

$$\mathcal{B}_0 = \left\{B = B^\top : \mathrm{tr}(B) = 0,\ \log(\underline{\lambda})I_d \preceq B \preceq \log(\overline{\lambda})I_d\right\}, \qquad M(B) = \exp(B).$$

Then $M(B) \in \mathcal{M}_0$. The spectral projection may be replaced in practice by a condition-number penalty, provided that the returned metric satisfies the localization condition used in the theory. The resulting bilevel metric-learning procedure is summarized in Algorithm 1.

Algorithm 1 Volume-based metric learning for MV-OTCP

Require: Estimation scores $S_1,\dots,S_N$, reference draws $U_1,\dots,U_N$, kernel bandwidth b, initial parameter $B_0 \in \mathcal{B}_0$, outer iterations T, inner steps K

1: **for** $t = 0,\dots,T-1$ **do**
2: Set $M_t = \exp(B_t)$.
3: Transform samples: $\bar{S}_i = M_t^{1/2}S_i$ and $\bar{U}_i = M_t^{1/2}U_i$.
4: Form the transformed-score KDE $\widehat{f}_{P,M_t}$.
5: Starting from warm-started θ, take K steps minimizing $\widehat{\mathcal{L}}^{\leftarrow}_{\mathrm{GN}}(M_t,\theta)$ in (10); denote the result by θ_t.
6: Evaluate $\widehat{\mathcal{J}}(M_t,\theta_t)$ using (6).
7: Compute the transport-response derivative by implicit differentiation through the fitted GradNet.
8: Update $B_{t+1} = \Pi_{\mathcal{B}_0}\left[B_t - \gamma_t\nabla_B\widehat{\mathcal{J}}\{M(B),\widehat{\theta}(M(B))\}|_{B=B_t}\right]$.
9: **end for**

Ensure: Learned determinant-one metric $\widehat{M} = M(B_T)$

A.3 General task-aware OTCP

We rewrite objective function here for convinence. For a decision $z \in \mathcal{Z}$ and an integrable loss density $\ell(z,y)$, define

$$\mathcal{L}(z,\mathcal{C}(x)) = \int_{\mathcal{C}(x)} \ell(z,y)\,dy, \qquad z^\star_M(x) \in \operatorname*{arg\,min}_{z\in\mathcal{Z}} \mathcal{L}(z,\mathcal{C}_M(x)). \tag{11}$$

The corresponding task-aware target for metric learning is

$$M^\star_{\text{task}} \in \arg\min_{M\in\mathcal{M}_0} \mathcal{J}_{\text{task}}(M), \qquad \mathcal{J}_{\text{task}}(M) = \mathbb{E}_X \mathcal{L}\{z^\star_M(X), \mathcal{C}_M(X)\}.$$

Using the same change-of-variables argument as in Section 3, the task loss can be rewritten as

$$\mathcal{J}_{\text{task}}(M) = \mathbb{E}_{X,U}\left[\ell\{z^\star_M(X), \widehat{\mu}(X) + G_M(U)\}\frac{\mathbb{1}\{U \in \mathbb{B}_{1-\alpha}\}\det\{\nabla G_M(U)\}}{f_Q(U)}\right].$$

Volume is recovered as the special case $\ell \equiv 1$, whereas a general task loss additionally accounts for where the accepted volume is placed in the response space.

Example A.1 (Integrated decision loss)**.** *The criterion* (11) *aggregates loss over the uncertainty set, as in averaged robust optimization (Bertsimas & Ma, 2023; Cortes-Gomez et al., 2025). It is an unnormalized integral, so both region size and location affect metric selection.*

Example A.2 (Smoothed contextual robust optimization)**.** *The contextual robust optimization (Chenreddy et al., 2022) objective* $\inf_{z\in\mathcal{Z}} \sup_{y\in\mathcal{C}(x)} \ell(z, y)$ *can be smoothed by the log-exponential transform*

$$\mathcal{L}_\tau(z, \mathcal{C}(x)) = \tau \log \int_{y\in\mathcal{C}(x)} \exp\{\ell(z, y)/\tau\}\, dy, \qquad \tau > 0.$$

For fixed τ*, minimizing above equation over* z *is equivalent to minimizing an integral of the form* (11) *with transformed loss density* $\ell^{\exp}_\tau(z, y) = \exp\{\ell(z, y)/\tau\}$*. As* $\tau \downarrow 0$*, under mild regularity conditions—e.g.,* $\mathcal{C}(x)$ *is compact and* $\ell(z, \cdot)$ *is continuous, the log-exponential objective approaches the worst-case loss.*

A.3.1 Empirical task-aware criterion

Let $\widehat{G}_M$ be a reference-to-score transport-map estimate obtained from the estimation sample, and define

$$\widehat{w}_M(u) = \frac{\mathbb{1}\{u \in \mathbb{B}_{1-\alpha}\}|\det \nabla_u \widehat{G}_M(u)|}{f_Q(u)}.$$

For estimation covariates $X_1, \ldots, X_N$ and a common independent reference sample $U_1, \ldots, U_K \sim \mathcal{Q}$, define

$$\widehat{\mathcal{L}}_i(z; M) := \frac{1}{K}\sum_{j=1}^{K} \widehat{w}_M(U_j)\ell\left\{z, \widehat{\mu}(X_i) + \widehat{G}_M(U_j)\right\}.$$

All K transported points are integrated jointly for each decision; optimizing a separate decision at every quadrature point would define a different, generally degenerate objective. Let $\widehat{z}^\star_i(M) \in \arg\min_{z\in\mathcal{Z}} \widehat{\mathcal{L}}_i(z; M)$, then

$$\widehat{M}_{\text{task}} \in \arg\min_{M\in\mathcal{M}_0} \widehat{\mathcal{J}}_{\text{task}}(M), \quad \widehat{\mathcal{J}}_{\text{task}}(M) := \frac{1}{N}\sum_{i=1}^{N} \widehat{\mathcal{L}}_i\{\widehat{z}^\star_i(M); M\}.$$

The estimated metric $\widehat{M}_{\text{task}}$ can be used in the full-conformal assignment (7), with the coverage guarantee of Theorem 3.1. For computational efficiency in the task experiments, we instead fit a score-to-reference map $\widehat{T}_{\widehat{M}_{\text{task}}}$ before calibration and calibrate its radius by the ordinary split-conformal rule, as in IMV-OTCP. This gives Task-Aware OT Conformal Prediction (TA-OTCP):

$$\widehat{\mathcal{C}}^{\text{TA}}(x) = \left\{y : \|\widehat{T}_{\widehat{M}_{\text{task}}}\{s(x, y)\}\| \le \widehat{q}_{1-\alpha}\left(\left\{\|\widehat{T}_{\widehat{M}_{\text{task}}}\{s(X_i, Y_i)\}\|\right\}_{i=1}^{n} \cup \{\infty\}\right)\right\}.$$

Both the metric and forward map are fixed before calibration, so exchangeability of the calibration and test scalar scores gives the usual split-conformal marginal guarantee.

Now consider the computation details. For an arbitrary pair (M, θ), define the transported point in the original score coordinates and its stabilized Jacobian weight by

$$\widehat{G}_{M,\theta}(u) := M^{-1/2}\Psi_\theta(M^{1/2}u), \tag{12}$$

$$w_{M,\theta}(u) := \frac{\mathbb{1}\{u \in \mathbb{B}_{1-\alpha}\}}{f_Q(u)}\det\{\nabla\Psi_\theta(M^{1/2}u) + \delta_N I_d\}. \tag{13}$$

With $X_1, \ldots, X_N$ and $U_1, \ldots, U_K$, the parameterized version is

$$\widehat{\mathcal{L}}_i(z; M, \theta) = \frac{1}{K} \sum_{j=1}^{K} w_{M,\theta}(U_j) \ell\{z, \widehat{\mu}(X_i) + \widehat{G}_{M,\theta}(U_j)\}.$$

Let $\widehat{z}_i^\star(M, \theta)$ minimize this criterion over $z \in \mathcal{Z}$, and set

$$\widehat{\mathcal{J}}_{\text{task}}(M, \theta) = \frac{1}{N} \sum_{i=1}^{N} \widehat{\mathcal{L}}_i\{\widehat{z}_i^\star(M, \theta); M, \theta\}.$$

The implemented outer problem is

$$\widehat{M}_{\text{task}} \in \operatorname*{arg\,min}_{M \in \mathcal{M}_0} \widehat{\mathcal{J}}_{\text{task}}\{M, \widehat{\theta}(M)\},$$

where $\widehat{\theta}(M)$ is obtained from (10). This is a bilevel problem because both the transported points and their Jacobian weights depend on the fitted map.

The metric derivative must retain this dependence. In particular,

$$\nabla_M \widehat{G}_{M,\widehat{\theta}(M)}(u) = \left\{ \partial_M \widehat{G}_{M,\theta}(u) + \partial_\theta \widehat{G}_{M,\theta}(u) \nabla_M \widehat{\theta}(M) \right\} \Big|_{\theta = \widehat{\theta}(M)}, \tag{14}$$

and the same decomposition applies to $w_{M,\widehat{\theta}(M)}(u)$. The direct term includes the derivatives of $M^{-1/2}$ and of the input $M^{1/2}u$; the response term accounts for the change in the fitted map. It can be evaluated by the implicit system in Appendix A.2 or by differentiating through the inner optimization.

Let $\widehat{y}_{ij} = \widehat{\mu}(X_i) + \widehat{G}_{M,\widehat{\theta}(M)}(U_j)$, $\widehat{w}_j = w_{M,\widehat{\theta}(M)}(U_j)$, and $\widehat{z}_i^\star = \widehat{z}_i^\star\{M, \widehat{\theta}(M)\}$. If $\widehat{z}_i^\star$ is an exact interior solution and the required derivatives exist, the envelope theorem gives

$$\nabla_M \widehat{\mathcal{J}}_{\text{task}}\{M, \widehat{\theta}(M)\} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} \left[\ell(\widehat{z}_i^\star, \widehat{y}_{ij}) \nabla_M \widehat{w}_j + \widehat{w}_j \nabla_y \ell(\widehat{z}_i^\star, \widehat{y}_{ij})^\top \nabla_M \widehat{y}_{ij} \right]. \tag{15}$$

The terms involving $\nabla_M \widehat{z}_i^\star$ cancel only after summing over j, because $\nabla_z \widehat{\mathcal{L}}_i(\widehat{z}_i^\star; M, \widehat{\theta}) = 0$. Thus the envelope theorem does not permit the fitted transport map to be held fixed. For a constrained convex task, the lower-level KKT system may instead be differentiated at regular solutions (Agrawal et al., 2019).

For smoothed CRO, define $\widehat{\mathcal{L}}_i^{\exp}$ by replacing ℓ with $\exp(\ell/\tau)$ in the preceding quadrature. The profiled metric objective is

$$\widehat{\mathcal{J}}_\tau(M) = \frac{1}{N} \sum_{i=1}^{N} \tau \log \widehat{\mathcal{L}}_i^{\exp}\{\widehat{z}_i^\star(M); M, \widehat{\theta}(M)\}.$$

For regular solutions, its total derivative weights each profiled exponential-integral derivative by $\tau / \widehat{\mathcal{L}}_i^{\exp}$. The logarithm can be omitted when finding each context's decision, but not when averaging over contexts to learn the metric. The transport-response terms in (14) remain present.

Example A.3 (Quadratic task with a closed-form decision)**.** *Let $H = H^\top \succ 0$, $\mathcal{Z} = \mathbb{R}^d$, and*

$$\ell_H(z, y) = \frac{1}{2}(z - y)^\top H (z - y).$$

Writing $\widehat{y}_{ij}(M) = \widehat{\mu}(X_i) + \widehat{G}_M(U_j)$ and $\widehat{w}_j(M) = \widehat{w}_M(U_j)$, the empirical decision is the weighted centroid

$$\widehat{z}_i^\star(M) = \frac{\sum_{j=1}^{K} \widehat{w}_j(M) \widehat{y}_{ij}(M)}{\sum_{j=1}^{K} \widehat{w}_j(M)}.$$

The decision is well defined whenever the denominator is positive. Thus the downstream optimization is available in closed form, although the task criterion still depends on the metric through the fitted transport map and its Jacobian.

A.4 Implementation details of IMV-OTCP

IMV-OTCP differs at the map-fitting stage: for each fixed x, it first parameterizes a determinant-one metric by an unconstrained matrix A_x,

$$B_x = \frac{A_x + A_x^\top}{2} - \frac{\mathrm{tr}(A_x)}{d} I_d, \quad M_x = \exp(B_x).$$

Then $B_x = {B_x}^\top$, $\mathrm{tr}(B_x) = 0$, $M_x \succ 0$, and $\det(M_x) = 1$. For a nonnegative covariate-localization kernel $K_{\tilde{h}}$ with bandwidth $\tilde{h}$, set $w_{i,\tilde{h}}(x) = K_{\tilde{h}}(X_i, x) / \sum_{k=1}^N K_{\tilde{h}}(X_k, x)$. The local empirical law is discrete, whereas GradNetOT is formulated for continuous source and target densities. We therefore use the weighted KDE of $(M^{1/2})_\# \mathcal{P}_x$:

$$\widehat{f}_{P,x,M}(z) = \sum_{i=1}^N w_{i,\tilde{h}}(x) \kappa_b(z - M^{1/2} S_i), \qquad \kappa_b(v) = b^{-d} \kappa(v/b).$$

Its score-space bandwidth b is distinct from the covariate-localization bandwidth $\tilde{h}$. Let $\Psi_{\theta,x}$ be a GradNet that maps the transformed reference distribution to this transformed local score distribution. Following GradNetOT (Chaudhari et al., 2025), we fit it through the Monge–Ampère density identity:

$$\widehat{\mathcal{L}}_x^{\leftarrow}(M, \theta_x) = \frac{1}{2N} \sum_{j=1}^N \Big[\log\det\{\nabla \Psi_{\theta,x}(M^{1/2} U_j)\} - \log f_Q(U_j) + \log \widehat{f}_{P,x,M}\{\Psi_{\theta,x}(M^{1/2} U_j)\} \Big]^2. \tag{16}$$

This loss directly parameterizes the monotone gradient map and matches its log-Jacobian determinant to the log density ratio. For a fixed M, let $\widehat{\theta}_x(M)$ minimize (16). The fitted map in the original coordinates is

$$\widehat{G}_{M,x}(u) = M^{-1/2} \Psi_{\widehat{\theta}_x(M),x}(M^{1/2} u). \tag{17}$$

The empirical counterpart of (8) is

$$\widehat{\mathcal{J}}_x(M, \theta_x) = \frac{1}{N} \sum_{j=1}^N \frac{\mathbb{1}\{U_j \in \mathbb{B}_{1-\alpha}\} \det\left\{\nabla \Psi_{\theta,x}(M^{1/2} U_j) + \delta_N I_d\right\}}{f_Q(U_j)}.$$

The pointwise metric estimate is

$$\widehat{M}_x \in \operatorname*{arg\,min}_{M \in \mathcal{M}_0} \widehat{\mathcal{J}}_x\{M, \widehat{\theta}_x(M)\}.$$

Because the fitted reference-to-score map depends on M_x through the profiled GradNetOT objective, we compute the total derivative of the empirical volume by implicitly differentiating through the first-order optimality conditions defining $\widehat{\theta}_x(M_x)$, following the implementation used for MV-OTCP in Appendix A.2. Thus, the metric update accounts for the response of the fitted transport map to a change in M_x.

After selecting $\widehat{M}_x$, fit a second GradNet F_{ψ_x} directly from the transformed local score law to the transformed reference law by minimizing

$$\widehat{\mathcal{L}}_{x,\mathrm{MA}}^{\rightarrow}(\psi_x) = \frac{1}{2} \sum_{i=1}^N w_{i,\tilde{h}}(x) \Bigg[\log\det\left\{\nabla F_{\psi_x}(\widehat{M}_x^{1/2} S_i)\right\} - \log \widehat{f}_{P,x,\widehat{M}_x}(\widehat{M}_x^{1/2} S_i) + \log f_Q\left\{\widehat{M}_x^{-1/2} F_{\psi_x}(\widehat{M}_x^{1/2} S_i)\right\} \Bigg]^2. \tag{18}$$

The original Monge–Ampère objective does not explicitly enforce the learned reference-to-score and score-to-reference maps to be mutual inverses. We therefore introduce inverse- and cycle-consistency

losses to encourage this property. In transformed coordinates, write $\bar{G}_x$ for the frozen reference-to-score GradNet obtained during metric learning, $\bar{T}_{\psi,x}$ for the score-to-reference GradNet being fitted, $\bar{U}_j = \widehat{M}_x^{1/2} U_j$, and $\bar{S}_i = \widehat{M}_x^{1/2} S_i$. We use

$$\widehat{\mathcal{L}}_{x,\mathrm{inv}}(\psi) = \frac{1}{N}\sum_{j=1}^{N} \|\bar{T}_{\psi,x}\{\bar{G}_x(\bar{U}_j)\} - \bar{U}_j\|_2^2,$$

$$\widehat{\mathcal{L}}_{x,\mathrm{cyc}}(\psi) = \sum_{i=1}^{N} w_{i,\tilde{h}}(x)\|\bar{G}_x\{\bar{T}_{\psi,x}(\bar{S}_i)\} - \bar{S}_i\|_2^2,$$

and fit the final map with

$$\widehat{\mathcal{L}}_{x,T}(\psi) = \widehat{\mathcal{L}}_{x,\mathrm{inv}}(\psi) + \widehat{\mathcal{L}}_{x,\mathrm{cyc}}(\psi) + \widehat{\mathcal{L}}^{\rightarrow}_{x,\mathrm{MA}}(\psi), \tag{19}$$

Let $\widehat{\psi}_x$ minimize (19). The fitted score-to-reference map is

$$\widehat{T}_{\widehat{M}_x,x}(s) = \widehat{M}_x^{-1/2} F_{\widehat{\psi}_x}(\widehat{M}_x^{1/2} s).$$

The empirical procedure is summarized in Algorithm 2.

Algorithm 2 Individualized volume-based metric learning for IMV-OTCP

Require: Independent training, estimation, and calibration samples; common reference draws; kernel $K_{\tilde{h}}$; level α

1: Fit the score function s on the training sample and compute the estimation scores.
2: **for** each calibration covariate X_i **do**
3: Form the weighted local estimation law and KDE at X_i.
4: **for** each outer metric-learning iteration **do**
5: Fit the reference-to-score GradNet by minimizing (16).
6: Update M through the profiled local task objective.
7: **end for**
8: Fix $\widehat{M}_{X_i}$ and fit the score-to-reference GradNet by minimizing (19).
9: Compute $\widehat{R}_N(X_i, Y_i)$ using the fitted map.
10: **end for**
11: Compute the global split-conformal cutoff in (9).
12: At a test covariate x, learn $\widehat{M}_x$ and construct $\widehat{T}_{\widehat{M}_x,x}$ by the same pointwise rule.

Ensure: The set (9); reuse $(\widehat{M}_x, \widehat{T}_{\widehat{M}_x,x})$ across all trial responses at that x.

Remark A.1. *A full-conformal extension of the MV-OTCP construction to localized transport is provided in Appendix C. We use the split-conformal version in* (9) *for computational efficiency, since the full construction requires repeated sample-dependent Kantorovich solves across calibration points, test points, and trial responses. The split construction fixes the localized score rule before calibration, enabling anchor-based amortization of the pointwise transport fits.*

Remark A.2 (Anchor-point implementation for localized OT methods)**.** *IMV-OTCP fit local transport objects at every covariate value. To reduce computation, we select a finite set of anchor covariates* $\{a_1, \ldots, a_L\}$ *using only the training and estimation samples, and fit* $(\widehat{M}_{a_j}, \widehat{T}_{\widehat{M}_{a_j},a_j})$ *once for each anchor. For a generic covariate value* x*, we use the fit associated with its nearest anchor,* $j(x) = \arg\min_{j\in[L]} \|x - a_j\|$*, so that* $(\widehat{M}_x, \widehat{T}_{\widehat{M}_x,x}) = (\widehat{M}_{a_{j(x)}}, \widehat{T}_{\widehat{M}_{a_{j(x)}},a_{j(x)}})$*. This amortizes the expensive pointwise fitting step while preserving the split-conformal validity argument.*

B ADDITIONAL THEORETICAL RESULTS

B.1 POLYHEDRAL REPRESENTATION OF MV-OTCP

Fix the calibration scores and regard the trial score as a query $s \in \mathbb{R}^d$. If the query is assigned to U_j, the remaining calibration scores are optimally matched to the other reference points with value

$$\kappa_{j,M} = \min_{\substack{\sigma:[n]\to[n+1]\setminus\{j\} \\ \sigma \text{ bijective}}} \sum_{i=1}^{n} (S_i - U_{\sigma(i)})^\top M (S_i - U_{\sigma(i)}).$$

Let $f_{j,M}(s) = (s-U_j)^\top M(s-U_j) + \kappa_{j,M}$. The index j is an optimal destination for the query if and only if s belongs to the candidate cell

$$\overline{\mathcal{R}}_{j,M} = \{s : f_{j,M}(s) \le f_{k,M}(s),\ k \ne j\}.$$

When several indices are optimal, the fixed tie rule partitions the overlap into disjoint cells $\mathcal{R}^{\text{tie}}_{j,M} \subseteq \overline{\mathcal{R}}_{j,M}$.

Proposition B.1 (Polyhedral representation of MV-OTCP)**.** *For any fixed $M = M^\top \succ 0$, the candidate cell $\overline{\mathcal{R}}_{j,M}$ is the convex polyhedron*

$$\overline{\mathcal{R}}_{j,M} = \left\{ s \in \mathbb{R}^d : \langle s, M(U_k - U_j)\rangle \le \beta_{j,k}(M), \quad k \ne j \right\},$$

where $\beta_{j,k}(M) = \frac{1}{2}\left\{U_k^\top M U_k - U_j^\top M U_j + \kappa_{k,M} - \kappa_{j,M}\right\}$. Let $\mathcal{I}_\alpha$ denote the indices of the accepted reference points, $\mathcal{I}_\alpha = \left\{ j : \|U_j\| \le \widehat{q}_{1-\alpha}\left(\{\|U_i\|\}_{i=1}^{n+1}\right)\right\}$. Under any deterministic priority rule for ties, $\mathcal{R}^{\text{tie}}_{j,M}$ is a possibly half-open polyhedral cell obtained by assigning the shared boundaries of the candidate cells. Consequently, the exact accepted score region is the finite union of polyhedral cells

$$\widehat{\mathcal{A}}_M = \bigcup_{j \in \mathcal{I}_\alpha} \mathcal{R}^{\text{tie}}_{j,M},$$

For $M = \widehat{M}$, the resulting MV-OTCP set is $\widehat{\mathcal{C}}^{\text{MV}}(x) = \widehat{\mu}(x) + \widehat{\mathcal{A}}_{\widehat{M}}$.

Proof. By definition, $\overline{\mathcal{R}}_{j,M}$ is the set of query scores for which assigning the query to U_j has no larger augmented OT objective than assigning it to any other reference point. Thus

$$\overline{\mathcal{R}}_{j,M} = \{s : f_{j,M}(s) \le f_{k,M}(s),\ k \ne j\}.$$

For each $k \ne j$, the inequality $f_{j,M}(s) \le f_{k,M}(s)$ is

$$(s-U_j)^\top M(s-U_j) + \kappa_{j,M} \le (s-U_k)^\top M(s-U_k) + \kappa_{k,M}.$$

Expanding both quadratic terms gives

$$s^\top M s - 2s^\top M U_j + U_j^\top M U_j + \kappa_{j,M} \le s^\top M s - 2s^\top M U_k + U_k^\top M U_k + \kappa_{k,M}.$$

The common term $s^\top M s$ cancels. Rearranging the remaining terms gives

$$2s^\top M(U_k - U_j) \le U_k^\top M U_k - U_j^\top M U_j + \kappa_{k,M} - \kappa_{j,M},$$

which is exactly the halfspace inequality in Proposition B.1. Intersecting these finitely many halfspaces over $k \ne j$ proves that $\overline{\mathcal{R}}_{j,M}$ is a convex polyhedron.

The deterministic tie rule assigns every common boundary to exactly one minimizing index. It therefore replaces weak by strict inequalities on selected facets, producing disjoint, possibly half-open polyhedral cells $\mathcal{R}^{\text{tie}}_{j,M}$. For the accepted index set $\mathcal{I}_\alpha$, the exact accepted score region is $\bigcup_{j\in\mathcal{I}_\alpha} \mathcal{R}^{\text{tie}}_{j,M}$. Finally, the affine map $s \mapsto \widehat{\mu}(x) + s$ preserves polyhedral cells and finite unions. □

B.2 Additional theoretical results of IMV-OTCP

We state the consistency results under high-level approximation assumptions on the reverse-map objective and the final forward-map. Define

$$\zeta_{M,x}(u) := \frac{\mathbb{1}\{u \in \mathbb{B}_{1-\alpha}\}|\det\{\nabla_u G_{M,x}(u)\}|}{f_Q(u)}.$$

Then $\mathcal{J}_x(M) = \mathbb{E}_U\{\zeta_{M,x}(U)\}$. Let $\mathcal{J}_{N,x}(M) := N^{-1}\sum_{j=1}^N \zeta_{M,x}(U_j)$ be its infeasible empirical average based on the same reference draws as $\widehat{\mathcal{J}}_x$.

Assumption B.1. *Fix an interior covariate value x.*

(i) There exist constants $0 < B_x < \infty$ and $0 \le L_x < \infty$ such that $|\zeta_{M,x}(U)| \le B_x$ almost surely for all $M \in \mathcal{M}_0$. Moreover, for every $\epsilon > 0$,

$$\sup_{M_c \in \mathcal{M}_0} \mathbb{E}_U\left[\sup_{M \in \mathcal{M}_0(M_c,\epsilon)} |\zeta_{M,x}(U) - \zeta_{M_c,x}(U)|\right] \le L_x\epsilon.$$

(ii) The fitted reverse-map objective satisfies $\mathcal{E}_N^{\mathrm{obj}}(x) = o_{\mathbb{P}}(1)$ as $N \to \infty$, where $\mathcal{E}_N^{\mathrm{obj}}(x) := \sup_{M \in \mathcal{M}_0} |\widehat{\mathcal{J}}_x\{M, \widehat{\theta}_x(M)\} - \mathcal{J}_{N,x}(M)|$.

(iii) For every $\epsilon > 0$, $\inf_{M \in \mathcal{M}_0:\ \|M - M_x^\star\|_F \ge \epsilon}\{\mathcal{J}_x(M) - \mathcal{J}_x(M_x^\star)\} > 0$.

Assumption B.1(i) imposes regularity on the local objective. Part (ii) is the localized counterpart of Assumption 3.1(ii), controlling the reverse-map approximation error that enters metric learning. Part (iii) parallels Assumption 3.1(iii) and provides the standard separation condition needed for local metric consistency.

Theorem B.1. *Suppose Assumption B.1(i) holds. For fixed d, there is a universal constant $C > 0$ such that, for every $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\sup_{M \in \mathcal{M}_0} |\widehat{\mathcal{J}}_x\{M, \widehat{\theta}_x(M)\} - \mathcal{J}_x(M)| \le \mathcal{E}_N^{\mathrm{obj}}(x) + CB_x\sqrt{\frac{d^2 \log\left\{e + 2\sqrt{d}\,\overline{\lambda}L_x N/B_x\right\} + \log(2/\delta)}{N}}.$$

If, in addition, parts (ii)–(iii) of Assumption B.1 hold, then $\|\widehat{M}_x - M_x^\star\|_F \xrightarrow{\mathbb{P}} 0$.

C Weighted full-conformal localized OTCP

This appendix extends the assignment construction of Section 3 to localized transport. Localization weights enter the source marginal of an augmented Kantorovich problem, and a learned local metric yields a full-conformal alternative to the split IMV-OTCP method in Section 4. We first describe the Euclidean-cost construction, then insert the learned metric.

C.1 Euclidean-cost weighted construction

Directly centering a local procedure at the distinguished test covariate X_{n+1} destroys the symmetry of the augmented sample. Inspired by Hore & Barber (2025), we instead let $H_h(\tilde{x} \mid x)$ be a localization density and draw a randomized center $\widetilde{X}_{n+1} \mid X_{n+1} \sim H_h(\cdot \mid X_{n+1})$. For a realized center $\tilde{x}$, define

$$p_i(\tilde{x}) = \frac{H_h(\tilde{x} \mid X_i)}{\sum_{k=1}^{n+1} H_h(\tilde{x} \mid X_k)}, \qquad i \in [n+1]. \tag{20}$$

Conditional on the unordered augmented sample and $\widetilde{X}_{n+1} = \tilde{x}$, the posterior probability that observation i occupies the test position is $p_i(\tilde{x})$. The same weights therefore serve as the source masses in the localized transport problem.

For each trial response y, let $\widehat{\pi}^y_{\tilde{x}} = (\widehat{\pi}^y_{\tilde{x},ij})_{i,j\in[n+1]}$ solve

$$\begin{aligned} &\widehat{\pi}^y_{\tilde{x}} \in \arg\min_{\pi\geq 0} \sum_{i=1}^{n+1}\sum_{j=1}^{n+1} \pi_{ij}\|S^y_i - U_j\|^2, \\ &\text{subject to} \quad \sum_{j=1}^{n+1}\pi_{ij} = p_i(\tilde{x}), \quad i\in[n+1], \qquad \sum_{i=1}^{n+1}\pi_{ij} = \frac{1}{n+1}, \quad j\in[n+1]. \end{aligned} \tag{21}$$

When $p_i(\tilde{x}) = 1/(n+1)$ for every i, an optimal plan can be chosen as the scaled FOTCP assignment. For nonuniform localization weights, (21) is generally a transportation linear program and a source row may split its mass across several reference points. This leads to two natural conformal constructions.

Randomized conditional transport. For a positive test-row mass, normalize the last row of the plan, $a^y_{\tilde{x},j} = \frac{\widehat{\pi}^y_{\tilde{x},n+1,j}}{p_{n+1}(\tilde{x})}$, $j\in[n+1]$, and draw $J^y_{\tilde{x}} \mid \widehat{\pi}^y_{\tilde{x}} \sim \mathrm{Categorical}\{a^y_{\tilde{x},1},\dots,a^y_{\tilde{x},n+1}\}$. The randomized LOTCP set is

$$\widehat{\mathcal{C}}^{\mathrm{LOTCP-rand}}(X_{n+1},\widetilde{X}_{n+1}) = \left\{ y\in\mathcal{Y} : \|U_{J^y_{\widetilde{X}_{n+1}}}\| \leq \widehat{q}_{1-\alpha}\left(\{\|U_i\|\}_{i=1}^{n+1}\right)\right\}.$$

A single $V\sim\mathrm{Unif}(0,1)$ is used for inverse-cdf sampling across all trial responses, so the construction defines one coherent random set rather than unrelated pointwise decisions.

Barycentric transport. Alternatively, for every positive-mass row define $\overline{U}^y_{\tilde{x},i} = \frac{1}{p_i(\tilde{x})}\sum_{j=1}^{n+1}\widehat{\pi}^y_{\tilde{x},ij}U_j$, and $R^y_{\tilde{x},i} = \|\overline{U}^y_{\tilde{x},i}\|$. Let

$$\widehat{q}^y_{1-\alpha}(\tilde{x}) = \inf\left\{ t : \sum_{i=1}^{n+1} p_i(\tilde{x})\mathbb{1}\{R^y_{\tilde{x},i}\leq t\}\geq 1-\alpha\right\}.$$

The deterministic LOTCP set is

$$\widehat{\mathcal{C}}^{\mathrm{LOTCP-bar}}(X_{n+1},\widetilde{X}_{n+1}) = \left\{ y\in\mathcal{Y} : R^y_{\widetilde{X}_{n+1},n+1} \leq \widehat{q}^y_{1-\alpha}(\widetilde{X}_{n+1})\right\}.$$

The weighted quantile must be formed from the barycentric radii themselves. In general, the barycenters do not follow $\mathcal{Q}$, so the empirical reference quantile used by the randomized construction is not valid here.

The quadratic cost also gives the localized construction a finite polyhedral representation. Fix $\tilde{x}$ and write $p_i = p_i(\tilde{x})$. For the following result, let $M = M^\top \succ 0$ be fixed and replace the Euclidean cost in (21) by $c_M(s,u) = (s-u)^\top M(s-u)$. The ordinary LOTCP construction corresponds to $M = I$, while the same result applies to IMV-OTCP after its local metric has been fixed. Let $\Pi_{\tilde{x}}$ denote the transportation polytope defined by the marginal constraints in (21), and let $\mathfrak{B}_{\tilde{x}}$ be a finite index set for its basic feasible plans, written $\{\pi^b : b\in\mathfrak{B}_{\tilde{x}}\}$.

For every $b\in\mathfrak{B}_{\tilde{x}}$, define the unnormalized first moment of its test row and its query-free objective term by

$$m_b = \sum_{j=1}^{n+1}\pi^b_{n+1,j}U_j, \quad \kappa_{b,M} = \sum_{i=1}^{n}\sum_{j=1}^{n+1}\pi^b_{ij}c_M(S_i,U_j) + \sum_{j=1}^{n+1}\pi^b_{n+1,j}U_j^\top MU_j.$$

The optimality cell associated with b is

$$\mathcal{R}_{b,M}(\tilde{x}) = \bigcap_{b'\in\mathfrak{B}_{\tilde{x}}}\left\{ s\in\mathbb{R}^d : 2s^\top M(m_{b'}-m_b)\leq \kappa_{b',M}-\kappa_{b,M}\right\}.$$

For the randomized construction and the common draw $V\sim\mathrm{Unif}(0,1)$, set

$$J_b(V) = \min\left\{ j : \frac{1}{p_{n+1}}\sum_{\ell=1}^{j}\pi^b_{n+1,\ell}\geq V\right\}.$$

For the barycentric construction, set $\overline{U}_i^b = \frac{1}{p_i}\sum_{j=1}^{n+1}\pi_{ij}^b U_j$, $R_i^b = \|\overline{U}_i^b\|$ and

$$q_b = \inf\left\{t : \sum_{i=1}^{n+1} p_i \mathbb{1}\{R_i^b \le t\} \ge 1-\alpha\right\}.$$

Proposition C.1 (Polyhedral representation of localized OTCP)**.** *Suppose the optimizer of the localized transportation problem is selected from its basic feasible plans. Conditional on $\tilde{x}$, the reference cloud, and V for the randomized construction, each $\mathcal{R}_{b,M}(\tilde{x})$ is a convex polyhedron. Apart from optimal-basis tie boundaries, the two accepted score regions are respectively*

$$\widehat{\mathcal{A}}_M^{\mathrm{LOTCP-rand}}(\tilde{x}, V) = \bigcup_{\substack{b\in\mathfrak{B}_{\tilde{x}}:\\ \|U_{J_b(V)}\|\le \widehat{q}_{1-\alpha}(\{\|U_i\|\}_{i=1}^{n+1})}} \mathcal{R}_{b,M}(\tilde{x}),$$

$$\widehat{\mathcal{A}}_M^{\mathrm{LOTCP-bar}}(\tilde{x}) = \bigcup_{\substack{b\in\mathfrak{B}_{\tilde{x}}:\\ R_{n+1}^b \le q_b}} \mathcal{R}_{b,M}(\tilde{x}).$$

Under a deterministic tie-breaking rule, assigning each tie boundary to its selected adjacent cell gives the corresponding exact, possibly half-open, finite polyhedral union. The prediction sets are obtained by translating these score regions by $\widehat{\mu}(X_{n+1})$.

Problem (21) is an unregularized discrete Kantorovich linear program and can be solved exactly by specialized transportation or minimum-cost-flow algorithms; in the equal-mass assignment special case, auction or Hungarian algorithms apply (Gabriel & Marco, 2019). This does not, however, remove the dependence on the trial response. Changing y changes the last row of the cost matrix, and may change the optimal transport basis. Thus a direct pointwise implementation must reoptimize the transport problem for each trial y. A parametric or warm-started transportation solver can reuse the current basis until a cell boundary is crossed, but enumerating all cells in $\mathfrak{B}_{\tilde{x}}$ in advance can be combinatorial and need not be cheaper than repeated optimization.

Theorem C.1 (Finite-sample validity of weighted full-conformal LOTCP)**.** *Suppose the $\{X_i, Y_i\}_{i=1}^{n+1}$ are exchangeable, the score construction is permutation-equivariant at the true response, the reference cloud is independent, and the solution of* (21) *is selected permutation-equivariantly. Then*

$$\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{\mathrm{LOTCP-rand}}(X_{n+1}, \widetilde{X}_{n+1})\} \ge 1-\alpha,$$
$$\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{\mathrm{LOTCP-bar}}(X_{n+1}, \widetilde{X}_{n+1})\} \ge 1-\alpha.$$

The first probability also averages over the common auxiliary draw V.

These are finite-sample marginal guarantees, where the probability averages over both the data and the randomized center.

C.2 Task-aware metric extension

We next record the task-aware extension of the weighted construction. Let $\mathcal{D}_{\mathrm{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ be independent of the data used to learn the metric, and let $U_{1:n+1}$ be an independent final reference cloud. Draw $\widetilde{X}_{n+1}$ and form the weights $p_i(\tilde{x})$ as in (20). For a realized center $\tilde{x}$, use the nearest anchor metric

$$\widehat{M}(\tilde{x}) = \widehat{M}_{a_{j(\tilde{x})}}.$$

For each trial response y, the IMV-OTCP plan $\widehat{\pi}_{\tilde{x}}^{y,\widehat{M}}$ solves

$$\begin{aligned} \widehat{\pi}_{\tilde{x}}^{y,\widehat{M}} &\in \arg\min_{\pi\ge 0} \sum_{i=1}^{n+1}\sum_{j=1}^{n+1} \pi_{ij}(S_i^y - U_j)^\top \widehat{M}(\tilde{x})(S_i^y - U_j), \\ \text{subject to} \quad & \sum_{j=1}^{n+1}\pi_{ij} = p_i(\tilde{x}), \quad i\in[n+1], \qquad \sum_{i=1}^{n+1}\pi_{ij} = \frac{1}{n+1}, \quad j\in[n+1]. \end{aligned} \tag{22}$$

Thus IMV-OTCP differs from LOTCP only through the ground cost: the Euclidean cost in (21) is replaced by $c_{\widehat{M}(\tilde{x})}$, while the localization weights and target masses remain unchanged.

Accordingly, the two LOTCP acceptance rules carry over without introducing a second fitted transport map. For randomized conditional transport, draw $J_{\tilde{x}}^{y,\widehat{M}}$ from the normalized last row of $\widehat{\pi}_{\tilde{x}}^{y,\widehat{M}}$, with probabilities $\widehat{\pi}_{\tilde{x},n+1,j}^{y,\widehat{M}}/p_{n+1}(\tilde{x})$ for $j \in [n+1]$. The resulting set is

$$\widehat{\mathcal{C}}^{\mathrm{IMV-rand}}(X_{n+1}, \widetilde{X}_{n+1}) = \left\{ y \in \mathcal{Y} : \| U_{J_{\widetilde{X}_{n+1}}^{y,\widehat{M}}} \| \leq \widehat{q}_{1-\alpha}\left(\{\|U_i\|\}_{i=1}^{n+1} \right) \right\}.$$

For barycentric transport, define $\overline{U}_{\tilde{x},i}^{y,\widehat{M}} = \frac{1}{p_i(\tilde{x})} \sum_{j=1}^{n+1} \widehat{\pi}_{\tilde{x},ij}^{y,\widehat{M}} U_j$ and $R_{\tilde{x},i}^{y,\widehat{M}} = \|\overline{U}_{\tilde{x},i}^{y,\widehat{M}}\|$. Let $\widehat{q}_{1-\alpha}^{y,\widehat{M}}(\tilde{x})$ be the weighted $(1-\alpha)$-quantile of these radii under the masses $p_i(\tilde{x})$. The barycentric IMV-OTCP set is

$$\widehat{\mathcal{C}}^{\mathrm{IMV-bar}}(X_{n+1}, \widetilde{X}_{n+1}) = \left\{ y \in \mathcal{Y} : R_{\widetilde{X}_{n+1},n+1}^{y,\widehat{M}} \leq \widehat{q}_{1-\alpha}^{y,\widehat{M}}(\widetilde{X}_{n+1}) \right\}.$$

Theorem C.2 (Finite-sample validity of weighted full-conformal IMV-OTCP)**.** *Suppose the assumptions of Theorem C.1 hold. Assume in addition that the estimation sample is independent of the augmented calibration-test sample, and that $\widehat{M}(\tilde{x})$ is an estimation-sample-measurable function of $\tilde{x}$ applied identically to all source rows. Then*

$$\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{\mathrm{IMV-rand}}(X_{n+1}, \widetilde{X}_{n+1})\} \geq 1-\alpha,$$
$$\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{\mathrm{IMV-bar}}(X_{n+1}, \widetilde{X}_{n+1})\} \geq 1-\alpha.$$

D Additional Experimental Results

D.1 Aspect-ratio ablation

We isolate the effect of the learned metric M in the Five-star and Clover settings by varying only the anisotropy of the linear transform, keeping the angular radial functions and rotations fixed. More precisely, let $R(\gamma)$ denote rotation through angle γ and write each base transform as $A_k^{(0)} = R(\gamma_k)\,\mathrm{diag}(a_k, b_k)$, where $k \in \{1, 2\}$ indexes five-star, and clover settings, respectively, and define the aspect-ratio family

$$A_k(\rho) = R(\gamma_k)\,\mathrm{diag}\left(\sqrt{a_k b_k \rho}, \sqrt{a_k b_k / \rho}\right), \qquad \rho > 0.$$

Thus the ratio between the two diagonal axes is exactly ρ, while $\det\{A_k(\rho)\} = a_k b_k$ is held fixed. In this way, increasing ρ makes the residual distribution more elongated without mechanically changing the overall area scale. We display $\rho \in \{2, 4, 6, 8, 10\}$. The $\rho = 6$ records are exactly those used for Five-star and Clover in Figure 2. The bars give the percentage improvement of MV-OTCP over FOTCP, so positive values favor MV-OTCP.

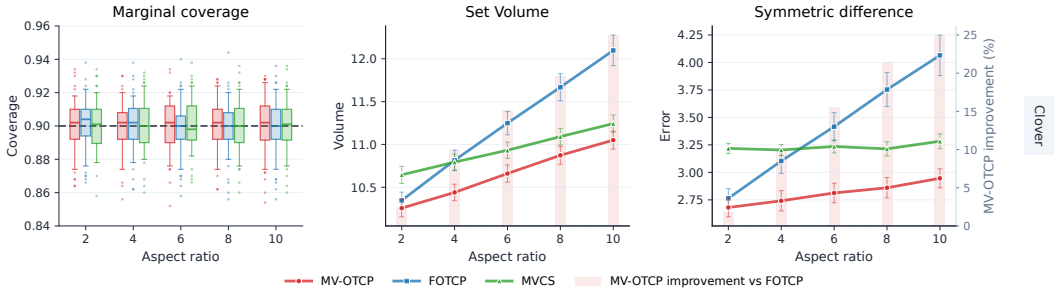


Figure 5: Aspect-ratio ablation for the clover setting. Translucent bars report the mean paired improvement of MV-OTCP over FOTCP on the right axis. The dashed line in the marginal-coverage panels marks the nominal target.

Figure 5 shows that marginal coverage remains near 0.9 throughout. MV-OTCP uniformly improves on MVCS in both efficiency metrics and is close to FOTCP at the smallest displayed aspect ratio. Its advantage over FOTCP increases with ρ, especially for symmetric difference, showing that the learned cost captures global anisotropy without sacrificing the nonconvex geometry supplied by the OT assignment.

D.2 Localized ground-cost and anchor ablation

We isolate the contribution of the learned local ground cost inside the IMV-OTCP pipeline on the heterogeneous Clover and Exp settings. We compare the reported method with IMV-OTCP($M = I$) over 100 paired replications and $L \in \{3, 5, 10, 15, 20\}$ empirical-quantile anchors. The two variants use the same localized fitting and calibration pipeline, differing only in the ground-cost metric: IMV-OTCP learns M at each anchor, whereas the baseline fixes $M = I$. This comparison isolates the contribution of metric learning in the individualized setting.

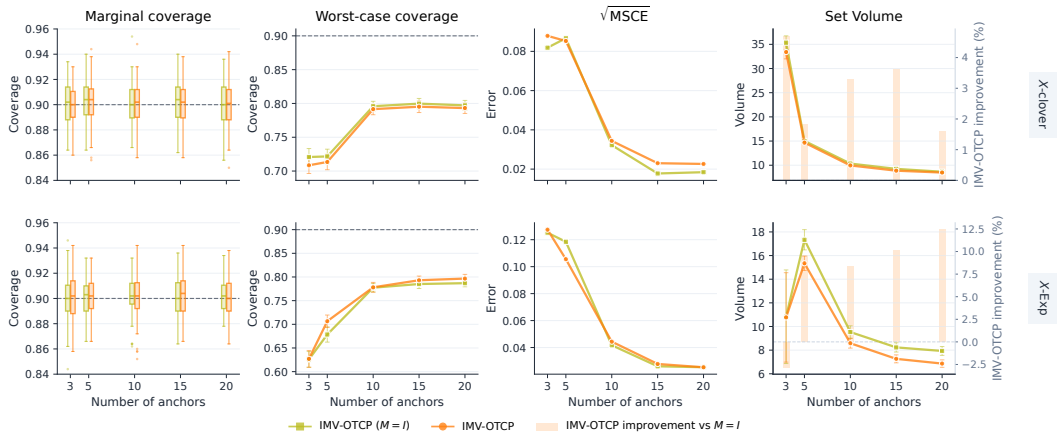


Figure 6: Localized ground-cost and anchor ablation. Translucent bars report the mean paired volume improvement of learned-metric IMV-OTCP over IMV-OTCP($M = I$).

Both branches retain marginal coverage near 0.9. Increasing the anchor count substantially improves the two local-coverage diagnostics, indicating that very coarse anchor grids are not adequate approximations of the varying conditional geometry. With too few anchors, the procedure may fail to capture the conditional structure adequately and can produce less stable prediction sets. Across anchor counts, IMV-OTCP consistently reduces volume relative to the $M = I$ baseline, while maintaining comparable marginal and local-coverage diagnostics. This provides evidence that learning the ground metric contributes additional efficiency beyond the localized construction itself.

D.3 General task objectives

We compare TA-OTCP with SCP and two decision-focused uncertainty quantification methods of Cortes-Gomez et al. (2025), denoted DFUQ-p and DFUQ-g. DFUQ-p incorporates the downstream set loss as a penalty in the conformal score and tunes the penalty weight on a separate estimation sample, whereas DFUQ-g constructs a loss-aware ordering through a greedy optimization procedure and subsequently conformalizes the resulting score. We consider the following two downstream objectives: a quadratic task loss for average RO in Example A.1 and a smoothed contextual robust-optimization objective in Example A.2. All methods use the same residual representation within a setting.

We consider the following two settings. For a prediction set $\mathcal{C}$, let $z_{\mathcal{C}}(x)$ minimize the corresponding set-induced task criterion over z.

(i) *Integrated quadratic task loss.* Let $X \sim \mathrm{Unif}[-1, 1]$, $Y = \mu(X) + \varepsilon$, and $\mu(x) = (x, -x)^\top$. Draw $C \in \{1, 2\}$ with probabilities $(2/3, 1/3)$ and $D = \sqrt{U}(\cos\Theta, \sin\Theta)^\top$, where $U \sim \mathrm{Unif}[0, 1]$ and $\Theta \sim \mathrm{Unif}[0, 2\pi]$ are independent, and set $\varepsilon = m_C + A_C D$, where

$$m_1 = (-0.22, 0)^\top, \quad m_2 = (0.44, 0)^\top, \quad A_1 = \begin{pmatrix} 0.85 & 0.20 \\ 0.20 & 0.18 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0.70 & 0.16 \\ 0.16 & 0.16 \end{pmatrix}.$$

For $Q = \mathrm{diag}(1, 100)$, the scaled task loss is $\ell_{\mathrm{I}}(z, y) = 10(z - y)^\top Q(z - y)$. The reported task objective is $\mathcal{J}_{\mathrm{I}}(\mathcal{C}) = \mathbb{E}_X[\int_{\mathcal{C}(X)} \ell_{\mathrm{I}}\{z_{\mathcal{C}}(X), y\}\, dy]$.

(ii) *Smoothed quadratic CRO.* Let $X \sim N\{(1,1)^\top, 2.25I_2\}$, $Y = \mu(X) + \varepsilon$, $\mu(x) = (-x_1 - x_2^2, -x_1^2 - x_2)^\top$, and $\varepsilon \sim N(0, I_2)$. We use $\ell_{\mathrm{II}}(z,y) = \frac{1}{2}(z-y)^\top H(z-y)$, where $z \in \mathbb{R}^2$ and $H = R_{45^\circ}\,\mathrm{diag}(20 \cdot 50/51, 20/51)R_{45^\circ}^\top$ with R_{45° the two-dimensional rotation matrix through 45°. The smoothed contextual robust objective is $\mathcal{J}_\tau(\mathcal{C}) = \mathbb{E}_X[\tau \log \int_{y \in \mathcal{C}(X)} \exp\{\ell_{\mathrm{II}}(z_{\mathcal{C}}(X), y)/\tau\}\, dy]$ with $\tau = 0.25$.

Evaluation metrics. Besides marginal coverage and set volume, we report the *Task objective* $\widehat{\mathcal{J}}_{\mathrm{I}}(\widehat{\mathcal{C}})$ or $\widehat{\mathcal{J}}_\tau(\widehat{\mathcal{C}})$, evaluated at the set-induced decision that minimizes the corresponding criterion. Lower values mean that the prediction set is better aligned with the downstream loss. The estimation of the task objective can be found in Appendix E.5.

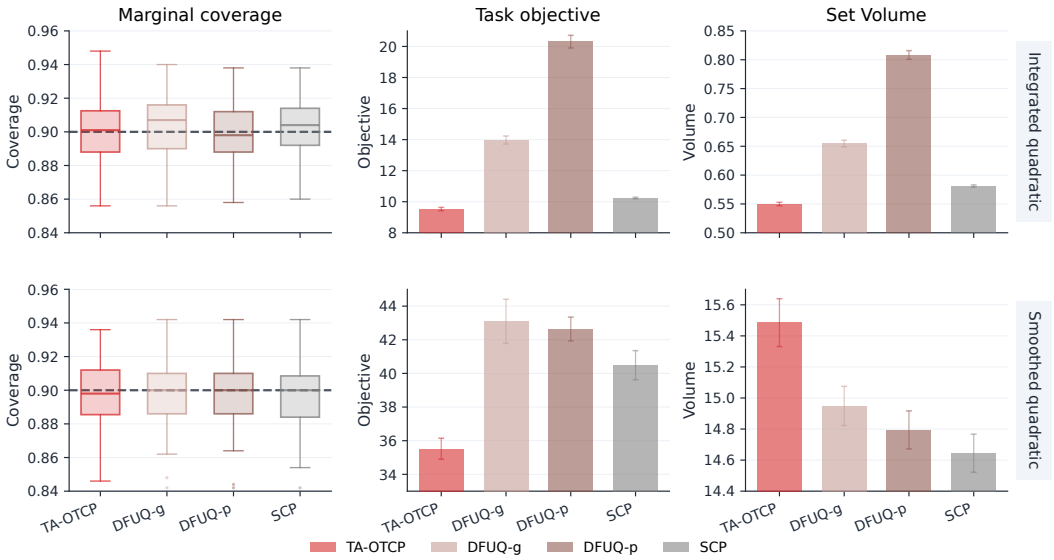


Figure 7: General task-objective experiments. The rows correspond to the integrated and smoothed quadratic settings. The dashed line in the marginal-coverage panels marks the nominal target; lower task objective and volume are preferable.

Figure 7 shows that all methods remain close to the nominal coverage level. Across both downstream objectives, TA-OTCP achieves the smallest task objective, outperforming SCP and the two DFUQ baselines. The volume results further highlight the distinction between task alignment and set size: under the integrated quadratic objective, TA-OTCP also yields the smallest volume, whereas under the smoothed quadratic objective it produces a larger set despite attaining the best task loss. This illustrates that the learned transport geometry is adapting to the downstream objective itself, rather than simply favoring smaller prediction sets.

D.4 U.S. STOCK PORTFOLIOS

Following the stock experiment of Chenreddy et al. (2022), we use daily U.S. equity observations from January 2012 through December 2017, where the response is the vector of percentage changes in adjusted closing prices and the covariates include individual trading volumes and broad market indices. In each replication, we randomly sample 15 stocks from the universe of 64 firms and apply all competing methods to the same selected subset. We use 2012–2014 for training, 2015 for task-aware metric estimation, 2016 for calibration, and 2017 for testing.

For return vector $y \in \mathbb{R}^{15}$ and long-only portfolio $z \in \mathcal{Z} := \{z \geq 0 : \mathbf{1}^\top z = 1\}$, let

$$\ell(z,y) = -y^\top z, \qquad z_C(x) \in \operatorname*{arg\,min}_{z \in \mathcal{Z}} \sup_{y \in C(x)} \ell(z,y), \qquad r_C(x) = \sup_{y \in C(x)} \ell\{z_C(x), y\}.$$

We compare against several existing methods at level $1 - \alpha = 0.9$. SCP uses an ellipsoidal score with covariance matrix $\widehat{\Sigma}$ estimated from the training data, while CRO is the ellipsoidal predict-then-

calibrate method of Sun et al. (2023). We additionally include the end-to-end method of Yeh et al. (2024), denoted E2E, which predicts a covariate-dependent Gaussian ellipsoid and differentiates the realized portfolio loss through both minibatch conformal calibration and the robust portfolio problem during training. Precise definitions are given in Appendix E.6.

Evaluation metrics. Besides marginal coverage, we report the *Certificate:* $n_{\text{test}}^{-1}\sum_{i=1}^{n_{\text{test}}} r_{\widehat{C}}(X_i)$ and the *Decision loss:* $n_{\text{test}}^{-1}\sum_{i=1}^{n_{\text{test}}}(-Y_i^\top z_{\widehat{C}}(X_i))$. The former averages the worst-case portfolio loss certified by the calibrated uncertainty set, whereas the latter evaluates the corresponding set-induced portfolio at the realized return. Lower values are better for both metrics.

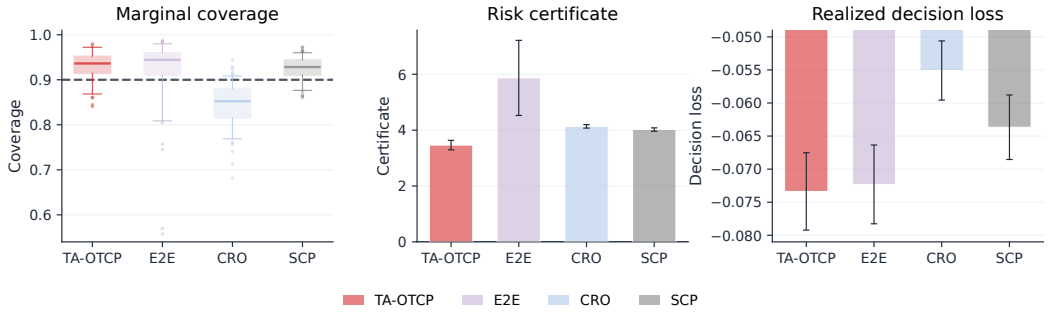


Figure 8: U.S. equity portfolio results for ellipsoidal uncertainty sets. The dashed line is the nominal level 0.9. Certificate and decision loss are means with normal-approximation 95% intervals (mean ± 1.96 standard errors). Lower is better for the certificate and decision loss.

Figure 8 reports the results. All methods exhibit some deviation from the nominal coverage level, which may be due to a certain degree of distribution shift in the data. Among the methods with coverage close to the nominal level, TA-OTCP achieves both the smallest mean certificate and the lowest mean decision loss, highlighting the benefit of tailoring the uncertainty-set geometry to the downstream portfolio objective.

E Additional Experimental Details

Simulations use independent training, calibration, and test samples of sizes $1{,}000/1{,}000/500$, with the training sample also serving as the estimation sample for metric learning. Real-data experiments use separate training, estimation, calibration, and test splits; the exact splitting strategy can be found in Appendix E.4.

E.1 Definition of Evaluation Metrics

We provide detailed definitions of the evaluation metrics used in our experiments below.

- Worst-case coverage: Let $\mathcal{B}_{\text{worst}}$ denote a collection of balls in $\mathcal{X}$ with $|\mathcal{B}_{\text{worst}}| = 100$. Specifically, each ball is centered at a randomly selected test point, and its radius is drawn independently from a uniform distribution between 0.04 and 0.12 times the width of the covariate support. We report

$$\textit{Worst-case} = \min_{B\in\mathcal{B}_{\text{worst}}} \frac{\sum_i \mathbb{1}(X_i \in B)C_i}{\sum_i \mathbb{1}(X_i \in B)}.$$

- $\sqrt{\text{MSCE}}$: For simulations, since $X \sim \text{Unif}(-2, 2)$, we partition the diagnostic support $[-2, 2]$ into 40 equal-width cells $\mathcal{I}_{\text{grid}}$. Because a single replication has only about 12–13 test observations per cell, we pool the cell counts across all R replications and define as:

$$\sqrt{\text{MSCE}}_{\text{sim}} = \left[\frac{1}{|\mathcal{I}_{\text{grid}}|}\sum_{J\in\mathcal{I}_{\text{grid}}}\left\{\frac{\sum_{r=1}^{R}\sum_i \mathbb{1}\{X_{ri}\in J\}\mathbb{1}\{Y_{ri}\notin \widehat{C}(X_{ri})\}}{\sum_{r=1}^{R}\sum_i \mathbb{1}\{X_{ri}\in J\}} - \alpha\right\}^2\right]^{1/2}.$$

For real data, because regular grids are impractical in high-dimensional covariate spaces, we assess $\sqrt{\mathrm{MSCE}}$ using a mass-controlled random-ball diagnostic. For replication r, draw a center index J_r randomly from the n_{test} test observations, and independently draw $P_r \sim \mathrm{Unif}(0.05, 0.20)$. Set $K_r = \min\{n_{\mathrm{test}}, \max(15, \lceil P_r n_{\mathrm{test}} \rceil)\}$ and let B_r contain the K_r test points nearest to X_{J_r}. Hence B_r is the empirical k-nearest-neighbor ball with random probability mass. Let $\mathcal{B}_{\mathrm{msce}}$ be the set of all B_r, with $|\mathcal{B}_{\mathrm{msce}}| = 100$. We report

$$\sqrt{\mathrm{MSCE}}_{\mathrm{real}} = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{1}{|\mathcal{B}_{\mathrm{msce}}|}\sum_{B_r \in \mathcal{B}_{\mathrm{msce}}}\left\{\frac{\sum_i \mathbb{1}\{X_{ri} \in B_r\}\mathbb{1}\{Y_{ri} \notin \widehat{C}(X_{ri})\}}{\sum_i \mathbb{1}\{X_{ri} \in B_r\}} - \alpha\right\}^2\right]^{1/2}.$$

E.2 Implementation details of homogeneous experiments

Let $X \sim \mathrm{Unif}(-2, 2)$ and generate

$$Y = m(X) + \varepsilon, \qquad m(x) = \begin{pmatrix} 1.2x + 0.35\sin(\pi x) \\ -0.8 + 0.45x^2 + 0.25\cos(\pi x) \end{pmatrix}.$$

Let $R(\gamma)$ denote rotation through angle γ and write $e(\theta) = (\cos\theta, \sin\theta)^\top$. For the first two settings, let $R_0 = \{-2\log V\}^{1/2}$ with $V \sim \mathrm{Unif}(0, 1)$, draw Θ with density proportional to $r_j(\theta)^2$, and set $\varepsilon = A_j^{(0)}\{R_0 r_j(\Theta)e(\Theta)\} + \xi$, where $\xi \sim N(0, 0.015^2 I_2)$. We consider the following three residual laws:

(i) *Five-star.*

$$r_1(\theta) = 0.70\{1 + 0.43\cos(5(\theta - 0.18))\},$$
$$A_1^{(0)} = R(0.34)\,\mathrm{diag}(1.55, 0.68).$$

(ii) *Clover.*

$$r_2(\theta) = 0.74\{1 + 0.40\cos(4(\theta + 0.24))\},$$
$$A_2^{(0)} = R(-0.46)\,\mathrm{diag}(1.42, 0.72).$$

(iii) *Exp Fixed.*

$$r_3(\theta) = 0.58\{1 + 0.42\cos[3(\theta - 0.27)] + 0.12\cos(\theta - 0.27)\},$$
$$A_3 = R(-0.43)\,\mathrm{diag}(2, 1/2).$$

Draw Θ_3 with density $r_3(\theta)^2 / \int_{-\pi}^{\pi} r_3(t)^2\,dt$ and independently draw $R_3 \sim \mathrm{Gamma}(2, 1)$. The resulting density with respect to area is proportional to $\exp(-R_3)$, so the population HPD region used as the oracle is available in exact form.

All methods use the same residual $s(x, y) = y - \widehat{\mu}(x)$, where $\widehat{\mu}$ is a fifth-degree polynomial ridge regression fit on the training sample. The SCP baseline calibrates the scalar Mahalanobis score $\|\widehat{\Sigma}^{-1/2}s(x, y)\|_2$ with an estimated global covariance $\widehat{\Sigma}$. The one-matrix MVCS baseline jointly learns a global matrix $\widehat{\Lambda}$ and scalar exponent $\widehat{p}$ by minimizing its log-volume objective, and then calibrates the score $\|\widehat{\Lambda}s(x, y)\|_{\widehat{p}}$. FOTCP uses the Euclidean OT cost, while MV-OTCP learns a positive definite metric cost $c_M(s, u) = (s - u)^\top M(s - u)$ before final calibration. For MV-OTCP, the reference-to-score map is represented by a GradNet (Chaudhari et al., 2025). Holding the learned metric fixed, final calibration uses the augmented assignment (7).

The SCP covariance ridge is $10^{-3}\{\mathrm{tr}(\widehat{\Sigma})/2\}I_2$. The MVCS matrix and scalar exponent are jointly trained for 200 epochs using minibatches of size 100, matrix learning rate 0.004, exponent learning rate 0.01, and cosine learning-rate decay, and are frozen before calibration. MV-OTCP uses a four-module GradNet of hidden width 24, 20 outer metric updates, 500 map steps in the first update and 200 thereafter, metric learning rate 0.002. The reference law $\mathcal{Q}$ is $U = R\Theta$, where $R \sim \mathrm{Unif}(0, 1)$ and Θ is uniform on the unit sphere. Empirically, we choose the reference cloud $U_{1:n+1}$ as in Ndiaye (2025), that is, let $n + 1 = n_R n_S + n_o$ and sample n_S points $\theta_1, \ldots, \theta_{n_S}$ on the unit sphere, then scale them to form n_R radial shells of the form $\{\frac{r}{n_R}\theta_s : r = 1, \ldots, n_R\}$ with n_o copies of the origin. In simulations, we use $n_R = 20$, $n_S = 50$ and $n_o = 1$, and is shared by FOTCP and MV-OTCP.

E.3 Implementation details of heterogeneous experiments

This appendix provides the coefficient functions used in the heterogeneous simulations of Subsection 5.2. Let

$$R(\gamma) = \begin{pmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{pmatrix}, \qquad m(x) = \big(1.2x + 0.25\sin(\pi x),\ -0.6 + 0.35x^2\big)^\top,$$

and define

$$s_{1/2}(x) = \frac{1+\sin(\pi x/2)}{2}, \qquad c_{1/2}(x) = \frac{1+\cos(\pi x/2)}{2}, \qquad s_1(x) = \frac{1+\sin(\pi x)}{2}.$$

X-varying mixture. Let

$$\{v_1, v_2, v_3, v_4\} = \big\{(1,0)^\top,\ (0,1)^\top,\ (-1,0)^\top,\ (0,-1)^\top\big\},$$

and define

$$\begin{aligned} p(x) &= \mathrm{softmax}\left[1.6\{\sin(\pi x/2), \cos(\pi x/2), -\sin(\pi x/2), -\cos(\pi x/2)\}\right], \\ \theta_{\mathrm{m}}(x) &= 0.65\pi x, \qquad a_{\mathrm{m}}(x) = 0.65 + 0.55 s_1(x), \qquad \tau(x) = 0.08 + 0.12 c_{1/2}(x). \end{aligned}$$

X-varying clover. The radial boundary and covariate-dependent deformation are given by

$$\begin{aligned} b(x) &= 0.48 + 0.30 s_1(x), \qquad \delta(x) = 0.35\pi x + 0.12\sin(\pi x), \\ q_x(\theta) &= 1 + b(x)\cos\{4(\theta - \delta(x))\}, \\ \rho_{\mathrm{c}}(x) &= 3 + \frac{7}{2}\{1+\cos(\pi x)\}, \qquad a_{\mathrm{c}}(x) = 0.65 + 0.30\{1+\sin(3\pi x/2 + 0.3)\}, \\ \gamma_{\mathrm{c}}(x) &= 0.85\pi x + 0.25\sin(\pi x), \\ A_{\mathrm{c}}(x) &= R\{\gamma_{\mathrm{c}}(x)\}\, a_{\mathrm{c}}(x)\, \mathrm{diag}\left\{\sqrt{\rho_{\mathrm{c}}(x)}, \rho_{\mathrm{c}}(x)^{-1/2}\right\}. \end{aligned}$$

Exp transformed. Let

$$\eta = (0.304717, -1.039984)^\top, \qquad \eta_r = 0.750451,$$

and define

$$\begin{aligned} m(x) &= 2\left[\sin(x\eta) + \frac{1}{2}\tanh(x^2\eta) + (x,0)^\top\right], \\ r(x) &= \frac{|x|}{2} + \frac{(\eta_r x)^2}{10} + 0.15, \end{aligned}$$

where the sine and hyperbolic tangent are applied componentwise. The rotation angle is obtained from the anchor locations

$$a = (-2.25, -1.55, -1.05, -0.60, -0.20, 0.20, 0.60, 1.05, 1.55, 2.25)$$

and fixed angles

$$\begin{aligned} \omega = (&1.721317, -0.384038, 2.253137, 1.240100, -2.549859, \\ &2.988423, 1.640789, 1.797395, -2.336631, -0.311734), \end{aligned}$$

drawn once from $\mathrm{Unif}[-\pi, \pi]$ using seed 42. Define

$$w_k(x) = \frac{(|x - a_k| + 10^{-6})^{-4}}{\sum_{\ell=1}^{10}(|x - a_\ell| + 10^{-6})^{-4}}, \qquad \theta_{\mathrm{e}}(x) = \sum_{k=1}^{10} w_k(x)\omega_k.$$

Finally, let $X \sim \mathrm{Unif}[-2, 2]$ and $E = (E_1, E_2)^\top$ with $E_1, E_2 \overset{\mathrm{iid}}{\sim} \mathrm{Exp}(1)$, independently, and generate

$$Y = g(X) + r(X) R\{\theta_{\mathrm{e}}(X)\} E.$$

All methods use the same residual $s(x, y) = y - \widehat{\mu}(x)$, where $\widehat{\mu}$ is a fifth-degree polynomial ridge regression fit on an independent training sample. RLCP uses $\|s(x, y)\|_2$ as its scalar score with bandwidth 0.2. For SCP, $\widehat{\Sigma}_x$ is represented by a Cholesky network with one hidden tanh layer of width 16 and then compute the score $\|\widehat{\Sigma}_x^{-1/2} s(x, y)\|_2$. MVCS learns an x-dependent matrix $\widehat{\Lambda}(x)$ and one scalar exponent p, and applies a global split-conformal quantile to $\|\widehat{\Lambda}(x)s(x, y)\|_p$ after freezing both fitted quantities. For IMV-OTCP, we use 15 empirical-quantile anchors and an RBF kernel with bandwidth of 0.1. The reference-to-score and score-to-reference maps are four-module GradNets of hidden width 32 (Chaudhari et al., 2025). The first outer iteration uses 600 map steps and subsequent iterations use 250; after selecting the best outer iterate, the final score-to-reference map uses 600 steps and the hybrid loss in (19). The reference law is the radial law $U = R\Theta$ above, represented by a 20×50 grid of equally spaced radial shells and common random directions, without an origin. Each calibration and test point is assigned to its nearest anchor before applying conformalied step.

E.4 Multivariate-regression volume study

Splitting and common representation. Each replication begins with one random permutation. `Energy` assigns 55% of the observations to training and 15% to each remaining split. The other datasets use 70% for training and 10% for each remaining split. Estimation and calibration samples are capped separately at $1{,}000$ observations, and the full test split is retained. Missing values are replaced by coordinatewise training medians. We then follow Dheur et al. (2025); Braun et al. (2026) to fit coordinatewise Gaussian quantile transforms $\widehat{Q}_X$ and $\widehat{Q}_Y$ on the training split, using at most $1{,}000$ empirical quantiles, and keep both transforms fixed.

Let $d_{\mathcal{X}}$ denote the original covariate dimension. To obtain a common covariate representation, let $\widehat{P}$ be the whitened PCA map fitted to $\widehat{Q}_X(X)$ on the training split, with

$$k = \min\{8, d_{\mathcal{X}}, n_{\mathrm{tr}} - 1\}, \qquad Z = \widehat{P}\{\widehat{Q}_X(X)\} \in \mathbb{R}^k.$$

The same frozen map is applied to estimation, calibration, and test covariates. Thus test observations do not affect either the PCA directions or their whitening scales. All methods receive Z as their covariate, and all localization and random-ball distances are Euclidean distances in this representation. The common center is ridge regression with penalty 1, fitted on $(Z, \widehat{Q}_Y(Y))$ from the training split. Define $S(z, y) = \widehat{Q}_Y(y) - \widehat{\mu}(z)$. Its residual $S = S(Z, Y)$ is the common multivariate score input. Set membership and volume are evaluated in these training-fitted response coordinates. Table 2 reports the effective sample sizes.

Table 2: Real multivariate-regression datasets and effective split sizes. Estimation and calibration are capped at $1{,}000$ observations; the test split is not truncated.

Dataset	n	$d_{\mathcal{X}}$	d	n_{tr}	n_{est}	n_{cal}	n_{test}
`Energy`	768	8	2	422	115	115	116
`Bias`	7,752	22	2	5,426	775	775	776
`Taxi`	61,286	6	2	42,900	1,000	1,000	6,130
`rf1`	9,125	64	8	6,387	912	912	914
`scm1d`	9,803	280	16	6,862	980	980	981

Method implementations. Whenever an ordinary split-conformal threshold is used below, it is the $\lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$th order statistic, with the usual $+\infty$ convention when the index exceeds n_{cal}. All method-specific hyperparameters were selected in pilot experiments whose random seeds are disjoint from those of the reported replications, and were fixed before formal evaluation.

SCP estimates a global covariance $\widehat{\Sigma}$ from the estimation residuals, adding $10^{-3}\,\mathrm{tr}(\widehat{\Sigma})/d$ to its eigenvalues with a numerical floor of 10^{-8}. It calibrates the Mahalanobis score $\|\widehat{\Sigma}^{-1/2} S\|_2$, giving the ellipsoid

$$\{y : S(z, y)^\top \widehat{\Sigma}^{-1} S(z, y) \le \widehat{q}^2\}.$$

MVCS uses the adaptive matrix score $\|\widehat{\Lambda}(z)S\|_{\widehat{p}}$. The matrix network has two ReLU hidden layers of width 64; the network and scalar exponent, initialized at $p = 2$, are jointly trained on the estimation residuals for 100 epochs with batch size 128, Adam learning rates 0.004 and 0.01, respectively, and cosine learning-rate decay. Computation uses double precision for $d \geq 8$ and single precision otherwise. The fitted matrix network and exponent are frozen before the common split-conformal calibration.

RLCP uses the Euclidean residual score $\|S\|_2$. For each test representation z, it draws $\widetilde{z} \sim N(z, h^2 I_k)$ and weights calibration point i by

$$w_i(\widetilde{z}) = \exp\{-\|Z_i - \widetilde{z}\|_2^2/(2h^2)\}.$$

The randomized threshold is the weighted $(1-\alpha)$ quantile after adding the test covariate with score $+\infty$. Writing $Z_{i,(k)}$ for the kth nearest estimation neighbor of Z_i, let $k_0 = \min\{n_{\text{est}} - 1, \max(20, \lceil n_{\text{est}}/4 \rceil)\}$. In the order `Energy`, `Bias`, `Taxi`, `rf1`, and `scm1d`, the RLCP bandwidth is

$$h = (1, 0.75, 0.75, 1, 0.75)\text{Median}_i d\{Z_i, Z_{i,(k_0)}\}.$$

For IMV-OTCP, the metric is parameterized by a trace-zero matrix exponential and learned with the implicit-gradient GradNetOT implementation. The common settings are four outer metric updates, two GradNet modules of hidden width 16, map-step counts $120/60$ for the first/later outer iterations, metric learning rate 0.02. The score KDE uses the weighted Scott bandwidth specified in Appendix A.2. The reference law is $U = R\Theta$ with $R \sim \text{Unif}(0,1)$ and Θ uniform on the unit sphere; an n_{est}-point stratified cloud uses radii i/n_{est} and independent random directions. Define

$$k_h(A) = \min\{n_{\text{est}} - 1, \max(20, \lceil n_{\text{est}}/A \rceil)\}, \qquad h_0(A) = \text{Median}_i d\{Z_i, Z_{i,(k_h(A))}\}.$$

The method fits A anchor-specific metric and transport-map pairs. The anchors are the centers returned by K-means on the estimation representations, using ten initializations, and the RBF localization bandwidth is $h = mh_0(A)$. Calibration and test observations are assigned to their nearest anchor; their transported radius is evaluated with that anchor's map, and one global split-conformal threshold is then calibrated from these assigned scores. In the dataset order above, the settings are

$$A = (6, 6, 8, 12, 16),$$
$$m = (0.35, 0.5, 1, 0.75, 0.75).$$

These dataset-specific bandwidth and anchor choices were fixed in the pilot experiments with disjoint random seeds described above; Table 1 uses this same fitted configuration.

E.5 Implementation details of general task experiments

For the integrated quadratic setting, $X \sim \text{Unif}[-1, 1]$, $\mu(x) = (x, -x)^\top$, and $Y = \mu(X) + m_C + A_C D$, where C has probabilities $(2/3, 1/3)$, D is uniform on the unit disk, and

$$m_1 = (-0.22, 0)^\top, \quad m_2 = (0.44, 0)^\top, \quad A_1 = \begin{pmatrix} 0.85 & 0.20 \\ 0.20 & 0.18 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0.70 & 0.16 \\ 0.16 & 0.16 \end{pmatrix}.$$

The loss is $10(z-y)^\top \text{diag}(1, 100)(z-y)$. Numerical integration uses a 61×61 residual grid on $[-2.5, 2.8] \times [-2, 2]$. TA-OTCP uses 40 outer updates, map-step counts $150/75$, metric learning rate 0.006, implicit damping 0.01, and 30 conjugate-gradient steps; the estimation and calibration reference factorizations are $(25, 40, 0)$ and $(25, 40, 1)$.

For the smoothed quadratic setting, $X \sim N\{(1, 1)^\top, 2.25 I_2\}$, $\mu(x) = (-x_1 - x_2^2, -x_1^2 - x_2)^\top$, and $Y = \mu(X) + \varepsilon$ with $\varepsilon \sim N(0, I_2)$. The quadratic curvature matrix is

$$H_{\text{Q}} = R_{45^\circ} \text{diag}(20 \cdot 50/51, 20/51) R_{45^\circ}^\top.$$

The smoothed robust objective uses temperature 0.25 and a 25×25 grid on $[-4, 4]^2$. TA-OTCP uses six outer updates, map-step counts $300/100$, metric learning rate 0.01, implicit damping 0.02, 20 conjugate-gradient steps, and reference factorizations $(250, 4, 0)$ and $(250, 4, 1)$. In both settings, the two finite-grid baselines receive a fresh split-conformal calibration after task-based selection.

E.6 Implementation details of U.S. stock portfolios

The chronological split is fixed as follows.

Table 3: Leakage-free chronological split for the portfolio experiment.

Training	Estimation	Calibration	Test
2012–2014	2015	2016	2017

The common ridge-regression center $\widehat{\mu}$ and residual covariance $\widehat{\Sigma}_{\mathrm{tr}}$ are estimated from training dates only. Using the symmetric positive-definite square root, the SCP score is $V_i^{\mathrm{SCP}} = \|\widehat{\Sigma}_{\mathrm{tr}}^{-1/2} S_i\|_2$. Its ordinary split-conformal quantile gives the corresponding ellipsoid. The task-aware metric and parametric score-to-reference map are fitted on 2015 using the empirical version of smoothed-CRO in Example A.2. Holding both fixed, TA-OTCP calibrates on 2016 the score $V_i^{\mathrm{TA}} = \|\widehat{T}_{\widehat{M}_{\mathrm{task}}}(S_i)\|_2$. For E2E, we adapt the ellipsoidal construction of Yeh et al. (2024) to the same four-way chronological split. A neural network predicts a conditional location and Cholesky factor,

$$h_\theta(x) = \{\mu_\theta(x), L_\theta(x)\}, \qquad \Sigma_\theta(x) = L_\theta(x) L_\theta(x)^\top,$$

where the diagonal of $L_\theta(x)$ is passed through softplus. All response standardization is estimated on 2012–2014 and the displayed parameters are mapped back to percentage-return units. The scalar nonconformity score and final uncertainty set are

$$V_i^{\mathrm{E2E}} = \{Y_i - \mu_\theta(X_i)\}^\top \Sigma_\theta(X_i)^{-1} \{Y_i - \mu_\theta(X_i)\},$$
$$\widehat{C}_{\mathrm{E2E}}(x) = \left\{ y : \{y - \mu_\theta(x)\}^\top \Sigma_\theta(x)^{-1} \{y - \mu_\theta(x)\} \le \widehat{q}_{\mathrm{E2E}} \right\},$$

where $\widehat{q}_{\mathrm{E2E}}$ is the ordinary plus-one split-conformal quantile computed on 2016.

We implement Algorithm 2 of Sun et al. (2023) using 2015 as their preliminary-calibration sample and 2016 as their scalar-adjustment sample. Write $S_i = Y_i - \widehat{\mu}(X_i)$. For CRO, $\widehat{g}(x)$ estimates the $(1-\alpha)$-quantile of $\|S_i\|_2$ using the 2015 data, and $\widehat{\Sigma}_{\mathrm{CRO}}$ is the regularized covariance matrix of the corresponding scaled residuals $S_i / \widehat{g}(X_i)$. The scalar calibration score is

$$V_i^{\mathrm{CRO}} = \frac{\{S_i^\top \widehat{\Sigma}_{\mathrm{CRO}}^{-1} S_i\}^{1/2}}{\widehat{g}(X_i)}.$$

Let $\widehat{\eta}$ be the $\lceil (n_{\mathrm{cal}}+1)(1-\alpha) \rceil$th order statistic of the 2016 scores. The resulting set is

$$\widehat{C}_{\mathrm{CRO}}(x) = \left\{ y : \ \{(y - \widehat{\mu}(x))^\top \widehat{\Sigma}_{\mathrm{CRO}}^{-1} (y - \widehat{\mu}(x))\}^{1/2} \le \widehat{\eta}\widehat{g}(x) \right\}.$$

F Additional proofs

F.1 Proof of Lemma 3.1

Proof. Let $A = M^{1/2}$, $\widetilde{\mathcal{P}} = A_\# \mathcal{P}$, and $\widetilde{\mathcal{Q}} = A_\# \mathcal{Q}$. For any measurable map T with $T_\# \mathcal{P} = \mathcal{Q}$, define the transformed map $\widetilde{T}(\tilde{s}) = AT(A^{-1}\tilde{s})$. Then $\widetilde{T}_\# \widetilde{\mathcal{P}} = \widetilde{\mathcal{Q}}$. Moreover, by the change of variables $\tilde{s} = As$,

$$\begin{aligned} \int (s - T(s))^\top M (s - T(s)) \, d\mathcal{P}(s) &= \int \|As - AT(s)\|^2 \, d\mathcal{P}(s) \\ &= \int \|\tilde{s} - \widetilde{T}(\tilde{s})\|^2 \, d\widetilde{\mathcal{P}}(\tilde{s}). \end{aligned}$$

Therefore T minimizes the metric-cost Monge problem from $\mathcal{P}$ to $\mathcal{Q}$ if and only if $\widetilde{T}$ minimizes the ordinary squared-Euclidean Monge problem from $\widetilde{\mathcal{P}}$ to $\widetilde{\mathcal{Q}}$.

Since $\mathcal{P}$ is absolutely continuous and A is invertible, $\widetilde{\mathcal{P}}$ is also absolutely continuous. Brenier's theorem implies that there exists a unique Monge solution

$$\widetilde{T} = \operatorname*{arg\,min}_{\widetilde{T}_\# \widetilde{\mathcal{P}} = \widetilde{\mathcal{Q}}} \int \|\tilde{s} - \widetilde{T}(\tilde{s})\|^2 \, d\widetilde{\mathcal{P}}(\tilde{s}).$$

Moreover, there exists a convex function ψ such that

$$\widetilde{T}(\tilde{s}) = \nabla\psi(\tilde{s}) \qquad \widetilde{\mathcal{P}}\text{-a.e. } \tilde{s}.$$

Substituting $\tilde{s} = As$ gives $AT(s) = \nabla\psi(As)$. Conversely, if a measurable map T satisfies $T_{\#}\mathcal{P} = \mathcal{Q}$ and $AT(s) = \nabla\psi(As)$ for some convex ψ, then $\widetilde{T}(\tilde{s}) = AT(A^{-1}\tilde{s}) = \nabla\psi(\tilde{s})$ pushes $\widetilde{\mathcal{P}}$ to $\widetilde{\mathcal{Q}}$. By Brenier's theorem, $\widetilde{T}$ is the ordinary squared-Euclidean Monge solution from $\widetilde{\mathcal{P}}$ to $\widetilde{\mathcal{Q}}$. The reduction above then implies that T is the metric-cost Monge solution from $\mathcal{P}$ to $\mathcal{Q}$.

It remains to prove the inverse statement. Suppose now that $\mathcal{Q}$ is also absolutely continuous with respect to Lebesgue measure. Since A is invertible, $\widetilde{\mathcal{Q}}$ is absolutely continuous as well. Let G be the solution of the reverse metric-cost Monge problem from $\mathcal{Q}$ to $\mathcal{P}$, and define

$$\widetilde{G}(\tilde{u}) = AG(A^{-1}\tilde{u}).$$

By the same reduction as above, $\widetilde{G}$ is the unique Brenier map from $\widetilde{\mathcal{Q}}$ to $\widetilde{\mathcal{P}}$ for the ordinary squared-Euclidean cost. The forward Brenier map $\widetilde{T}$ induces the optimal coupling

$$\widetilde{\pi} = (\mathrm{Id}, \widetilde{T})_{\#}\widetilde{\mathcal{P}}$$

between $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{Q}}$. Since the squared-Euclidean cost is symmetric, the reversed coupling

$$\widetilde{\pi}^{\mathrm{rev}} = (\widetilde{T}, \mathrm{Id})_{\#}\widetilde{\mathcal{P}}$$

is an optimal coupling between $\widetilde{\mathcal{Q}}$ and $\widetilde{\mathcal{P}}$. Hence this reversed optimal coupling must be induced by $\widetilde{G}$, and therefore

$$\widetilde{G} \circ \widetilde{T}(\tilde{s}) = \tilde{s} \qquad \widetilde{\mathcal{P}}\text{-a.e. } \tilde{s}, \qquad \widetilde{T} \circ \widetilde{G}(\tilde{u}) = \tilde{u} \qquad \widetilde{\mathcal{Q}}\text{-a.e. } \tilde{u}.$$

Returning to the original variables $\tilde{s} = As$ and $\tilde{u} = Au$ gives

$$G \circ T(s) = s \qquad \mathcal{P}\text{-a.e. } s, \qquad T \circ G(u) = u \qquad \mathcal{Q}\text{-a.e. } u.$$

This proves that T and G are a.e. inverses of each other. □

F.2 Proof of Theorem 3.1

Proof. Let $U_{1:n+1} = \{U_1, \ldots, U_{n+1}\}$. Condition on the pre-calibration fit $(\widehat{\mu}, \widehat{M})$ and on $U_{1:n+1}$. Set $y = Y_{n+1}$ and write $Z_i = (X_i, Y_i)$. The augmented observations $Z_1, \ldots, Z_{n+1}$ are exchangeable. Let $\mathcal{D}_{\mathrm{aug}} = \{Z_1, \ldots, Z_{n+1}\}$ denote the unordered augmented sample, and fix an ordered representative $z_{1:n+1}$ of this multiset.

For any permutation τ of $[n+1]$, let $z^{\tau}_{1:n+1} = (z_{\tau(1)}, \ldots, z_{\tau(n+1)})$. Because the objective in (7) applies the same fixed metric $\widehat{M}$ to every augmented score and because ties are resolved permutation-equivariantly, the selected assignment satisfies

$$\widehat{\sigma}_{\widehat{M}}(z^{\tau}_{1:n+1}; U_{1:n+1}) = \widehat{\sigma}_{\widehat{M}}(z_{1:n+1}; U_{1:n+1}) \circ \tau.$$

Thus relabeling the augmented sample relabels the source index in exactly the same way, while preserving the reference point assigned to each unordered observation.

Conditional on $\mathcal{D}_{\mathrm{aug}}$ and $U_{1:n+1}$, exchangeability implies that the distinguished test observation is uniformly distributed over the $n+1$ elements of $\mathcal{D}_{\mathrm{aug}}$. Since the assignment is a permutation, the reference point assigned to a uniformly selected source index is uniformly distributed over $U_1, \ldots, U_{n+1}$. Therefore

$$U_{\widehat{\sigma}_{\widehat{M}}^{Y_{n+1}}(n+1)} \overset{d}{=} U_J \mid \mathcal{D}_{\mathrm{aug}}, U_{1:n+1}, \qquad J \sim \mathrm{Unif}\{1, \ldots, n+1\}.$$

By construction of $\widehat{q}_{1-\alpha}\left(\{\|U_i\|\}_{i=1}^{n+1}\right)$, at least $\lceil (n+1)(1-\alpha) \rceil$ of these values are no larger than the threshold. Hence

$$\mathbb{P}\left\{ \left\| U_{\widehat{\sigma}_{\widehat{M}}^{Y_{n+1}}(n+1)} \right\|_2 \le \widehat{q}_{1-\alpha}\left(\{\|U_i\|\}_{i=1}^{n+1}\right) \mid \mathcal{D}_{\mathrm{aug}}, U_{1:n+1} \right\} \ge 1 - \alpha.$$

The event inside the probability is exactly $Y_{n+1} \in \widehat{\mathcal{C}}^{\mathrm{MV}}(X_{n+1})$. Averaging over the conditioning variables gives the lower bound. If the reference radii are distinct, exactly $\lceil (n+1)(1-\alpha) \rceil$ destinations are accepted, giving the stated equality and upper bound. □

F.3 Proof of Theorem 3.2

We first introduce several auxiliary lemmas.

Lemma F.1 (Covering number of regularized positive definite metrics)**.** *Let*

$$\mathcal{M}_{\mathrm{sp}}(\underline{\lambda}, \overline{\lambda}) = \left\{ M = M^\top \succ 0 : \lambda_{\min}(M) \geq \underline{\lambda}, \lambda_{\max}(M) \leq \overline{\lambda} \right\},$$

where $0 < \underline{\lambda} \leq \overline{\lambda} < \infty$*. Then, with* $m = d(d+1)/2$ *and* $R_{\mathcal{M}} = \sqrt{d}\,\overline{\lambda}$*,*

$$N(\epsilon, \mathcal{M}_{\mathrm{sp}}, \|\cdot\|_F) \leq \left(1 + \frac{2R_{\mathcal{M}}}{\epsilon}\right)^m, \qquad \epsilon > 0. \tag{23}$$

Proof. The vector space of real symmetric $d \times d$ matrices has dimension $m = d(d+1)/2$ under the Frobenius inner product. If $M \in \mathcal{M}_{\mathrm{sp}}(\underline{\lambda}, \overline{\lambda})$, then

$$\|M\|_F = \left(\sum_{j=1}^{d} \lambda_j^2(M)\right)^{1/2} \leq \sqrt{d}\,\overline{\lambda} = R_{\mathcal{M}}.$$

Thus $\mathcal{M}_{\mathrm{sp}}$ is contained in an m-dimensional Euclidean ball of radius $R_{\mathcal{M}}$. The standard volumetric covering bound for a Euclidean ball gives (23). □

Assumption F.1 (Primitive global objective regularity)**.** *(i) On* $\bigcup_{M \in \mathcal{M}_0} G_M(\mathbb{B}_{1-\alpha})$*, the score density is* L_{f_P}*-Lipschitz and satisfies* $f_P \geq f_{P,\min} > 0$*.*

(ii) The maps G_M *are differentiable almost everywhere on* $\mathbb{B}_{1-\alpha}$ *and satisfy the Monge–Ampère identity. For every* $\epsilon > 0$*,*

$$\sup_{M_c \in \mathcal{M}_0} \mathbb{E}_U \left[\sup_{M \in \mathcal{M}_0(M_c, \epsilon)} \|G_M(U) - G_{M_c}(U)\| \mathbb{1}\{U \in \mathbb{B}_{1-\alpha}\} \right] \leq L_G \epsilon.$$

Lemma F.2 (Verification of global objective regularity)**.** *Under Assumption F.1, define*

$$B_\zeta = \frac{1}{f_{P,\min}}, \qquad L_\zeta = \frac{L_{f_P} L_G}{f_{P,\min}^2}.$$

Then $\sup_{M \in \mathcal{M}_0} |\zeta_M(U)| \leq B_\zeta$ *almost surely and, for every* $\epsilon > 0$*,*

$$\sup_{M_c \in \mathcal{M}_0} \mathbb{E} \left[\sup_{\substack{M \in \mathcal{M}_0: \\ \|M - M_c\|_F \leq \epsilon}} |\zeta_M(U) - \zeta_{M_c}(U)| \right] \leq L_\zeta \epsilon.$$

Consequently, Assumption 3.1(i) holds with these constants.

Proof. The metric Brenier representation gives

$$\nabla_u G_M(u) = M^{-1/2} \nabla^2 g_M(M^{1/2}u) M^{1/2}, \qquad |\det \nabla_u G_M(u)| = \det\{\nabla^2 g_M(M^{1/2}u)\}.$$

Since $(G_M)_{\#}\mathcal{Q} = \mathcal{P}$, the Monge–Ampère identity gives, for $\mathcal{Q}$-a.e. u,

$$f_Q(u) = f_P(G_M(u)) |\det \nabla_u G_M(u)| = f_P(G_M(u)) \det\{\nabla^2 g_M(M^{1/2}u)\}.$$

Therefore the Jacobian factor in (4) can be rewritten as

$$\zeta_M(u) = \mathbb{1}\{u \in \mathbb{B}_{1-\alpha}\} \frac{1}{f_P(G_M(u))}. \tag{24}$$

Then the envelope bound follows immediately from (24) and Assumption F.1(i):

$$|\zeta_M(U)| \leq 1/f_{P,\min} = B_\zeta.$$

For the modulus, fix $M_c \in \mathcal{M}_0$ and $M \in \mathcal{M}_0$ with $\|M - M_c\|_F \leq \epsilon$. On $\{U \in \mathbb{B}_{1-\alpha}\}$, write $S = G_M(U)$ and $S_c = G_{M_c}(U)$. Using (24),

$$|\zeta_M(U) - \zeta_{M_c}(U)| \leq \mathbb{1}\{U \in \mathbb{B}_{1-\alpha}\} \frac{L_{f_P}}{f_{P,\min}^2} \|S - S_c\|.$$

Taking the supremum over all such M, then expectation over U, and applying Assumption F.1(ii) proves the result. □

Proof of Theorem 3.2. **Step 1: Reduction to a finite metric net.** It is enough to control the infeasible empirical process

$$\sup_{M\in\mathcal{M}_0} |\mathcal{J}_N(M) - \mathcal{J}(M)|,$$

because the definition of $\mathcal{E}_N^{\mathrm{obj}} = \sup_{M\in\mathcal{M}_0} |\widehat{\mathcal{J}}(M) - \mathcal{J}_N(M)|$ and the triangle inequality give

$$\sup_{M\in\mathcal{M}_0} |\widehat{\mathcal{J}}(M) - \mathcal{J}(M)| \le \mathcal{E}_N^{\mathrm{obj}} + \sup_{M\in\mathcal{M}_0} |\mathcal{J}_N(M) - \mathcal{J}(M)|. \tag{25}$$

Fix $\epsilon > 0$ and let $\mathcal{N}_\epsilon = \{M_1, \dots, M_K\}$ be an ϵ-net of $\mathcal{M}_0$ under $\|\cdot\|_F$. By Lemma F.1 and the spectral bounds defining $\mathcal{M}_0$,

$$K \le \left(1 + \frac{2R_{\mathcal{M}}}{\epsilon}\right)^m. \tag{26}$$

For each center M_j, define the local oscillation envelope

$$H_j(U) = \sup_{\substack{M\in\mathcal{M}_0:\\ \|M-M_j\|_F\le\epsilon}} |\zeta_M(U) - \zeta_{M_j}(U)|.$$

For any $M \in \mathcal{M}_0$, there exists M_j such that $\|M - M_j\|_F \le \epsilon$ by the definition of $\mathcal{N}_\epsilon$. Then

$$\begin{aligned}
|\mathcal{J}_N(M) - \mathcal{J}(M)| &\le |\mathcal{J}_N(M_j) - \mathcal{J}(M_j)| + |\mathcal{J}_N(M) - \mathcal{J}_N(M_j)| + |\mathcal{J}(M) - \mathcal{J}(M_j)| \\
&= |\mathcal{J}_N(M_j) - \mathcal{J}(M_j)| + \left|\frac{1}{N}\sum_{i=1}^N \{\zeta_M(U_i) - \zeta_{M_j}(U_i)\}\right| + \left|\mathbb{E}\{\zeta_M(U) - \zeta_{M_j}(U)\}\right| \\
&\le |\mathcal{J}_N(M_j) - \mathcal{J}(M_j)| + \frac{1}{N}\sum_{i=1}^N |\zeta_M(U_i) - \zeta_{M_j}(U_i)| + \mathbb{E}|\zeta_M(U) - \zeta_{M_j}(U)| \\
&\le |\mathcal{J}_N(M_j) - \mathcal{J}(M_j)| + \frac{1}{N}\sum_{i=1}^N H_j(U_i) + \mathbb{E}H_j.
\end{aligned}$$

Assumption 3.1(i) gives $0 \le H_j \le 2B_\zeta$ a.s. and $\mathbb{E}H_j \le L_\zeta\epsilon$. Consequently,

$$\sup_{M\in\mathcal{M}_0} |\mathcal{J}_N(M) - \mathcal{J}(M)| \le \max_{1\le j\le K} |\mathcal{J}_N(M_j) - \mathcal{J}(M_j)| + \max_{1\le j\le K} \frac{1}{N}\sum_{i=1}^N H_j(U_i) + L_\zeta\epsilon. \tag{27}$$

Step 2: Concentration on the net. For a fixed j, the variables $\zeta_{M_j}(U_i)$ are independent and bounded by B_ζ. Hoeffding's inequality implies

$$\mathbb{P}\left\{|\mathcal{J}_N(M_j) - \mathcal{J}(M_j)| > t\right\} \le 2\exp\left(-\frac{Nt^2}{2B_\zeta^2}\right).$$

Taking a union bound over $j \in [K]$ gives, with probability at least $1 - \delta/2$,

$$\max_{1\le j\le K} |\mathcal{J}_N(M_j) - \mathcal{J}(M_j)| \le B_\zeta\sqrt{\frac{2\{\log(4K) + \log(1/\delta)\}}{N}}.$$

Similarly, for each j, Hoeffding's inequality applied to $H_j \in [0, 2B_\zeta]$ gives

$$\mathbb{P}\left\{\frac{1}{N}\sum_{i=1}^N H_j(U_i) > \mathbb{E}H_j + t\right\} \le \exp\left(-\frac{Nt^2}{2B_\zeta^2}\right).$$

Another union bound gives, with probability at least $1 - \delta/2$,

$$\max_{1\le j\le K} \frac{1}{N}\sum_{i=1}^N H_j(U_i) \le L_\zeta\epsilon + B_\zeta\sqrt{\frac{2\{\log(2K) + \log(1/\delta)\}}{N}}.$$

Step 3: Combine above results. Combining the last two displays with (27), and using (26), yields

$$\sup_{M\in\mathcal{M}_0} |\mathcal{J}_N(M) - \mathcal{J}(M)| \le CB_\zeta \sqrt{\frac{m\log(1+2R_{\mathcal{M}}/\epsilon) + \log(2/\delta)}{N}} + 2L_\zeta\epsilon.$$

If $L_\zeta > 0$, choose $\epsilon = B_\zeta/(L_\zeta N)$; then $2L_\zeta\epsilon = 2B_\zeta/N$ is absorbed into the displayed constant after replacing the logarithm by $\log\{e + 2R_{\mathcal{M}}L_\zeta N/B_\zeta\}$. By Eq. (25), this gives the uniform-deviation bound in the theorem.

Step 4: Argmin consistency. Let $\Delta_N = \sup_{M\in\mathcal{M}_0} |\widehat{\mathcal{J}}(M) - \mathcal{J}(M)|$. Taking the tail probability $\eta_N^{\text{tail}} = N^{-1}$ shows that the empirical-process term is $o_{\mathbb{P}}(1)$ for fixed d. Assumption 3.1(ii) therefore implies $\Delta_N = o_{\mathbb{P}}(1)$. Since $\widehat{M}$ minimizes $\widehat{\mathcal{J}}$ over $\mathcal{M}_0$ and $M^\star \in \mathcal{M}_0$,

$$\begin{aligned}\mathcal{J}(\widehat{M}) - \mathcal{J}(M^\star) &\le \{\mathcal{J}(\widehat{M}) - \widehat{\mathcal{J}}(\widehat{M})\} + \{\widehat{\mathcal{J}}(M^\star) - \mathcal{J}(M^\star)\} \\ &\le 2\Delta_N.\end{aligned}$$

For any $\epsilon > 0$, define

$$\text{gap}(\epsilon) := \inf_{M\in\mathcal{M}_0:\ \|M-M^\star\|_F\ge\epsilon} \{\mathcal{J}(M) - \mathcal{J}(M^\star)\}.$$

Assumption 3.1(iii) gives $\text{gap}(\epsilon) > 0$, and hence

$$\mathbb{P}\{\|\widehat{M} - M^\star\|_F \ge \epsilon\} \le \mathbb{P}\{2\Delta_N \ge \text{gap}(\epsilon)\} \to 0.$$

Thus $\|\widehat{M} - M^\star\|_F \xrightarrow{\mathbb{P}} 0$. Since $\widehat{M}, M^\star \in \mathcal{M}_0$, $\|\widehat{M} - M^\star\|_F \le 2\sqrt{d}\,\overline{\lambda}$ almost surely; boundedness together with convergence in probability gives $\mathbb{E}\|\widehat{M} - M^\star\|_F \to 0$. □

F.4 Proof of Theorem 3.3

Proof. **Step 1: Control set disagreement.** For the true augmented sample, write $S_i = Y_i - \widehat{\mu}(X_i)$ for $i \in [n+1]$. For a fixed metric M, let $\widehat{T}_{n+1,M}(S_i) = U_{\widehat{\sigma}_M^{Y_{n+1}}(i)}$, $i \in [n+1]$ be the empirical Monge map solution of (7) with $y = Y_{n+1}$ and $\widehat{M}$ replaced by M. Define

$$\widehat{p}_{\triangle,n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\left\{\mathbb{1}\{\|T_{M^\star}(S_i)\| \le q_\star\} \ne \mathbb{1}\{\|\widehat{T}_{n+1,\widehat{M}}(S_i)\| \le \widehat{q}_{1-\alpha}\left(\{\|U_i\|\}_{i=1}^{n+1}\right)\}\right\}.$$

For each i, set

$$e_i = \left|\|\widehat{T}_{n+1,\widehat{M}}(S_i)\| - \|T_{M^\star}(S_i)\|\right|.$$

Write

$$\widehat{q}_n = \widehat{q}_{1-\alpha}\left(\{\|U_i\|\}_{i=1}^{n+1}\right), \qquad \delta_n = |\widehat{q}_n - q_\star|.$$

If the two indicators in the definition of $\widehat{p}_{\triangle,n+1}$ differ, then

$$\left|\|T_{M^\star}(S_i)\| - q_\star\right| \le e_i + \delta_n.$$

Therefore, for any fixed $\tau > 0$,

$$\begin{aligned}\widehat{p}_{\triangle,n+1} &\le \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{\left|\|T_{M^\star}(S_i)\| - q_\star\right| \le e_i + \delta_n\} \\ &\le \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{\left|\|T_{M^\star}(S_i)\| - q_\star\right| \le \tau\} + \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{e_i > \tau/2\} + \mathbb{1}\{\delta_n > \tau/2\}.\end{aligned} \tag{28}$$

Step 2: Control the average map error. Let

$$A_{n+1}(M) = \left\{\frac{1}{n+1}\sum_{i=1}^{n+1} \|\widehat{T}_{n+1,M}(S_i) - T_M(S_i)\|^2\right\}^{1/2}.$$

For a fixed $M \in \mathcal{M}_0$, set $Z_i = MS_i$. For every permutation σ,

$$\sum_{i=1}^{n+1} c_M(S_i, U_{\sigma(i)}) = C_{M,S,U} - 2\sum_{i=1}^{n+1}(MS_i)^\top U_{\sigma(i)},$$
$$\sum_{i=1}^{n+1} \|Z_i - U_{\sigma(i)}\|^2 = C'_{M,S,U} - 2\sum_{i=1}^{n+1}(MS_i)^\top U_{\sigma(i)},$$

where neither $C_{M,S,U}$ nor $C'_{M,S,U}$ depends on σ. Hence the c_M-assignment of $S_{1:n+1}$ to $U_{1:n+1}$ is exactly the Euclidean discrete–discrete optimal transport assignment from $Z_{1:n+1}$ to $U_{1:n+1}$. The same identity at the population level shows that the Euclidean optimal transport map from $M_\#\mathcal{P}$ to $\mathcal{Q}$ is

$$z \longmapsto T_M(M^{-1}z).$$

Assumption 3.2(ii) and $\|M^{-1}\|_{\mathrm{op}} \le \underline{\lambda}^{-1}$ imply that these transformed maps are Lipschitz uniformly over $M \in \mathcal{M}_0$.

Let $\widehat{\mu}_{n+1,M} = (n+1)^{-1}\sum_{i=1}^{n+1}\delta_{Z_i}$, and $\widehat{\nu}_{n+1} = (n+1)^{-1}\sum_{i=1}^{n+1}\delta_{U_i}$. The selected permutation induces an optimal coupling between $\widehat{\mu}_{n+1,M}$ and $\widehat{\nu}_{n+1}$, and its barycentric projection at Z_i is $U_{\widehat{\sigma}_M^{Y_{n+1}}(i)}$. The supports of $M_\#\mathcal{P}$, $M \in \mathcal{M}_0$, are contained in a common compact set, and the transformed population maps have a common Lipschitz bound. Hence the constant in Corollary 2.3 of Deb et al. (2021) can be chosen uniformly over M, giving

$$\sup_{M\in\mathcal{M}_0} \mathbb{E}\{A_{n+1}(M)^2\} \le C r_d^{(n+1,n+1)} \longrightarrow 0, \tag{29}$$

where $r_d^{(n+1,n+1)} \to 0$ is their discrete–discrete empirical transport rate. Consequently, Jensen's inequality gives

$$\sup_{M\in\mathcal{M}_0} \mathbb{E}A_{n+1}(M) \longrightarrow 0. \tag{30}$$

We can now bound the second term of (28) by

$$\begin{aligned}
\frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{1}\{e_i > \tau/2\} &\overset{(i)}{\le} \frac{2}{\tau}\frac{1}{n+1}\sum_{i=1}^{n+1} e_i \\
&\overset{(ii)}{\le} \frac{2}{\tau}\frac{1}{n+1}\sum_{i=1}^{n+1}\|\widehat{T}_{n+1,\widehat{M}}(S_i) - T_{M^\star}(S_i)\| \\
&\overset{(iii)}{\le} \frac{2}{\tau}\left\{A_{n+1}(\widehat{M}) + \frac{1}{n+1}\sum_{i=1}^{n+1}\|T_{\widehat{M}}(S_i) - T_{M^\star}(S_i)\|\right\},
\end{aligned}$$

where (i) uses $\mathbb{1}\{e_i > \tau/2\} \le 2e_i/\tau$, (ii) is the reverse triangle inequality, and (iii) uses the triangle inequality and Cauchy–Schwarz.

Step 3: Bound the expected symmetric difference. Define

$$\pi_{n,N} = \mathbb{P}\{Y_{n+1} \in \mathcal{C}_\star^{\mathrm{MV}}(X_{n+1}) \triangle \widehat{\mathcal{C}}^{\mathrm{MV}}(X_{n+1})\}.$$

Write $\omega_\star(t) = \mathbb{P}\{\big|\|T_{M^\star}(S)\| - q_\star\big| \le t\}$ for a fresh residual $S \sim \mathcal{P}$. By exchangeability of the augmented calibration-test sample and permutation equivariance of the assignment,

$$\begin{aligned}
\pi_{n,N} &= \mathbb{E}[\widehat{p}_{\triangle,n+1}] \\
&\le \omega_\star(\tau) + \frac{2}{\tau}\left[\mathbb{E}\left\{A_{n+1}(\widehat{M})\right\} + \mathbb{E}\left\{\frac{1}{n+1}\sum_{i=1}^{n+1}\|T_{\widehat{M}}(S_i) - T_{M^\star}(S_i)\|\right\}\right] + \mathbb{P}(\delta_n > \tau/2).
\end{aligned}$$

Since $\widehat{M}$ is trained on the independent dataset $\mathcal{D}_{\mathrm{est}}$, condition on $\widehat{M}$, Assumption 3.2(iii) yields

$$\mathbb{E}\left[\frac{1}{n+1}\sum_{i=1}^{n+1}\|T_{\widehat{M}}(S_i) - T_{M^\star}(S_i)\| \,\middle|\, \widehat{M}\right] \le L_T\|\widehat{M} - M^\star\|_F.$$

Because $(T_{M^\star})_\# \mathcal{P} = \mathcal{Q}$, this modulus is determined by the radial law of the reference distribution. Assumption 3.2(i) implies that this radial law is continuous at $q_\star$, so $\omega_\star(t) \to 0$ as $t \to 0$. Then, for any fixed $\tau > 0$,

$$\pi_{n,N} \le \omega_\star(\tau) + \frac{2}{\tau}\left\{\mathbb{E}A_{n+1}(\widehat{M}) + L_T \mathbb{E}\|\widehat{M} - M^\star\|_F\right\} + \mathbb{P}(\delta_n > \tau/2). \tag{31}$$

Independence of $\widehat{M}$ from the augmented calibration and reference samples, together with (30), gives

$$\mathbb{E}A_{n+1}(\widehat{M}) \le \sup_{M \in \mathcal{M}_0} \mathbb{E}A_{n+1}(M) \longrightarrow 0 \qquad \text{as } n \to \infty, \tag{32}$$

uniformly in N. By Theorem 3.2, $\mathbb{E}\|\widehat{M} - M^\star\|_F \to 0$ as $N \to \infty$.

Step 4: Establish cutoff consistency. To control the remaining cutoff term, let

$$\widehat{F}_{n+1}(t) = \frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{\|U_i\| \le t\}.$$

By the theorem's i.i.d. reference condition, the variables $\|U_i\|$ are i.i.d. with distribution function F_Q. The Glivenko–Cantelli theorem gives

$$\sup_{t \in \mathbb{R}} |\widehat{F}_{n+1}(t) - F_Q(t)| \longrightarrow 0 \qquad \text{almost surely.}$$

Since F_Q is continuously differentiable near $q_\star$ and $F_Q'(q_\star) > 0$, it is strictly increasing there; empirical-quantile consistency therefore gives $\widehat{q}_n \to q_\star$ almost surely and hence $\delta_n \to 0$ in probability. Combining this fact with (31), (32), and Theorem 3.2, for every fixed $\tau > 0$,

$$\limsup_{\min\{N,n\}\to\infty} \pi_{n,N} \le \omega_\star(\tau).$$

Finally let $\tau \downarrow 0$. The continuity of F_Q at $q_\star$ gives $\omega_\star(t) \to 0$, proving the claimed joint-limit result. □

F.5 Proof of Theorem 4.1

Proof. Let $\mathcal{H}_N$ be the σ-algebra generated by $\mathcal{D}_{\text{est}}$, reference draws, and pretrained predictor $\widehat{\mu}$. Conditional on $\mathcal{H}_N$, the score function $(x, y) \mapsto \widehat{R}_N(x, y)$ is fixed. By the theorem's independence and exchangeability assumptions, $V_i := \widehat{R}_N(X_i, Y_i)$, $i \in [n+1]$, are exchangeable. Put $k = \lceil (n+1)(1-\alpha) \rceil$, and let q_{cal} be the kth order statistic of $\{V_1, \dots, V_n, \infty\}$ and q_{aug} the kth order statistic of $\{V_1, \dots, V_{n+1}\}$. Since $V_{n+1} \le \infty$, $q_{\text{aug}} \le q_{\text{cal}}$ for every realization. Hence

$$\begin{aligned}
\mathbb{P}\{V_{n+1} \le q_{\text{cal}} \mid \mathcal{H}_N\} &\ge \mathbb{P}\{V_{n+1} \le q_{\text{aug}} \mid \mathcal{H}_N\} \\
&= \mathbb{E}\left[\frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\{V_i \le q_{\text{aug}}\} \,\middle|\, \mathcal{H}_N\right] \\
&\ge \frac{k}{n+1} \ge 1 - \alpha.
\end{aligned}$$

The first event is exactly $\{Y_{n+1} \in \widehat{\mathcal{C}}^{\text{IMV}}(X_{n+1})\}$ by (9). Taking expectations over $\mathcal{H}_N$ proves the claim. □

F.6 Proof of Theorem B.1

Proof. **Step 1: Reduce the local empirical process to a finite metric net.** Fix x. It is enough to control

$$A_{N,x}^{\text{obj}} := \sup_{M \in \mathcal{M}_0} |\mathcal{J}_{N,x}(M) - \mathcal{J}_x(M)|,$$

since the triangle inequality give

$$\sup_{M \in \mathcal{M}_0} |\widehat{\mathcal{J}}_x(M, \widehat{\theta}_x(M)) - \mathcal{J}_x(M)| \le \mathcal{E}_N^{\text{obj}}(x) + A_{N,x}^{\text{obj}}.$$

Let $m = d(d+1)/2$ and $R_{\mathcal{M}} = \sqrt{d}\,\overline{\lambda}$. For $\varepsilon > 0$, let $\{M_1, \dots, M_K\}$ be an ε-net of $\mathcal{M}_0$ in Frobenius norm. Since $\mathcal{M}_0$ is contained in the radius-$R_{\mathcal{M}}$ ball in the m-dimensional space of symmetric matrices, Lemma F.1 gives

$$K \le \left(1 + \frac{2R_{\mathcal{M}}}{\varepsilon}\right)^m.$$

For every net point, define the local oscillation envelope

$$H_j(u) := \sup_{M \in \mathcal{M}_0(M_j, \varepsilon)} |\zeta_{M,x}(u) - \zeta_{M_j,x}(u)|.$$

Assumption B.1(i) implies

$$0 \le H_j(U) \le 2B_x, \qquad \mathbb{E}_U H_j(U) \le L_x \varepsilon.$$

For each $M \in \mathcal{M}_0$, choose $j(M)$ such that $\|M - M_{j(M)}\|_F \le \varepsilon$. Then

$$\begin{aligned}
|\mathcal{J}_{N,x}(M) - \mathcal{J}_x(M)| &\le |\mathcal{J}_{N,x}(M_{j(M)}) - \mathcal{J}_x(M_{j(M)})| \\
&\quad + |\mathcal{J}_{N,x}(M) - \mathcal{J}_{N,x}(M_{j(M)})| + |\mathcal{J}_x(M) - \mathcal{J}_x(M_{j(M)})| \\
&\le |\mathcal{J}_{N,x}(M_{j(M)}) - \mathcal{J}_x(M_{j(M)})| \\
&\quad + \frac{1}{N}\sum_{i=1}^N H_{j(M)}(U_i) + \mathbb{E}_U H_{j(M)}(U).
\end{aligned}$$

Consequently,

$$A_{N,x}^{\mathrm{obj}} \le \max_{1 \le j \le K} |\mathcal{J}_{N,x}(M_j) - \mathcal{J}_x(M_j)| + \max_{1 \le j \le K} \frac{1}{N}\sum_{i=1}^N H_j(U_i) + L_x\varepsilon. \tag{33}$$

Step 2: Apply concentration on the net. Because $\zeta_{M_j,x}(U) \in [-B_x, B_x]$ and $H_j(U) \in [0, 2B_x]$, Hoeffding's inequality and a union bound give, for every $t > 0$,

$$\mathbb{P}\left\{\max_{1 \le j \le K} |\mathcal{J}_{N,x}(M_j) - \mathcal{J}_x(M_j)| > t\right\} \le 2K \exp\left(-\frac{Nt^2}{2B_x^2}\right),$$

$$\mathbb{P}\left\{\max_{1 \le j \le K}\left\{\frac{1}{N}\sum_{i=1}^N H_j(U_i) - \mathbb{E}_U H_j(U)\right\} > t\right\} \le K \exp\left(-\frac{Nt^2}{2B_x^2}\right).$$

Taking

$$t = B_x\sqrt{\frac{2\{\log(3K) + \log(1/\delta)\}}{N}}$$

and applying another union bound in (33) shows that, with probability at least $1 - \delta$,

$$A_{N,x}^{\mathrm{obj}} \le CB_x\sqrt{\frac{m\log(1 + 2R_{\mathcal{M}}/\varepsilon) + \log(2/\delta)}{N}} + 2L_x\varepsilon \tag{34}$$

for a universal constant C.

Step 3: Choose the net resolution. Suppose first that $L_x > 0$ and take $\varepsilon_N = B_x/(L_xN)$. Then $2L_x\varepsilon_N = 2B_x/N$, $m \le d^2$, and $R_{\mathcal{M}} = \sqrt{d}\,\overline{\lambda}$. Enlarging C in (34) therefore gives

$$A_{N,x}^{\mathrm{obj}} \le CB_x\left[\frac{d^2}{N}\log\left\{e + \frac{2\sqrt{d}\,\overline{\lambda}L_xN}{B_x}\right\} + \frac{\log(2/\delta)}{N}\right]^{1/2}.$$

Step 4: Conclude pointwise metric consistency. Taking the failure probability N^{-1} for $N \ge 2$ in the preceding empirical-process bound gives a deterministic right-hand side converging to zero for fixed d. Therefore

$$A_{N,x}^{\mathrm{obj}} = o_{\mathbb{P}}(1).$$

Assumption B.1(ii) now gives

$$D_{N,x} := \sup_{M \in \mathcal{M}_0} |\widehat{\mathcal{J}}_x\{M, \widehat{\theta}_x(M)\} - \mathcal{J}_x(M)| = o_{\mathbb{P}}(1).$$

Finally, empirical optimality of $\widehat{M}_x$ and $M_x^\star \in \mathcal{M}_0$ imply

$$\begin{aligned}\mathcal{J}_x(\widehat{M}_x) &\le \widehat{\mathcal{J}}_x\{\widehat{M}_x, \widehat{\theta}_x(\widehat{M}_x)\} + D_{N,x} \\ &\le \widehat{\mathcal{J}}_x\{M_x^\star, \widehat{\theta}_x(M_x^\star)\} + D_{N,x} \\ &\le \mathcal{J}_x(M_x^\star) + 2D_{N,x}.\end{aligned}$$

For every $\varepsilon > 0$, define

$$\mathrm{gap}_x(\varepsilon) := \inf_{M \in \mathcal{M}_0:\ \|M - M_x^\star\|_F \ge \varepsilon} \{\mathcal{J}_x(M) - \mathcal{J}_x(M_x^\star)\}.$$

Assumption B.1(iii) gives $\mathrm{gap}_x(\varepsilon) > 0$, and therefore

$$\{\|\widehat{M}_x - M_x^\star\|_F \ge \varepsilon\} \subseteq \{\mathrm{gap}_x(\varepsilon) \le 2D_{N,x}\}.$$

Consequently,

$$\mathbb{P}\{\|\widehat{M}_x - M_x^\star\|_F \ge \varepsilon\} \le \mathbb{P}\{2D_{N,x} \ge \mathrm{gap}_x(\varepsilon)\} \longrightarrow 0.$$

□

F.7 Proof of Theorem 4.2

Proof. **Step 1: Control the average radius error.** Let

$$R_i^\star = R^\star(X_i, Y_i), \qquad \widehat{R}_i = \widehat{R}_N(X_i, Y_i), \qquad i \in [n].$$

Define the average radius error

$$A_{n,N} := \frac{1}{n}\sum_{i=1}^{n} |\widehat{R}_i - R_i^\star|.$$

By Assumption 4.1, for almost every x, $\mathcal{E}_N^{\mathrm{rad}}(x) \to 0$ in probability and $0 \le \mathcal{E}_N^{\mathrm{rad}}(x) \le B_{\mathrm{rad}}(x)$. Bounded convergence in probability, followed by dominated convergence over P_X, therefore gives

$$\mathbb{E}\left\{\int \mathcal{E}_N^{\mathrm{rad}}(x)\, dP_X(x)\right\} \longrightarrow 0.$$

Conditional on $\mathcal{H}_N$, the calibration observations are i.i.d. and

$$\begin{aligned}\mathbb{E}(A_{n,N} \mid \mathcal{H}_N) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(|\widehat{R}_N(X_i, Y_i) - R^\star(X_i, Y_i)| \mid \mathcal{H}_N\right) \\ &= \int \mathcal{E}_N^{\mathrm{rad}}(x)\, dP_X(x).\end{aligned}$$

Consequently, for every $\varepsilon > 0$, Markov's inequality gives

$$\mathbb{P}(A_{n,N} > \varepsilon) \le \frac{1}{\varepsilon}\mathbb{E}\left\{\int \mathcal{E}_N^{\mathrm{rad}}(x)\, dP_X(x)\right\} \longrightarrow 0,$$

and hence $A_{n,N} = o_{\mathbb{P}}(1)$.

Step 2: Compare the fitted and oracle empirical CDFs. For $t \in \mathbb{R}$, define the two empirical distribution functions

$$F_n^\star(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{R_i^\star \le t\}, \qquad \widehat{F}_{n,N}(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\widehat{R}_i \le t\}.$$

For $\eta > 0$, also define the fraction of radius errors exceeding η:

$$D_{n,N}(\eta) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{|\widehat{R}_i - R_i^\star| > \eta\}.$$

Since $\eta \mathbb{1}\{|\widehat{R}_i - R_i^\star| > \eta\} \le |\widehat{R}_i - R_i^\star|$, we have

$$D_{n,N}(\eta) \le \frac{A_{n,N}}{\eta} = o_{\mathbb{P}}(1)$$

for every fixed $\eta > 0$. Moreover, pointwise in the calibration sample,

$$\mathbb{1}\{R_i^\star \le t - \eta\} - \mathbb{1}\{|\widehat{R}_i - R_i^\star| > \eta\} \le \mathbb{1}\{\widehat{R}_i \le t\},$$
$$\mathbb{1}\{\widehat{R}_i \le t\} \le \mathbb{1}\{R_i^\star \le t + \eta\} + \mathbb{1}\{|\widehat{R}_i - R_i^\star| > \eta\}.$$

After averaging over i,

$$F_n^\star(t - \eta) - D_{n,N}(\eta) \le \widehat{F}_{n,N}(t) \le F_n^\star(t + \eta) + D_{n,N}(\eta).$$

Step 3: Control the oracle empirical CDF. For P_X-almost every X_i, the oracle transport satisfies $T_{M^\star_{X_i}, X_i}\{s(X_i, Y_i)\} \mid X_i \sim \mathcal{Q}$. Consequently,

$$\mathbb{P}(R_i^\star \le t \mid X_i, \mathcal{H}_N) = \mathbb{P}_{U \sim \mathcal{Q}}(\|U\| \le t) = F_Q(t),$$
$$\mathbb{P}(R_i^\star \le t \mid \mathcal{H}_N) = F_Q(t).$$

Thus, conditionally on $\mathcal{H}_N$, $R_1^\star, \dots, R_n^\star$ are i.i.d. with distribution function F_Q. The Dvoretzky–Kiefer–Wolfowitz inequality gives, for every $z > 0$,

$$\mathbb{P}\left\{ \sup_{t \in \mathbb{R}} |F_n^\star(t) - F_Q(t)| > z \;\middle|\; \mathcal{H}_N \right\} \le 2e^{-2nz^2}.$$

Therefore,

$$\sup_{t \in \mathbb{R}} |F_n^\star(t) - F_Q(t)| = o_{\mathbb{P}}(1).$$

Step 4: Prove consistency of the calibration cutoff. Let $k_n := \lceil (n+1)(1-\alpha) \rceil$ with $\tau_n := \frac{k_n}{n}$. For all sufficiently large n, $k_n \le n$, and $\tau_n \longrightarrow 1 - \alpha$. Let $\widehat{q}_{n,N}$ denote the empirical cutoff in (9). By the $+\infty$ convention,

$$\widehat{q}_{n,N} = \inf\{t : \widehat{F}_{n,N}(t) \ge \tau_n\}.$$

Fix $\varepsilon > 0$ and put $\eta = \varepsilon/2$. Strict increase of F_Q at $q_\star$ gives

$$g_-(\varepsilon) := (1 - \alpha) - F_Q(q_\star - \varepsilon/2) > 0,$$
$$g_+(\varepsilon) := F_Q(q_\star + \varepsilon/2) - (1 - \alpha) > 0.$$

The empirical-CDF sandwich at $t = q_\star - \varepsilon$ gives

$$\begin{aligned} \widehat{F}_{n,N}(q_\star - \varepsilon) &\le F_n^\star(q_\star - \varepsilon/2) + D_{n,N}(\varepsilon/2) \\ &\le F_Q(q_\star - \varepsilon/2) + \sup_t |F_n^\star(t) - F_Q(t)| + D_{n,N}(\varepsilon/2) \\ &= 1 - \alpha - g_-(\varepsilon) + o_{\mathbb{P}}(1) < \tau_n \end{aligned}$$

with probability tending to one. Similarly,

$$\begin{aligned} \widehat{F}_{n,N}(q_\star + \varepsilon) &\ge F_n^\star(q_\star + \varepsilon/2) - D_{n,N}(\varepsilon/2) \\ &\ge F_Q(q_\star + \varepsilon/2) - \sup_t |F_n^\star(t) - F_Q(t)| - D_{n,N}(\varepsilon/2) \\ &= 1 - \alpha + g_+(\varepsilon) + o_{\mathbb{P}}(1) > \tau_n \end{aligned}$$

with probability tending to one. Hence

$$\mathbb{P}\left\{ q_\star - \varepsilon < \widehat{q}_{n,N} \le q_\star + \varepsilon \right\} \longrightarrow 1,$$

which proves

$$\widehat{q}_{n,N} \xrightarrow{\mathbb{P}} q_\star.$$

Step 5: Prove conditional set convergence. Now fix an interior x satisfying the pointwise conditions and condition on $X_{n+1} = x$. The test response Y_{n+1} is independent of the calibration sample and $\mathcal{H}_N$ under this conditioning. Abbreviate

$$R_x^\star := R^\star(x, Y_{n+1}), \qquad \widehat{R}_{N,x} := \widehat{R}_N(x, Y_{n+1}).$$

For every $\varepsilon > 0$, the following pointwise indicator inequality holds:

$$\begin{aligned}&\left|\mathbb{1}\{\widehat{R}_{N,x} \le \widehat{q}_{n,N}\} - \mathbb{1}\{R_x^\star \le q_\star\}\right| \\ &\quad \le \mathbb{1}\{|R_x^\star - q_\star| \le 2\varepsilon\} + \mathbb{1}\{|\widehat{R}_{N,x} - R_x^\star| > \varepsilon\} + \mathbb{1}\{|\widehat{q}_{n,N} - q_\star| > \varepsilon\}.\end{aligned}$$

Indeed, on the complement of the three events, if $R_x^\star < q_\star - 2\varepsilon$, then

$$\widehat{R}_{N,x} \le R_x^\star + \varepsilon < q_\star - \varepsilon \le \widehat{q}_{n,N},$$

whereas if $R_x^\star > q_\star + 2\varepsilon$, then

$$\widehat{R}_{N,x} \ge R_x^\star - \varepsilon > q_\star + \varepsilon \ge \widehat{q}_{n,N}.$$

Taking expectations in the indicator inequality conditional on $X = x$ gives

$$\begin{aligned}&\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}_\star^{\mathrm{IMV}}(x) \triangle \widehat{\mathcal{C}}^{\mathrm{IMV}}(x) \;\middle|\; X_{n+1} = x\right\} \\ &\quad \le F_Q(q_\star + 2\varepsilon) - F_Q((q_\star - 2\varepsilon)-) \\ &\qquad + \mathbb{E}\left[\min\left\{1, \frac{\mathcal{E}_N^{\mathrm{rad}}(x)}{\varepsilon}\right\}\right] + \mathbb{P}(|\widehat{q}_{n,N} - q_\star| > \varepsilon).\end{aligned}$$

Here conditional Markov's inequality gives

$$\mathbb{P}\left(|\widehat{R}_{N,x} - R_x^\star| > \varepsilon \mid \mathcal{H}_N, X_{n+1} = x\right) \le \min\left\{1, \frac{\mathcal{E}_N^{\mathrm{rad}}(x)}{\varepsilon}\right\}.$$

Since $\mathcal{E}_N^{\mathrm{rad}}(x) = o_{\mathbb{P}}(1)$ by Assumption 4.1, bounded convergence in probability implies

$$\mathbb{E}\left[\min\left\{1, \frac{\mathcal{E}_N^{\mathrm{rad}}(x)}{\varepsilon}\right\}\right] \longrightarrow 0.$$

The quantile term also tends to zero. Thus

$$\begin{aligned}&\limsup_{N,n\to\infty} \mathbb{P}\left\{Y_{n+1} \in \mathcal{C}_\star^{\mathrm{IMV}}(x) \triangle \widehat{\mathcal{C}}^{\mathrm{IMV}}(x) \;\middle|\; X_{n+1} = x\right\} \\ &\qquad \le F_Q(q_\star + 2\varepsilon) - F_Q((q_\star - 2\varepsilon)-).\end{aligned}$$

Letting $\varepsilon \downarrow 0$ and using continuity of F_Q at $q_\star$ proves the symmetric-difference convergence.

Step 6: Deduce asymptotic conditional coverage. Finally,

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}_\star^{\mathrm{IMV}}(x) \mid X_{n+1} = x\} = \mathbb{P}(R_x^\star \le q_\star \mid X_{n+1} = x) = F_Q(q_\star) = 1 - \alpha.$$

For any two events A and B,

$$|\mathbb{P}(A) - \mathbb{P}(B)| \le \mathbb{P}(A \triangle B).$$

Applying this inequality conditionally on $X_{n+1} = x$ gives

$$\begin{aligned}&\left|\mathbb{P}\{Y_{n+1} \in \widehat{\mathcal{C}}^{\mathrm{IMV}}(x) \mid X_{n+1} = x\} - (1-\alpha)\right| \\ &\quad \le \mathbb{P}\left\{Y_{n+1} \in \mathcal{C}_\star^{\mathrm{IMV}}(x) \triangle \widehat{\mathcal{C}}^{\mathrm{IMV}}(x) \;\middle|\; X_{n+1} = x\right\} \longrightarrow 0.\end{aligned}$$

□

F.8 Proof of Proposition C.1

Proof. The feasible set $\Pi_{\tilde{x}}$ is a bounded transportation polytope, and its objective is linear in π. Hence an optimal plan can be selected from the finite collection of basic feasible plans indexed by $\mathfrak{B}_{\tilde{x}}$. If the trial score is s, the objective value at the plan π^b is

$$\begin{aligned}L_{b,M}(s) &= \sum_{i=1}^{n}\sum_{j=1}^{n+1} \pi_{ij}^b c_M(S_i, U_j) + \sum_{j=1}^{n+1} \pi_{n+1,j}^b c_M(s, U_j) \\ &= p_{n+1} s^\top M s - 2 s^\top M m_b + \kappa_{b,M},\end{aligned}$$

where the last equality uses $\sum_j \pi^b_{n+1,j} = p_{n+1}$. Consequently, $L_{b,M}(s) \le L_{b',M}(s)$ if and only if

$$2s^\top M(m_{b'} - m_b) \le \kappa_{b',M} - \kappa_{b,M}.$$

The common quadratic term has canceled. Intersecting these halfspaces over $b' \in \mathfrak{B}_{\tilde{x}}$ gives precisely $\mathcal{R}_{b,M}(\tilde{x})$, so its closure is a convex polyhedron.

Within the relative interior of this cell, the selected plan is π^b. Therefore $J_b(V)$ and the randomized acceptance decision are fixed throughout the cell. Likewise, all barycenters $\overline{U}^b_i$, their weighted quantile q_b, and the barycentric acceptance decision are fixed throughout the cell. Taking the union over the accepting bases gives the two displayed representations. On a hyperplane where several bases are optimal, a deterministic tie-breaking rule assigns the boundary to the selected cell. Translation by $\widehat{\mu}(X_{n+1})$ then gives the response-space representation. □

F.9 Proof of Theorem C.1

Proof. Set $y = Y_{n+1}$, $Z_i = (X_i, Y_i)$, and let $\mathcal{D}_{\text{aug}} = \{Z_1, \dots, Z_{n+1}\}$ be the unordered augmented multiset. Choose an arbitrary ordering of $\mathcal{D}_{\text{aug}}$, and let $I \in [n+1]$ denote the index of the occurrence in this ordering corresponding to the original test observation Z_{n+1}. Let $\mathcal{G} = \sigma(\mathcal{D}_{\text{aug}}, \widetilde{X}_{n+1}, U_{1:n+1})$ be the σ-algebra generated by the data. Conditional on $\widetilde{X}_{n+1} = \tilde{x}$, exchangeability and Bayes' rule give

$$\mathbb{P}(I = i \mid \mathcal{G}) = \frac{H_h(\tilde{x} \mid x_i)}{\sum_{k=1}^{n+1} H_h(\tilde{x} \mid x_k)} = p_i(\tilde{x}).$$

By permutation equivariance, the fitted scores and plan are fixed in this ordering. Write $p_i = p_i(\tilde{x})$ and $\widehat{\pi}_{ij}$ for the plan at the true response which satisfies

$$\sum_j \widehat{\pi}_{ij} = p_i, \qquad \sum_i \widehat{\pi}_{ij} = \frac{1}{n+1}.$$

For the randomized rule, only rows with $p_i > 0$ can be selected. The common uniform draw used for inverse-cdf sampling yields

$$\mathbb{P}(J = j \mid I = i, \mathcal{G}) = \frac{\widehat{\pi}_{ij}}{p_i}, \qquad p_i > 0.$$

Hence,

$$\mathbb{P}(J = j \mid \mathcal{G}) = \sum_{i: p_i > 0} p_i \frac{\widehat{\pi}_{ij}}{p_i} = \sum_i \widehat{\pi}_{ij} = \frac{1}{n+1}.$$

With $\widehat{q}_U = \widehat{q}_{1-\alpha}\left(\{\|U_j\|\}_{j=1}^{n+1}\right)$,

$$\begin{aligned}\mathbb{P}\{\|U_J\| \le \widehat{q}_U \mid \mathcal{G}\} &= \sum_{j=1}^{n+1} \mathbb{P}(J = j \mid \mathcal{G}) \mathbb{1}\{\|U_j\| \le \widehat{q}_U\} \\ &= \frac{1}{n+1} \sum_{j=1}^{n+1} \mathbb{1}\{\|U_j\| \le \widehat{q}_U\} \ge 1 - \alpha.\end{aligned}$$

For the barycentric rule, let $R_i = R^{Y_{n+1}}_{\tilde{x},i}$ for $p_i > 0$, and let $\widehat{q} = \widehat{q}^{Y_{n+1}}_{1-\alpha}(\tilde{x})$. Then

$$\begin{aligned}\mathbb{P}\{R_I \le \widehat{q} \mid \mathcal{G}\} &= \sum_{i: p_i > 0} \mathbb{P}(I = i \mid \mathcal{G}) \mathbb{1}\{R_i \le \widehat{q}\} \\ &= \sum_{i: p_i > 0} p_i \mathbb{1}\{R_i \le \widehat{q}\} \ge 1 - \alpha.\end{aligned}$$

Taking expectations over $\mathcal{G}$ proves both claims. □

F.10 Proof of Theorem C.2

Proof. $\widehat{q}_U = \widehat{q}_{1-\alpha}\left(\{\|U_j\|\}_{j=1}^{n+1}\right)$. Since $\mathcal{D}_{\text{est}}$ is independent of the augmented sample, conditioning additionally on $\mathcal{D}_{\text{est}}$ does not change the conditional distribution of the test index. Hence

$$\mathbb{P}(I = i \mid \mathcal{G}, \mathcal{D}_{\text{est}}) = \mathbb{P}(I = i \mid \mathcal{G}) = p_i(\tilde{x}).$$

Moreover, conditional on $(\mathcal{G}, \mathcal{D}_{\text{est}})$, $\widehat{M}(\tilde{x})$ and the corresponding transport plan $\widehat{\pi}$ are fixed. By permutation equivariance of the construction, the arbitrary ordering of the augmented observations is immaterial, and the transport plan satisfies

$$\sum_j \widehat{\pi}_{ij} = p_i(\tilde{x}), \qquad \sum_i \widehat{\pi}_{ij} = \frac{1}{n+1}.$$

For the randomized metric rule,

$$\mathbb{P}(J^{\widehat{M}} = j \mid \mathcal{G}, \mathcal{D}_{\text{est}}) = \sum_{i:p_i>0} p_i \frac{\widehat{\pi}_{ij}}{p_i} = \frac{1}{n+1},$$

$$\mathbb{P}\{\|U_{J^{\widehat{M}}}\| \le \widehat{q}_U \mid \mathcal{G}, \mathcal{D}_{\text{est}}\} = \frac{1}{n+1}\sum_{j=1}^{n+1} \mathbb{1}\{\|U_j\| \le \widehat{q}_U\} \ge 1-\alpha.$$

For the barycentric metric rule, write $R_i^{\widehat{M}} = R_{\tilde{x},i}^{Y_{n+1},\widehat{M}}$ and $\widehat{q}^{\widehat{M}} = \widehat{q}_{1-\alpha}^{Y_{n+1},\widehat{M}}(\tilde{x})$. Then

$$\mathbb{P}\{R_I^{\widehat{M}} \le \widehat{q}^{\widehat{M}} \mid \mathcal{G}, \mathcal{D}_{\text{est}}\} = \sum_{i:p_i>0} p_i \mathbb{1}\{R_i^{\widehat{M}} \le \widehat{q}^{\widehat{M}}\} \ge 1-\alpha.$$

Taking expectations over $(\mathcal{G}, \mathcal{D}_{\text{est}})$ proves both claims. □