LEBESGUE-CONTRASTIVE PREDICTION FOR MULTIVARIATE CONFORMAL REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

The minimum-volume prediction region with conditional coverage $1-\alpha$ is a superlevel set of the conditional density, known as the conditional highest-density region (HDR), so a scalar score reproducing this level-set geometry suffices for optimal multivariate regression. We propose Lebesgue-contrastive prediction (LECP), which learns such a score by logistic discrimination of observed responses against a Lebesgue reference measure, requiring neither a normalized density model nor an invertible parameterization. An alternating procedure fits the contrastive score and a conditional-quantile level that tracks the target density contour, followed by split conformal calibration on an independent hold-out set. The resulting regions enjoy finite-sample marginal coverage by construction and, under estimation consistency, converge to the conditional HDR. On synthetic and real multivariate regression benchmarks, LECP yields substantial volume reductions over existing methods while maintaining competitive conditional reliability.

1 INTRODUCTION

Let $X \in \mathcal{X}$ denote covariates, let $Y \in \mathcal{Y} \subseteq \mathbb{R}^d$ be a multivariate response, and let μ denote Lebesgue measure on $\mathcal{Y}$. For a target miscoverage level $\alpha \in (0,1)$, consider the conditional minimum-volume problem over Lebesgue-measurable subsets $C \subseteq \mathcal{Y}$:

$$\underset{C \subseteq \mathcal{Y}}{\text{minimize}} \quad \mu(C) \qquad \text{subject to} \quad \mathbb{P}(Y \in C \mid X = x) \geq 1-\alpha. \tag{1}$$

When the conditional law admits a density $p(\cdot \mid x)$ with a regular target contour, the solution is the conditional highest-density region (HDR) $C^\star_\alpha(x) = \{y : p(y \mid x) \geq c_\alpha(x)\}$, where $c_\alpha(x)$ is the largest level such that the superlevel set has conditional probability at least $1-\alpha$ (Hyndman, 1996; Polonik, 1995; Polonik & Yao, 2000). Conditional density level sets therefore define the population-optimal geometry for this problem.

This characterization suggests a direct route to conformal prediction (CP) regions (Papadopoulos et al., 2002; Vovk et al., 2005; Shafer & Vovk, 2008): a scalar score whose sublevel sets coincide with conditional density superlevel sets recovers $C^\star_\alpha(x)$ at an appropriate covariate-dependent level. Existing methods obtain such geometry from a fitted conditional distribution or generative model (Dheur et al., 2025), or by optimizing prediction-set volume within a restricted set class (Braun et al., 2026a). However, note that level-set geometry is invariant to the x-dependent normalizing constant, so contrastive objectives, which learn unnormalized log-densities from reference comparisons, recover the same geometry without density normalization.

We propose Lebesgue-contrastive prediction (LECP). Instead of fitting a normalized density by maximum likelihood, LECP parameterizes a scalar score $s_\theta(x,y)$ and learns it through logistic contrast between observed responses and a Lebesgue reference measure with intensity $\lambda > 0$. For a fixed temperature $\beta > 0$, the population Bayes solution is

$$s^\star(x,y) = \beta \log \frac{\lambda}{p(y \mid x)}, \tag{2}$$

whose sublevel sets are exactly the conditional density superlevel sets, yet requiring neither density normalization nor invertibility. Since the minimum-volume contour $c_\alpha(x)$ corresponds to a covariate-dependent level of this score, LECP jointly fits a conditional-quantile network $t_\phi(x)$

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.

that tracks the $(1-\alpha)$-quantile of $s_\theta(x, Y)$ via a pinball loss. The centered score $g_{\theta,\phi}(x, y) := s_\theta(x, y) - t_\phi(x)$ then has its zero sublevel set aligned with the minimum-volume boundary. Training alternates contrastive updates of s_θ with quantile updates of t_ϕ; the resulting centered score serves as the nonconformity score for split conformal calibration on an independent hold-out set.

We make three main contributions: (1) we introduce LECP, a contrastive framework that learns conditional density-level geometry without explicit density normalization or invertible parameterizations, together with an alternating training procedure and an iterative residual variant (Res-LECP); (2) we show that the population minimizer of the contrastive risk recovers the conditional density geometry and, under estimation consistency, the calibrated regions converge in set to the conditional minimum-volume oracle, while finite-sample marginal coverage follows from standard split conformal calibration; (3) on the evaluated synthetic and real benchmarks, LECP yields substantial volume reductions over existing methods while maintaining competitive conditional reliability.

2 Related Work

Multivariate conformal regression. Beyond previous copula-based approaches (Messoudi et al., 2021), recent work has developed diverse conformity scores for flexible multi-output prediction regions. HPD-split thresholds an estimated conditional density into HDRs (Izbicki et al., 2022); PCP builds regions from conditional generative-model samples (Wang et al., 2023); and CP2-PCP extends this construction toward approximate conditional validity (Plassier et al., 2025b). Dheur et al. (2025) unify these approaches and introduce CDF-based and latent-space conformity scores, yielding the variants C-HDR, C-PCP, and L-CP. Building on normalizing flows, CONTRA (Fang et al., 2025) and JAPAN (English & Lippert, 2026) define density- or distance-based scores in the learned latent space, while DSPS (Luo & Zhou, 2026) and latent recalibration (Dheur & Ben Taieb, 2025) further extend this flow-based direction. Moreover, transport-based methods map multivariate outputs to an ordered reference space (Thurin et al., 2025; Klein et al., 2025; Ndiaye, 2025), offering an alternative route to flexible region geometry.

Minimum-volume and highest-density prediction. HDRs are the minimum-volume sets of prescribed probability content (Hyndman, 1996; Polonik, 1995); Polonik & Yao (2000) study their conditional analogue. Scott & Nowak (2006) estimate minimum-volume sets over restricted set classes, and Lei et al. (2013) connect density level sets to distribution-free prediction. More recently, conformal efficiency has been studied explicitly through volume optimization. Gao et al. (2025) characterize volume optimality under structured prediction-set classes, and Braun et al. (2026a) learn minimum-volume conformal sets within adaptive norm-based families. Most closely related to our objective is super-level-set regression (SLS) (Braun et al., 2026c), which directly parameterizes conditional level-set boundaries and optimizes a surrogate of their volume using volume-preserving flow components. LECP instead represents the target region through a scalar score and a covariate-dependent level, learned with alternating contrastive and conditional-quantile updates.

Contrastive density estimation and conditional score calibration. Learning a density or density ratio by classifying data against a known reference is a classical idea. Noise-contrastive estimation (NCE) fits unnormalized models by discriminating data from noise (Gutmann & Hyvärinen, 2010; Gutmann & Hyvärinen, 2012); conditional and density-ratio extensions appear in Ceylan & Gutmann (2018) and Sugiyama et al. (2010). CINDES applies the same principle to neural conditional density estimation (Dai et al., 2025). LECP shares the classification-to-density reduction but differs in two respects: it draws localized reference samples via importance sampling to concentrate estimation effort on data-supported regions, while its population objective remains a Lebesgue contrast; and it pairs the contrastive score with a conditional-level learner that identifies the covariate-dependent level defining the minimum-volume region. This is similar to the principle of alternating quantile-based interval optimization (Su et al., 2026) and estimating conditional quantiles of conformity scores as in score rectification (Plassier et al., 2025a).

3 Lebesgue-Contrastive Prediction

LECP learns a scalar score $s_\theta(x, y)$ together with a covariate-dependent level $t_\phi(x)$. The two components are fitted on a proper-training set $\mathcal{D}_{\text{train}}$, selected jointly on a validation set $\mathcal{D}_{\text{val}}$, and fi-

nally conformalized on an independent calibration set $\mathcal{D}_{\text{cal}}$. During training, s_θ and t_ϕ are updated alternately: the former is learned by a Lebesgue-contrastive objective, while the latter tracks the conditional $(1-\alpha)$ quantile of the evolving score.

3.1 Lebesgue-Contrastive Score Learning

We learn s_θ by discriminating observed responses from a Lebesgue reference measure. For temperature $\beta > 0$, define $\ell_\beta^+(u) = \beta \log(1 + e^{u/\beta})$ and $\ell_\beta^-(u) = \beta \log(1 + e^{-u/\beta})$. Contrasting the conditional data distribution against the scaled reference measure $\lambda\mu$ leads to the population risk

$$\mathcal{R}_{\lambda,\beta}(s) = \mathbb{E}\left[\ell_\beta^+\left(s(X,Y)\right)\right] + \lambda\, \mathbb{E}_X \int_{\mathcal{Y}} \ell_\beta^-\left(s(X,y)\right)\, d\mu(y) \tag{3}$$

$$= \mathbb{E}_X \int_{\mathcal{Y}} \left\{ p(y \mid X)\ell_\beta^+(s(X,y)) + \lambda \ell_\beta^-(s(X,y)) \right\} d\mu(y). \tag{4}$$

As shown in Section 1, the population minimizer is density-equivalent: lower scores correspond exactly to higher conditional density. The first term in (3) pushes scores down at observed responses, while the Lebesgue term penalizes assigning low scores over unnecessarily large regions of $\mathcal{Y}$.

Monte Carlo approximation. The Lebesgue integral is generally unavailable in closed form. Let $\mathcal{D}_{\text{train}} = \{(X_i, Y_i)\}_{i=1}^n$. For each training observation, we draw $Y_{ij}^- \sim q_i$, $j = 1, \ldots, M$, from a proposal density q_i whose support covers $\mathcal{Y}$ up to μ-null sets, and estimate the integral by importance sampling. The resulting score loss is

$$\widehat{\mathcal{R}}_{n,M}(\theta) = \frac{1}{n}\sum_{i=1}^n \left[\ell_\beta^+(s_\theta(X_i,Y_i)) + \frac{\lambda}{M}\sum_{j=1}^M \frac{\ell_\beta^-\left(s_\theta(X_i,Y_{ij}^-)\right)}{q_i(Y_{ij}^-)} \right]. \tag{5}$$

When the integral is finite, a full-support proposal gives an unbiased importance-sampling estimate; the proposal affects only its Monte Carlo properties. To concentrate reference samples near the conditional support, q_i is typically localized around Y_i. For instance, q_i can be a Gaussian $\mathcal{N}(Y_i, \Sigma_i)$ or a mixture of Gaussians $\sum_{k=1}^K \pi_k \mathcal{N}(Y_i, \Sigma_k)$ with varying covariance scales, so that wider components explore the tails while narrow ones concentrate samples near the observed response.

3.2 Conditional Level Tracking

The contrastive score determines the level-set geometry, but the level enclosing conditional probability $1-\alpha$ varies with x. Given the current score network, we therefore train $t_\phi(x)$ to estimate the conditional $(1-\alpha)$ quantile of $s_\theta(X,Y)$ using the pinball loss:

$$\widehat{\mathcal{Q}}_n(\phi \mid \theta) = \frac{1}{n}\sum_{i=1}^n \rho_{1-\alpha}(s_\theta(X_i,Y_i) - t_\phi(X_i)), \qquad \rho_\tau(u) = u\big(\tau - \mathbf{1}\{u<0\}\big). \tag{6}$$

Because the target quantile changes as s_θ evolves, the two networks are not fitted in separate stages. Instead, within each mini-batch, we first update the score parameters θ via (5) while holding ϕ fixed, then update the level parameters ϕ via (6) using the newly updated score, holding θ fixed. This produces a sequence of score–level pairs (s_k, t_k) whose level estimate remains aligned with the corresponding score. For any fitted pair, the centered score

$$g(x,y) = s(x,y) - t(x) \tag{7}$$

is used as the nonconformity score. Its zero sublevel set is the nominal prediction region before conformal calibration; at the population target, this zero level coincides with the conditional minimum-volume boundary derived in Section 1.

3.3 Checkpoint Selection and Conformal Calibration

We save a collection of score–level checkpoints $\{(s_k, t_k) : k \in \mathcal{K}\}$ during training. Let $\widehat{q}_{1-\alpha}(\{r_i\}_{i=1}^N)$ denote the split-conformal empirical quantile, i.e., the $\lceil (N+1)(1-\alpha) \rceil$-th smallest element of $\{r_1, \ldots, r_N, +\infty\}$. Each checkpoint k is temporarily conformalized on $\mathcal{D}_{\text{val}}$:

$$\widehat{\gamma}_k^{\text{val}} = \widehat{q}_{1-\alpha}(\{g_k(X_i,Y_i) : (X_i,Y_i) \in \mathcal{D}_{\text{val}}\}), \quad g_k = s_k - t_k, \tag{8}$$

Algorithm 1 Lebesgue-contrastive prediction (LECP)

Require: $\mathcal{D}_{\mathrm{train}}$, $\mathcal{D}_{\mathrm{val}}$, $\mathcal{D}_{\mathrm{cal}}$, miscoverage level α, reference intensity λ, temperature β, proposal family $\{q_i\}$, reference-sample size M, checkpoint set $\mathcal{K}$.

```
Initialize score parameters θ and level parameters φ.
for each training epoch do
    for each mini-batch B ⊂ D_train do
        For (X_i, Y_i) ∈ B, draw Y_ij^- ~ q_i, j = 1, ..., M.
        Update θ using (5) on B.
        Update φ using (6) on B, with θ fixed.
    end for
    if the epoch is in K then
        Save (s_θ, t_φ).
    end if
end for
for each saved checkpoint k do
    Set g_k = s_k − t_k, compute γ̂_k^val, and estimate V̂_k^val.
end for
Select k* ∈ arg min_{k∈K} V̂_k^val.
Freeze ŝ = s_{k*} and refine t_{k*} on D_train, yielding t̂.
Set ĝ = ŝ − t̂ and compute γ̂^cal on D_cal.
return Ĉ_α(x) = {y : ĝ(x, y) ≤ γ̂^cal}.
```

and the checkpoint $k^\star$ with the smallest estimated mean volume $\widehat{V}_k^{\mathrm{val}}$ of $\{y : g_k(x,y) \le \widehat{\gamma}_k^{\mathrm{val}}\}$ is selected. We then freeze $\widehat{s} = s_{k^\star}$ and refine the level network on $\mathcal{D}_{\mathrm{train}}$ with $\widehat{s}$ held fixed, reducing any lag between the selected score and its quantile estimate accumulated during alternating updates. The refined centered score $\widehat{g} = \widehat{s} - \widehat{t}$ is calibrated on the independent set $\mathcal{D}_{\mathrm{cal}}$:

$$\widehat{\gamma}^{\mathrm{cal}} = \widehat{q}_{1-\alpha}(\{\widehat{g}(X_i, Y_i) : (X_i, Y_i) \in \mathcal{D}_{\mathrm{cal}}\}), \tag{9}$$

yielding the final region $\widehat{C}_\alpha(x) = \{y \in \mathcal{Y} : \widehat{g}(x,y) \le \widehat{\gamma}^{\mathrm{cal}}\}$. The complete pipeline of LECP is summarized in Algorithm 1.

3.4 Residual Refinement Variant: Res-LECP

The fitted LECP score may retain systematic error due to finite data, limited model capacity, or imperfect optimization. We therefore introduce a stagewise refinement called Residual LECP (Res-LECP). Starting from the fitted LECP pair $(s^{[0]}, t^{[0]}) = (\widehat{s}, \widehat{t})$, stage b freezes the cumulative score $s^{[b-1]}$ and constructs the self-normalized reference

$$\pi_{b-1}(y \mid x) \propto w_{b-1}(x,y) \exp\Big\{-s^{[b-1]}(x,y)/\beta\Big\}, \tag{10}$$

where the positive weight w_{b-1} concentrates mass on responses near the current level-set boundary. A residual network δ_b is then fitted by weighted logistic contrast between the conditional data distribution and this reference. Because the same weight w_{b-1} multiplies both the data measure and the unnormalized reference, its y-dependence cancels in the density ratio; the reference normalization contributes only an x-dependent offset. The unrestricted Bayes-optimal correction therefore satisfies

$$s^{[b-1]}(x,y) + \delta_b^\star(x,y) = -\beta \log p(y \mid x) + a_b(x), \tag{11}$$

for an x-dependent offset $a_b(x)$, which is subsequently absorbed by the conditional level. The cumulative score is updated as $s^{[b]} = s^{[b-1]} + \eta\, \delta_b$ with a step size $\eta \in (0, 1]$; at unit step size, a single population residual stage removes the remaining log-density error, and the boundary weight concentrates estimation effort near the contour that determines the prediction region.

In practice, reference samples are drawn from the same proposal family as LECP and reweighted by self-normalized importance weights $\bar{W}_{ij}^{(b)} \propto w_{b-1}(X_i, Y_{ij}^-) \exp\{-s^{[b-1]}(X_i, Y_{ij}^-)/\beta\}/q_i(Y_{ij}^-)$, giving the empirical residual loss

$$\widehat{\mathcal{R}}_b(\delta) = \frac{1}{n} \sum_{i=1}^{n} \left[w_{b-1}(X_i, Y_i)\, \ell_\beta^+(\delta(X_i, Y_i)) + \sum_{j=1}^{M} \bar{W}_{ij}^{(b)}\, \ell_\beta^-(\delta(X_i, Y_{ij}^-)) \right]. \tag{12}$$

After fitting B residual corrections, Res-LECP aligns each prefix $b \in \{0, \dots, B\}$ with a conditional-quantile correction, temporarily conformalizes each on $\mathcal{D}_{\mathrm{val}}$, and selects the prefix with the smallest estimated volume for final calibration on $\mathcal{D}_{\mathrm{cal}}$ (Algorithm 2; full derivation in Appendix A).

4 Theoretical Analysis

Retain the notation of Section 1; write P_X for the marginal law of X and fix jointly measurable versions of all conditional densities, distributions, and quantiles. For fixed $\lambda, \beta > 0$, define

$$D_s(x, y) := \sigma\left(-\frac{s(x, y)}{\beta}\right), \qquad \sigma(u) = (1 + e^{-u})^{-1}. \tag{13}$$

4.1 Population Target and Finite-Sample Validity

We first characterize the population target learned by the Lebesgue-contrastive loss and its relation to the conditional minimum-volume oracle in the following Proposition 1.

Proposition 1 (Population target and conditional minimum volume)**.** *Suppose $Y \mid X = x$ admits a conditional density $p(\cdot \mid x)$ and there exists a measurable $c_\alpha : \mathcal{X} \to (0, \infty)$ such that*

$$\mathbb{P}(p(Y \mid X) \geq c_\alpha(X) \mid X = x) = 1 - \alpha. \tag{14}$$

Let $s^\star(x, y) = \beta \log\{\lambda / p(y \mid x)\}$, with $s^\star = +\infty$ where $p = 0$, and $t^\star_\alpha(x) = \beta \log\{\lambda / c_\alpha(x)\}$. Then $s^\star$ minimizes $\mathcal{R}_{\lambda,\beta}$ over extended-valued measurable scores and, when $\mathcal{R}_{\lambda,\beta}(s^\star) < \infty$, is unique up to $P_X \otimes \mu$-null sets. Moreover, $t^\star_\alpha(x)$ is a conditional $(1 - \alpha)$-quantile of $s^\star(x, Y)$, and

$$C^\star_\alpha(x) := \{y : p(y \mid x) \geq c_\alpha(x)\} = \{y : s^\star(x, y) \leq t^\star_\alpha(x)\}. \tag{15}$$

For P_X-almost every x, $C^\star_\alpha(x)$ has finite volume and is the unique minimizer of (1) up to μ-null sets.

The final split-conformal step provides finite-sample marginal validity regardless of estimation accuracy. Let $\mathcal{F}_n$ be the σ-field generated by the training and validation data and the algorithmic randomness used to construct the pre-calibration score $\widehat{g}$. Checkpoint selection, level refinement, and residual-prefix selection are included in $\mathcal{F}_n$; calibration data are not.

Theorem 1 (Finite-sample marginal coverage)**.** *Let $\widehat{g}$ be the pre-calibration score returned by Algorithm 1 or Algorithm 2. Suppose that, conditional on $\mathcal{F}_n$, the n_{cal} calibration observations and a fresh test observation $(X_{n_{\mathrm{cal}}+1}, Y_{n_{\mathrm{cal}}+1})$ are exchangeable. Then the calibrated region $\widehat{C}_\alpha$ satisfies*

$$1 - \alpha \leq \mathbb{P}\Big(Y_{n_{\mathrm{cal}}+1} \in \widehat{C}_\alpha(X_{n_{\mathrm{cal}}+1})\Big) \leq 1 - \alpha + \frac{1}{n_{\mathrm{cal}} + 1}, \tag{16}$$

where the upper bound additionally requires the $n_{\mathrm{cal}}+1$ scores $\widehat{g}(X_i, Y_i)$ to be almost surely distinct.

4.2 Asymptotic Conditional Coverage and Minimum-Volume Optimality

Retain $\mathcal{F}_n$ for the information used to fit the final pre-calibration pair $(\widehat{s}_n, \widehat{t}_n)$, and set $\mathcal{H}_n := \sigma(\mathcal{F}_n, \mathcal{D}_{\mathrm{cal}})$. Write $\widehat{g}_n = \widehat{s}_n - \widehat{t}_n$ and $\widehat{C}_{\alpha,n}(x) = \{y : \widehat{g}_n(x, y) \leq \widehat{\gamma}^{\mathrm{cal}}_n\}$. Following Lei et al. (2018); Sesia & Candès (2020), we formulate asymptotic conditional coverage on covariate sets whose probability tends to one.

Definition 1 (Asymptotic conditional coverage)**.** *The regions $\widehat{C}_{\alpha,n}$ are said to achieve asymptotic conditional coverage at level $1 - \alpha$ if there exist $\mathcal{H}_n$-measurable random sets $\Lambda_n \subseteq \mathcal{X}$ such that, for a fresh (X, Y) independent of $\mathcal{H}_n$,*

$$\mathbb{P}(X \in \Lambda_n \mid \mathcal{H}_n) = 1 - o_P(1), \quad \sup_{x \in \Lambda_n} \left|\mathbb{P}\Big(Y \in \widehat{C}_{\alpha,n}(x) \,\Big|\, X = x, \mathcal{H}_n\Big) - (1 - \alpha)\right| = o_P(1). \tag{17}$$

Assumption 1 (Data splitting and calibration)**.** *The observations are i.i.d. from P_{XY}. The calibration observations and a fresh test observation are jointly independent of $\mathcal{F}_n$, and $n_{\mathrm{cal}} \to \infty$ as $n \to \infty$. The fitted maps $(x, y) \mapsto \widehat{s}_n(x, y)$ and $x \mapsto \widehat{t}_n(x)$ are jointly measurable with the fitting randomness and are real-valued.*

Assumption 2 (Oracle regularity)**.** *The conditional density and threshold in Proposition 1 exist. Let* $g^\star_\alpha(x,y) := s^\star(x,y) - t^\star_\alpha(x)$*. For* P_X*-almost every* x*, the target score quantile is identifiable: for every* $\varepsilon > 0$*,*

$$\mathbb{P}(g^\star_\alpha(X,Y) \le -\varepsilon \mid X = x) < 1-\alpha < \mathbb{P}(g^\star_\alpha(X,Y) \le \varepsilon \mid X = x)\,. \tag{18}$$

In addition, $\mu\{y : p(y \mid x) = c_\alpha(x)\} = 0$ *for* P_X*-almost every* x*.*

Together with (14), condition (18) states that the conditional $(1-\alpha)$-quantile set of $s^\star(x,Y)$ is the singleton $\{t^\star_\alpha(x)\}$; the zero-contour condition is equivalent to $\mathbb{P}(g^\star_\alpha(X,Y) = 0 \mid X = x) = 0$, since $\mathbb{P}(p(Y \mid X) = c_\alpha(X) \mid X = x) = c_\alpha(x)\,\mu\{y : p(y \mid x) = c_\alpha(x)\}$.

For the population target, write $D^\star := D_{s^\star}$, so that $D^\star(x,y) = p(y \mid x)/\{p(y \mid x) + \lambda\}$, with $D^\star = 0$ where $p = 0$. For the fitted score, define $\widehat{D}_n(x,y) := D_{\widehat{s}_n}(x,y)$, $\widehat{F}_{n,x}(u) := \mathbb{P}(\widehat{s}_n(X,Y) \le u \mid X = x, \mathcal{F}_n)$, and $\mathcal{T}_n(x) := \Big\{u \in \mathbb{R} : \widehat{F}_{n,x}(u^-) \le 1-\alpha \le \widehat{F}_{n,x}(u)\Big\}$.

Assumption 3 (Estimation consistency)**.** *Let* $\mathrm{dist}(u,A) := \inf_{v\in A}|u-v|$ *for* $u \in \mathbb{R}$ *and* $A \subseteq \mathbb{R}$*. With expectations taken over a fresh observation independent of* $\mathcal{F}_n$*, the pre-calibration pair satisfies*

$$\varepsilon_{D,n} := \mathbb{E}\Big[(\widehat{D}_n(X,Y) - D^\star(X,Y))^2 \,\Big|\, \mathcal{F}_n\Big] = o_P(1), \tag{19}$$

$$\varepsilon_{\mu,n} := \mathbb{E}\left[\int_{\mathcal{Y}} (\widehat{D}_n(X,y) - D^\star(X,y))^2 \, d\mu(y) \,\middle|\, \mathcal{F}_n\right] = o_P(1), \tag{20}$$

$$\varepsilon_{t,n} := \mathbb{E}\big[\min\big\{1, \mathrm{dist}^2\big(\widehat{t}_n(X), \mathcal{T}_n(X)\big)\big\} \,\big|\, \mathcal{F}_n\big] = o_P(1). \tag{21}$$

Conditions (19) and (21) suffice for asymptotic conditional coverage. The Lebesgue-domain condition (20) implies (19) and is additionally needed for geometric recovery (Theorem 2(ii)); on unbounded $\mathcal{Y}$ it requires $\widehat{D}_n(x,\cdot) - D^\star(x,\cdot) \in L^2(\mu)$, automatic on bounded $\mathcal{Y}$ but demanding sufficient tail growth of $\widehat{s}_n$ on $\mathbb{R}^d$. The bounded error in (21) captures convergence to the fitted-score quantile set in P_X-probability and follows from the untruncated mean-square analogue. All these three conditions concern the final pair after checkpoint selection and level refinement.

Theorem 2 (Asymptotic conditional coverage and minimum-volume optimality)**.** *Suppose Assumptions 1 and 2 hold.*

(i) ***Asymptotic conditional coverage.*** *If (19) and (21) hold, then* $\widehat{\gamma}^{\mathrm{cal}}_n \xrightarrow{P} 0$*, and* $\widehat{C}_{\alpha,n}$ *achieves asymptotic conditional coverage at level* $1-\alpha$ *in the sense of Definition 1.*

(ii) ***Asymptotic minimum-volume optimality.*** *If, in addition, (20) holds, then, for a fresh* $X \sim P_X$ *independent of* $\mathcal{H}_n$*,*

$$\mu\Big(\widehat{C}_{\alpha,n}(X) \triangle C^\star_\alpha(X)\Big) \xrightarrow{P} 0. \tag{22}$$

The convergences are joint over fitting, calibration, and an independent test covariate. Complete proofs and sufficient conditions in terms of population excess risks are given in Appendix B.

5 Experimental Studies

Experimental setup. We study three questions: whether LECP can recover conditional regions with nonconvex or disconnected topology, whether this flexibility translates into smaller regions at comparable conditional reliability, and whether the gain persists beyond two response dimensions. The synthetic benchmarks comprise one-dimensional Comb and two-dimensional Rings, Spiral, and Lemniscate. For the real multi-output benchmarks, $d = 2$ for Bio and House (Feldman et al., 2023) and Taxi (Wang et al., 2023); $d = 4$ for Households (Camehl et al., 2025); and $d = 8$ for RF1 and RF2 (Dheur et al., 2025). We compare LECP and Res-LECP with SLS (Braun et al., 2026c), L-CP and C-PCP (Dheur et al., 2025), C-HDR (Izbicki et al., 2022), PCP (Wang et al., 2023) and CP2-PCP (Plassier et al., 2025b). Every method uses the same random $60\%/10\%/10\%/20\%$ train/validation/calibration/test split and a final split-conformal calibration. We repeat the comparison over ten seeds at $\alpha \in \{0.05, 0.10, 0.20\}$ and focus below on $\alpha = 0.10$; data, implementation, and additional results are reported in Appendix C.

Evaluation metrics. Let $\pi_{\widehat{C}}(x) = \mathbb{P}(Y \in \widehat{C}(X) \mid X = x)$, and let n_{test} be the test-sample size. Our efficiency measure and synthetic conditional-reliability metric are, respectively,

$$\text{Volume} = \left\{ \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mu\big(\widehat{C}(X_i)\big) \right\}^{1/d}, \quad \text{CC-MAE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left| \pi_{\widehat{C}}(X_i) - (1-\alpha) \right|. \tag{23}$$

Unless otherwise noted, all volumes reported below are dimension-scaled (the d-th root of the mean volume), facilitating comparison across response dimensions. The synthetic generator permits direct Monte Carlo estimation of CC-MAE. On real data, we use L_1 excess risk of target coverage (ERT) and worst-slab coverage (WSC) (Braun et al., 2026b) with minimum slab mass 0.10; their definitions and estimators are in Appendix C. Smaller volume, CC-MAE, and ERT are preferred, whereas larger WSC is preferred. Note that volume and conditional diagnostics should be interpreted jointly.

5.1 Conditional topology and efficiency

Table 1 shows a consistent efficiency gain without a corresponding loss in conditional reliability. Res-LECP has the smallest mean volume on all four problems, reducing volume relative to SLS (the strongest baseline by this criterion) by 34% on Comb, 19% on Rings, 6% on Spiral, and 23% on Lemniscate. LECP is nearly indistinguishable from its residual refinement. Both variants attain the smallest or tied smallest CC-MAE throughout.

Figure 1 illustrates the source of this improvement. The learned sublevel sets adapt their topology to the covariate: LECP transitions from one to several intervals on Comb, retains both annular bands on Rings, and resolves the narrow Spiral and Lemniscate supports. These examples demonstrate the flexible geometric behavior of density-level learning.

Table 1: Synthetic results at $\alpha = 0.10$. Entries are mean (standard deviation) over ten splits. Boldface marks the smallest mean in each row. All methods are split-conformalized.

Dataset	Metric	LECP	Res-LECP	SLS	L-CP	C-HDR	C-PCP	CP2-PCP	PCP
Comb	Volume (↓)	$1.23_{(0.03)}$	$\mathbf{1.22}_{(0.04)}$	$1.84_{(0.05)}$	$3.86_{(0.04)}$	$3.99_{(0.08)}$	$5.19_{(0.12)}$	$5.17_{(0.11)}$	$4.79_{(0.15)}$
	CC-MAE (↓)	$\mathbf{0.02}_{(0.01)}$	$\mathbf{0.02}_{(0.01)}$	$0.03_{(0.00)}$	$0.07_{(0.01)}$	$0.08_{(0.01)}$	$0.05_{(0.00)}$	$0.05_{(0.00)}$	$0.09_{(0.01)}$
Rings	Volume (↓)	$\mathbf{1.13}_{(0.02)}$	$\mathbf{1.13}_{(0.02)}$	$1.39_{(0.01)}$	$1.44_{(0.01)}$	$1.42_{(0.03)}$	$1.62_{(0.03)}$	$1.60_{(0.02)}$	$1.58_{(0.02)}$
	CC-MAE (↓)	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$0.03_{(0.00)}$	$0.03_{(0.00)}$	$0.03_{(0.00)}$	$0.04_{(0.00)}$
Spiral	Volume (↓)	$\mathbf{2.63}_{(0.05)}$	$\mathbf{2.63}_{(0.05)}$	$2.80_{(0.07)}$	$3.50_{(0.10)}$	$3.31_{(0.10)}$	$3.87_{(0.06)}$	$3.85_{(0.06)}$	$3.81_{(0.06)}$
	CC-MAE (↓)	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$0.03_{(0.00)}$	$0.03_{(0.00)}$	$0.03_{(0.00)}$	$0.05_{(0.00)}$
Lemnis-cate	Volume (↓)	$0.37_{(0.01)}$	$\mathbf{0.36}_{(0.01)}$	$0.47_{(0.07)}$	$0.52_{(0.02)}$	$0.48_{(0.02)}$	$0.57_{(0.02)}$	$0.57_{(0.02)}$	$0.56_{(0.01)}$
	CC-MAE (↓)	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$0.02_{(0.01)}$	$\mathbf{0.01}_{(0.00)}$	$0.04_{(0.01)}$	$0.03_{(0.00)}$	$0.03_{(0.00)}$	$0.03_{(0.00)}$

5.2 Real multi-output regression

LECP or Res-LECP has the smallest mean volume on every real benchmark (Table 2). Against SLS, Res-LECP reduces volume by about 15% on House, 7% on Households, 4% on Bio, 9% on Taxi, and more than 25% on RF1 and RF2. The conditional diagnostics present a more nuanced picture: density- and sample-based baselines occasionally attain better ERT or WSC, but typically at the cost of appreciably larger regions. LECP and Res-LECP occupy the small-volume end of the empirical efficiency–reliability frontier, with ERT and WSC of the same order as the strongest competitors. Residual refinement yields modest but consistent volume improvements, indicating that the base contrastive score already captures most of the useful geometry.

The Taxi slices in Figure 2 make this trade-off visible. The baselines each impose topological constraints: CP2-PCP is restricted to unions of fixed-radius balls, flow-based C-HDR typically yields connected regions that struggle to separate distinct modes, and SLS requires the maximum number of components (flows) to be specified in advance. LECP, by contrast, constructs prediction regions whose topology is entirely data-driven: it can discover an arbitrary number of disconnected components without pre-specifying a mode count, producing compact regions that faithfully track the multimodal structure of the conditional distribution as it shifts across covariate slices.

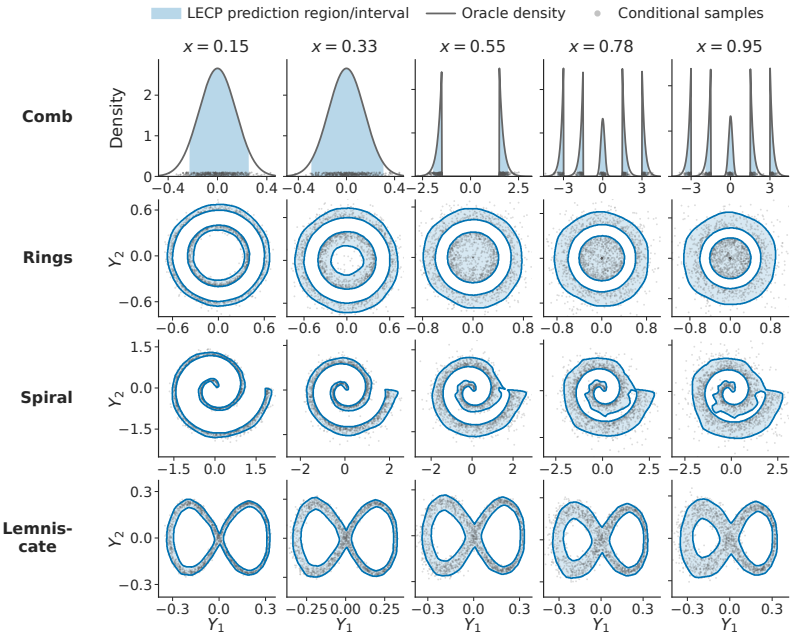


Figure 1: LECP prediction regions at five covariate values for the four synthetic problems ($\alpha = 0.10$, one representative split). The Comb row overlays the prediction interval on the oracle conditional density; the remaining rows show two-dimensional regions and conditional draws. LECP follows changes in the number of connected components, preserves holes, and tracks curved supports.

Table 2: Real-data results at $\alpha = 0.10$. Entries are mean (standard deviation) over ten splits. Boldface marks the best mean in each row. All methods are split-conformalized.

Dataset	Metric	LECP	Res-LECP	SLS	L-CP	C-HDR	C-PCP	CP2-PCP	PCP
Bio	Volume (↓)	**0.85**$_{(0.02)}$	**0.85**$_{(0.03)}$	0.89$_{(0.02)}$	1.05$_{(0.05)}$	0.94$_{(0.03)}$	1.06$_{(0.03)}$	1.03$_{(0.03)}$	0.99$_{(0.02)}$
	ERT (↓)	0.03$_{(0.00)}$	**0.02**$_{(0.00)}$	**0.02**$_{(0.00)}$	**0.02**$_{(0.00)}$	**0.02**$_{(0.00)}$	**0.02**$_{(0.00)}$	**0.02**$_{(0.00)}$	0.04$_{(0.00)}$
	WSC (↑)	0.83$_{(0.02)}$	0.83$_{(0.02)}$	0.83$_{(0.01)}$	0.84$_{(0.01)}$	**0.85**$_{(0.01)}$	**0.85**$_{(0.01)}$	0.84$_{(0.01)}$	0.79$_{(0.01)}$
House	Volume (↓)	**0.86**$_{(0.03)}$	**0.86**$_{(0.03)}$	1.01$_{(0.02)}$	1.47$_{(0.07)}$	1.50$_{(0.09)}$	1.66$_{(0.08)}$	1.62$_{(0.07)}$	1.52$_{(0.06)}$
	ERT (↓)	0.04$_{(0.01)}$	0.04$_{(0.01)}$	0.04$_{(0.01)}$	**0.03**$_{(0.01)}$	**0.03**$_{(0.01)}$	**0.03**$_{(0.01)}$	**0.03**$_{(0.01)}$	0.04$_{(0.01)}$
	WSC (↑)	0.75$_{(0.02)}$	0.75$_{(0.02)}$	0.76$_{(0.03)}$	0.75$_{(0.02)}$	**0.77**$_{(0.03)}$	**0.77**$_{(0.02)}$	0.76$_{(0.02)}$	**0.77**$_{(0.02)}$
Taxi	Volume (↓)	2.49$_{(0.03)}$	**2.48**$_{(0.03)}$	2.72$_{(0.04)}$	2.92$_{(0.05)}$	2.90$_{(0.04)}$	3.39$_{(0.05)}$	3.35$_{(0.04)}$	3.32$_{(0.04)}$
	ERT (↓)	0.01$_{(0.00)}$	0.01$_{(0.00)}$	0.01$_{(0.00)}$	0.01$_{(0.00)}$	0.01$_{(0.00)}$	0.01$_{(0.00)}$	**0.00**$_{(0.00)}$	0.01$_{(0.00)}$
	WSC (↑)	0.85$_{(0.01)}$	0.85$_{(0.01)}$	0.86$_{(0.01)}$	0.86$_{(0.01)}$	0.86$_{(0.01)}$	**0.87**$_{(0.01)}$	0.86$_{(0.01)}$	0.85$_{(0.01)}$
House-holds	Volume (↓)	2.56$_{(0.07)}$	**2.55**$_{(0.08)}$	2.73$_{(0.06)}$	2.67$_{(0.08)}$	2.66$_{(0.07)}$	2.94$_{(0.07)}$	2.92$_{(0.05)}$	2.82$_{(0.05)}$
	ERT (↓)	**0.01**$_{(0.02)}$	**0.01**$_{(0.01)}$	**0.01**$_{(0.01)}$	0.02$_{(0.01)}$	**0.01**$_{(0.01)}$	**0.01**$_{(0.01)}$	**0.01**$_{(0.01)}$	**0.01**$_{(0.01)}$
	WSC (↑)	0.76$_{(0.03)}$	0.76$_{(0.03)}$	0.76$_{(0.01)}$	0.76$_{(0.02)}$	**0.77**$_{(0.02)}$	**0.77**$_{(0.01)}$	0.76$_{(0.01)}$	**0.77**$_{(0.01)}$
RF1	Volume (↓)	0.34$_{(0.03)}$	**0.33**$_{(0.03)}$	0.45$_{(0.03)}$	0.34$_{(0.02)}$	0.34$_{(0.02)}$	0.45$_{(0.03)}$	0.45$_{(0.03)}$	0.35$_{(0.03)}$
	ERT (↓)	0.05$_{(0.01)}$	0.05$_{(0.01)}$	0.05$_{(0.01)}$	0.05$_{(0.01)}$	0.05$_{(0.01)}$	**0.03**$_{(0.01)}$	**0.03**$_{(0.01)}$	0.08$_{(0.01)}$
	WSC (↑)	0.75$_{(0.05)}$	0.76$_{(0.03)}$	**0.77**$_{(0.02)}$	0.74$_{(0.02)}$	0.75$_{(0.02)}$	0.76$_{(0.03)}$	0.75$_{(0.03)}$	0.52$_{(0.06)}$
RF2	Volume (↓)	0.32$_{(0.00)}$	**0.31**$_{(0.01)}$	0.44$_{(0.05)}$	0.32$_{(0.01)}$	0.33$_{(0.02)}$	0.43$_{(0.03)}$	0.43$_{(0.03)}$	0.33$_{(0.02)}$
	ERT (↓)	0.05$_{(0.01)}$	0.05$_{(0.01)}$	0.06$_{(0.01)}$	0.05$_{(0.01)}$	0.05$_{(0.01)}$	**0.04**$_{(0.01)}$	**0.04**$_{(0.01)}$	0.09$_{(0.01)}$
	WSC (↑)	0.76$_{(0.05)}$	0.76$_{(0.03)}$	0.76$_{(0.03)}$	0.74$_{(0.02)}$	0.76$_{(0.02)}$	**0.77**$_{(0.03)}$	0.75$_{(0.03)}$	0.52$_{(0.06)}$

Supplementary analyses. Taken together, LECP consistently produces the smallest or near-smallest regions, while baselines that occasionally surpass it in ERT or WSC do so at appreciably larger volume. This ranking persists at $\alpha \in \{0.05, 0.20\}$ (Appendix D.1) and is robust to moderate perturbations of all main hyperparameters, with an overly narrow proposal scale as the sole clear failure mode (Appendix D.2). Mechanism ablations attribute the efficiency gains primarily to the importance-sampling correction that enforces a Lebesgue effective reference, with proposal local-

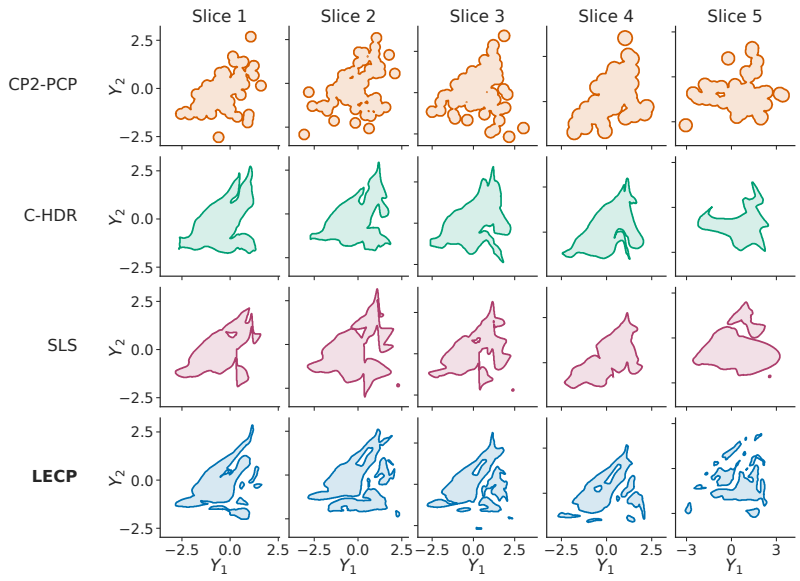


Figure 2: Prediction regions on Taxi at five test covariates ($\alpha = 0.10$, one representative split). LECP retains the nonconvex structure of the conditional response while avoiding the diffuse components and excess area visible in several competing regions.

ity contributing additionally in higher dimensions (Appendix D.3). The narrow gap between LECP and Res-LECP on the present benchmarks reflects the adequacy of the base contrastive score. A controlled stress test amplifying covariate-dependent response scale ($Y = X^{\gamma}Z$, $\gamma \in \{0, \ldots, 4\}$) exposes the regime where residual refinement matters: as scale heterogeneity grows, Res-LECP improves the IoU between the calibrated region and the oracle HDR by up to $+0.28$ over the base fit while maintaining comparable conditional reliability (Appendix D.5). Although LECP's training cost exceeds that of a single flow fit, its inference requires only two scalar forward passes and is an order of magnitude faster than the cheapest baseline (Appendix D.4).

6 Conclusions

We introduced LECP, a contrastive framework for multivariate conformal regression. By learning density-level geometry through a lightweight contrastive objective paired with conditional-quantile tracking, LECP consistently yields smaller prediction regions than both flow-based and direct volume-optimization baselines across all synthetic and real benchmarks, while maintaining competitive conditional reliability. The framework requires neither density normalization nor invertible parameterizations; split conformal calibration provides finite-sample marginal coverage by construction, and under estimation consistency of the score and level networks, the calibrated regions converge to the conditional minimum-volume oracle. Residual refinement (Res-LECP) offers additional gains, particularly when the base score retains systematic error near the target contour.

Limitations and future work. While localized mixture proposals perform well in moderate dimensions, importance-weight variance grows inevitably with response dimension; designing adaptive or amortized sampling strategies is critical for high-dimensional applicability. The reference intensity λ currently requires manual tuning, motivating data-driven selection. Our asymptotic guarantees rest on estimation consistency conditions that cannot be directly verified, and establishing finite-sample convergence rates for the coupled score and level learners remains open. Although LECP achieves inference times an order of magnitude below flow-based alternatives, its training cost is higher, favoring settings where low-latency prediction justifies the upfront expense. Finally, the Lebesgue reference restricts the framework to continuous responses in $\mathbb{R}^d$; extending the contrastive construction to mixed, discrete, or manifold-valued output spaces is a natural next step.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation, debugging, and organization of portions of the experimental code, and with the development, checking, and refinement of parts of the mathematical proofs. Generative AI was also used to edit the manuscript for language, grammar, clarity, and presentation.

All AI-assisted outputs were critically reviewed and verified by the authors. In particular, AI-assisted experimental code was manually inspected, executed, and tested for consistency with the intended algorithms and experimental protocols, and AI-assisted mathematical arguments were independently checked by the authors for logical correctness and consistency with the stated assumptions and claims. The final mathematical statements, proofs, experimental methodology, results, and scientific conclusions were determined and validated by the authors.

We take responsibility for the final content of this work, including all text, mathematical claims and proofs, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops a general-purpose methodology for multivariate conformal regression and does not involve newly recruited human participants, user studies, interventions, or the collection of new personal data. Our experiments use synthetic data and previously published benchmark datasets, with their sources and preprocessing procedures documented in the paper and supplementary material. A relevant limitation for downstream use is that the finite-sample guarantee provided by split conformal prediction is marginal rather than conditional; it should therefore not be interpreted as guaranteeing subgroup- or individual-level reliability. Applications in consequential settings should additionally consider distribution shift, data representativeness, fairness, privacy, safety, and domain-specific validation.

REPRODUCIBILITY STATEMENT

The methodological and theoretical components of this work are documented in detail. The LECP methodology is described in Section 3 and Algorithm 1, while the complete Res-LECP procedure is provided in Appendix A and Algorithm 2. The theoretical assumptions and main guarantees are stated in Section 4, with complete proofs and supporting results given in Appendix B.

Experimental details are provided in Appendix C, including the datasets, synthetic data-generating mechanisms, preprocessing procedures, data splits, model architectures, training protocols, hyperparameters, and evaluation metrics. Appendix D reports results across additional miscoverage levels, sensitivity analyses, mechanism ablations, computational-cost comparisons, and a controlled conditional-scale stress test designed to assess when residual refinement (Res-LECP) improves over the base LECP fit. We additionally provide the experimental code in the supplementary material, including implementations for data generation, LECP and Res-LECP, conformal calibration, baseline evaluation, and computation of the reported metrics.

REFERENCES

Sacha Braun, Liviu Aolaritei, Michael I. Jordan, and Francis Bach. Minimum volume conformal sets for multivariate regression. *Journal of the American Statistical Association*, 0(ja):1–27, 2026a. doi: 10.1080/01621459.2026.2692078.

Sacha Braun, David Holzmüller, Michael I. Jordan, and Francis Bach. Conditional coverage diagnostics for conformal prediction. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=vaApZm6MKM.

Sacha Braun, Michael I. Jordan, and Francis Bach. Super-level-set regression: Conditional quantiles via volume minimization, 2026c. URL https://arxiv.org/abs/2605.06210.

Annika Camehl, Dennis Fok, and Kathrin Gruber. On superlevel sets of conditional densities and multivariate quantile regression. *Journal of Econometrics*, 249:105807, 2025. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2024.105807.

Ciwan Ceylan and Michael U. Gutmann. Conditional noise-contrastive estimation of unnormalised models. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 726–734. PMLR, 10–15 Jul 2018.

Dehao Dai, Jianqing Fan, Yihong Gu, and Debarghya Mukherjee. Cindes: Classification induced neural density estimator and simulator, 2025. URL https://arxiv.org/abs/2510.00367.

Victor Dheur and Souhaib Ben Taieb. Multivariate latent recalibration for conditional normalizing flows. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 4946–4988. Curran Associates, Inc., 2025.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *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.

Eshant English and Christoph Lippert. JAPAN: Joint adaptive prediction areas with normalising flow. In *The Fourteenth International Conference on Learning Representations*, volume 2026, pp. 131010–131047, 2026.

Zhenhan Fang, Aixin Tan, and Jian Huang. CONTRA: Conformal prediction region via normalizing flow transformation. In *International Conference on Learning Representations*, volume 2025, pp. 91833–91850, 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.

Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18495–18527. PMLR, 13–19 Jul 2025.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13 (11):307–361, 2012.

Harlfoxem. House sales in king county, usa, 2016. URL https://www.kaggle.com/datasets/harlfoxem/housesalesprediction. Accessed: 2026-03-01.

Rob J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50 (2):120–126, 1996. ISSN 00031305.

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.

Michal Klein, Louis Bethune, Eugene Ndiaye, and Marco Cuturi. Multivariate conformal prediction using optimal transport, 2025. URL https://arxiv.org/abs/2502.03609.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013. doi: 10.1080/01621459.2012.751873.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Rui Luo and Zhixin Zhou. Density-sorted prediction set: Efficient conformal prediction for multi-target regression. *Pattern Recognition*, 172:112513, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.112513.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, 120:108101, 2021. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.108101.

Eugene Ndiaye. Beyond uncertainty sets: Leveraging optimal transport to extend conformal predictive distribution to multivariate settings, 2025. URL https://arxiv.org/abs/2511.15146.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning*, ECML'02, pp. 345–356, Berlin, Heidelberg, 2002. Springer-Verlag. ISBN 3540440364. doi: 10.1007/3-540-36755-1_29.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49459–49492. PMLR, 13–19 Jul 2025a.

Vincent Plassier, Alexander Fishkov, Mohsen Guizani, Maxim Panov, and Eric Moulines. Probabilistic conformal prediction with approximate conditional validity. In *International Conference on Learning Representations*, volume 2025, pp. 62896–62928, 2025b.

Wolfgang Polonik. Measuring Mass Concentrations and Estimating Density Contour Clusters-An Excess Mass Approach. *The Annals of Statistics*, 23(3):855 – 881, 1995. doi: 10.1214/aos/1176324626.

Wolfgang Polonik and Qiwei Yao. Conditional minimum volume predictive regions for stochastic processes. *Journal of the American Statistical Association*, 95(450):509–519, 2000. ISSN 01621459, 1537274X.

Prashant Rana. Physicochemical Properties of Protein Tertiary Structure. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5QW3H.

Clayton D. Scott and Robert D. Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7(24):665–704, 2006.

Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. *Stat*, 9(1):e261, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

Xiaoyi Su, Zhixin Zhou, and Rui Luo. Co-optimization for adaptive conformal prediction, 2026. URL https://arxiv.org/abs/2603.01719.

Masashi Sugiyama, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. Conditional density estimation via least-squares density ratio estimation. In Yee Whye Teh and Mike Titterington (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 781–788, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. Optimal transport-based conformal prediction. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59509–59527. PMLR, 13–19 Jul 2025.

Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12(71):2411–2414, 2011.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David Blei. Probabilistic conformal prediction using conditional random samples. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8814–8836. PMLR, 25–27 Apr 2023.

A RESIDUAL LECP

This appendix specifies the residual objective and training procedure used by Res-LECP. Let $(s^{[0]}, t^{[0]}) = (\widehat{s}, \widehat{t})$ denote the fitted LECP score and level. During residual fitting, both components are frozen, as are all previously fitted residual networks.

Cumulative score and boundary weight. At stage b, the frozen cumulative score is

$$s^{[b-1]}(x, y) = s^{[0]}(x, y) + \eta \sum_{\ell=1}^{b-1} \delta_\ell(x, y), \tag{24}$$

where $\eta \in (0, 1]$ is the residual step size. When boundary focusing is enabled, we set

$$w_{b-1}(x, y) = w_{\min} + (1 - w_{\min}) \exp\left[-\frac{|s^{[b-1]}(x, y) - t^{[0]}(x)|}{h\beta}\right], \tag{25}$$

with bandwidth $h > 0$ and floor $w_{\min} \in (0, 1]$. This weight concentrates estimation effort near the frozen base LECP level under the current cumulative score.

Reference distribution and residual contrast. At every residual stage, we require the frozen reference to be normalizable:

$$0 < Z_{b-1}(x) := \int_{\mathcal{Y}} w_{b-1}(x, u) e^{-s^{[b-1]}(x,u)/\beta} \, d\mu(u) < \infty \tag{26}$$

for P_X-almost every x. The stage-b reference density is then

$$\pi_{b-1}(y \mid x) = \frac{w_{b-1}(x, y) e^{-s^{[b-1]}(x,y)/\beta}}{Z_{b-1}(x)}. \tag{27}$$

The residual network δ_b minimizes

$$\mathcal{R}_b(\delta) = \mathbb{E}\Big[w_{b-1}(X, Y)\ell_\beta^+(\delta(X, Y))\Big] + \mathbb{E}_X \mathbb{E}_{Y^- \sim \pi_{b-1}(\cdot|X)}\Big[\ell_\beta^-(\delta(X, Y^-))\Big]. \tag{28}$$

Population target. Because the same positive weight $w_{b-1}(x, y)$ appears in both the data measure and the unnormalized reference, it cancels from their pointwise density ratio. The unrestricted minimizer therefore satisfies

$$\delta_b^\star(x, y) = -s^{[b-1]}(x, y) - \beta \log p(y \mid x) - \beta \log Z_{b-1}(x). \tag{29}$$

In particular, with unit step size ($\eta = 1$), we have $s^{[b-1]} + \delta_b^\star = -\beta \log p + a_b(x)$: the residual corrects the conditional log-density representation up to an x-dependent offset that is subsequently absorbed by the conditional level estimation.

Monte Carlo objective. The reference expectation in (28) is approximated using the same proposal family as LECP. For $Y_{ij}^- \sim q_i$, define

$$W_{ij}^{(b)} = \frac{w_{b-1}(X_i, Y_{ij}^-)\exp\{-s^{[b-1]}(X_i, Y_{ij}^-)/\beta\}}{q_i(Y_{ij}^-)}, \qquad \bar{W}_{ij}^{(b)} = \frac{W_{ij}^{(b)}}{\sum_{m=1}^{M} W_{im}^{(b)}}. \tag{30}$$

Since the Res-LECP reference $\pi_{b-1}(\cdot \mid x)$ is a normalized density, there is no separate intensity multiplier λ, unlike the base LECP loss (5) which contrasts against the unnormalized measure $\lambda\mu$. The self-normalized loss takes the form

$$\widehat{\mathcal{R}}_b(\delta) = \frac{1}{n}\sum_{i=1}^{n}\left[w_{b-1}(X_i, Y_i)\ell_\beta^+(\delta(X_i, Y_i)) + \sum_{j=1}^{M}\bar{W}_{ij}^{(b)}\ell_\beta^-(\delta(X_i, Y_{ij}^-))\right]. \tag{31}$$

The self-normalized reference term is generally biased for finite M; Proposition 6 establishes its almost-sure convergence as $M \to \infty$.

Residual checkpoint selection. Each residual network is initialized at zero and fitted while all earlier components remain frozen. A checkpoint is selected by validation residual loss, with the zero correction (no residual update) retained as a candidate.

Level alignment and prefix selection. No level network is trained between residual stages. After fitting the residuals, prefix b has cumulative score $s^{[b]}$ given by (24) with $b-1$ replaced by b. Fit a conditional pinball correction $u_b(x)$ to the $(1-\alpha)$-quantile of $s^{[b]}(X,Y) - t^{[0]}(X)$, and define

$$t^{[b]}(x) = t^{[0]}(x) + u_b(x), \qquad g^{[b]}(x,y) = s^{[b]}(x,y) - t^{[b]}(x), \qquad b = 1, \ldots, B. \tag{32}$$

The unrefined candidate is $g^{[0]} = s^{[0]} - t^{[0]}$.

Each $g^{[b]}$, $b = 0, \ldots, B$, is temporarily conformalized on $\mathcal{D}_{\mathrm{val}}$ and evaluated by the same validation-volume estimator, using common Monte Carlo samples across prefixes. Res-LECP selects the prefix with the smallest validation volume. Only after this choice is fixed do we compute the split-conformal quantile on $\mathcal{D}_{\mathrm{cal}}$ and return the corresponding calibrated region.

Algorithm 2 Residual LECP (Res-LECP)

Require: Fitted LECP pair $(s^{[0]}, t^{[0]})$, training set $\mathcal{D}_{\mathrm{train}}$, validation set $\mathcal{D}_{\mathrm{val}}$, calibration set $\mathcal{D}_{\mathrm{cal}}$, number of residual stages B.

1: **for** $b = 1, \ldots, B$ **do**
2: Form $s^{[b-1]}$ as in (24) and freeze all previous components.
3: Fit δ_b using (31); select its checkpoint by validation residual loss.
4: **end for**
5: Set $g^{[0]} = s^{[0]} - t^{[0]}$.
6: **for** $b = 1, \ldots, B$ **do**
7: Fit a conditional-quantile correction u_b on $\mathcal{D}_{\mathrm{train}}$ and form $t^{[b]}$ and $g^{[b]}$ as in (32).
8: Temporarily conformalize $g^{[b]}$ on $\mathcal{D}_{\mathrm{val}}$ and estimate its validation volume.
9: **end for**
10: Include $g^{[0]}$ in the same validation-volume comparison and select the best prefix $b^\star$.
11: Set $\widehat{g} \leftarrow g^{[b^\star]}$, and calibrate $\widehat{g}$ on $\mathcal{D}_{\mathrm{cal}}$, yielding $\widehat{\gamma}^{\mathrm{cal}}$.
12: **return** $\widehat{C}_\alpha(x) = \{y : \widehat{g}(x,y) \le \widehat{\gamma}^{\mathrm{cal}}\}$.

B Theoretical Proofs

B.1 Notation and preliminary identities

Basic conventions. Set $\psi(u) := \sigma(-u/\beta) = (1+e^{u/\beta})^{-1}$, so $D_s = \psi \circ s$. We extend ψ continuously by $\psi(+\infty) = 0$ and $\psi(-\infty) = 1$; on $(0,1)$, $\psi^{-1}(v) = \beta\log\{(1-v)/v\}$ and $\sup_u |\psi'(u)| = 1/(4\beta)$. Pointwise risks use $0 \cdot \infty = 0$. The σ-fields $\mathcal{F}_n$ and $\mathcal{H}_n$ are as in Section 4; fresh observations are always independent of the relevant fitting σ-field. For a random measurable set $\Lambda \subseteq \mathcal{X}$, write $P_X(\Lambda) := \int \mathbf{1}\{x \in \Lambda\}\, dP_X(x)$.

Quantile endpoints. If F is a CDF, define its lower and upper $(1-\alpha)$-quantiles by

$$\underline{t}(F) := \inf\{u : F(u) \geq 1-\alpha\}, \qquad \bar{t}(F) := \inf\{u : F(u) > 1-\alpha\}.$$

Since $1-\alpha \in (0,1)$, both are finite and satisfy

$$\{u : F(u^-) \leq 1-\alpha \leq F(u)\} = [\underline{t}(F), \bar{t}(F)].$$

For a jointly measurable family of CDFs, these endpoints are measurable since the infima may be taken over rationals. Thus, the fitted-score quantile sets $\mathcal{T}_n(x)$ are measurable nonempty compact intervals, and distance to them is measurable and attained. Write $\underline{t}_n(x) := \underline{t}(\widehat{F}_{n,x}) \in \mathcal{T}_n(x)$.

Oracle gap and good covariate set. Write $F^\star_x(u) := \mathbb{P}(s^\star(X,Y) \leq u \mid X = x)$. Under Assumption 2, define the oracle CDF gap, conditional score-error probability, and good covariate set by

$$\kappa_\varepsilon(x) := \min\left\{(1-\alpha) - F^\star_x(t^\star_\alpha(x) - \varepsilon),\ F^\star_x(t^\star_\alpha(x) + \varepsilon) - (1-\alpha)\right\}, \tag{33}$$

$$r_{n,\varepsilon}(x) := \mathbb{P}\left(|\widehat{s}_n(X,Y) - s^\star(X,Y)| \geq \varepsilon \mid X = x, \mathcal{F}_n\right), \tag{34}$$

$$\mathcal{X}_{n,\varepsilon} := \{x : \kappa_\varepsilon(x) > 0,\ r_{n,\varepsilon}(x) \leq \kappa_\varepsilon(x)/2\}.$$

All three are measurable, and $\kappa_\varepsilon(x) > 0$ for P_X-a.e. x by (18).

Discriminator-scale comparison. Set comparisons are simpler on the discriminator scale. Define

$$\vartheta^\star_\alpha(x) := \psi(t^\star_\alpha(x)) = \frac{c_\alpha(x)}{c_\alpha(x) + \lambda},$$
$$\widehat{\vartheta}_n(x) := \psi(\widehat{t}_n(x) + \widehat{\gamma}^{\mathrm{cal}}_n).$$

Monotonicity of ψ gives

$$\widehat{C}_{\alpha,n}(x) = \{y : \widehat{D}_n(x,y) \geq \widehat{\vartheta}_n(x)\}, \qquad C^\star_\alpha(x) = \{y : D^\star(x,y) \geq \vartheta^\star_\alpha(x)\}. \tag{35}$$

The following deterministic comparison is used in the set-consistency proofs:

$$\left.\begin{aligned} |\widehat{D}_n(x,y) - D^\star(x,y)| &\leq \delta, \\ |\widehat{\vartheta}_n(x) - \vartheta^\star_\alpha(x)| &\leq \delta, \\ |D^\star(x,y) - \vartheta^\star_\alpha(x)| &> 2\delta \end{aligned}\right\} \Longrightarrow \mathbf{1}\{y \in \widehat{C}_{\alpha,n}(x)\} = \mathbf{1}\{y \in C^\star_\alpha(x)\}.$$

Indeed, if $D^\star(x,y) - \vartheta^\star_\alpha(x) > 2\delta$, then

$$\widehat{D}_n(x,y) \geq D^\star(x,y) - \delta > \vartheta^\star_\alpha(x) + \delta \geq \widehat{\vartheta}_n(x),$$

so both indicators equal one; if $D^\star(x,y) - \vartheta^\star_\alpha(x) < -2\delta$, the reversed inequalities show that both indicators equal zero.

B.2 Proof of Proposition 1

Proof. Fix x. Since $\ell^+_\beta(s) = -\beta \log D_s$ and $\ell^-_\beta(s) = -\beta \log(1 - D_s)$, the pointwise contrastive integrand, divided by β, is

$$-p(y \mid x) \log D - \lambda \log(1 - D) \quad \text{for} \quad D \in [0,1].$$

If $p(y \mid x) > 0$, this function is strictly convex on $(0,1)$ and has the unique minimizer $D^\star(x,y) = p(y \mid x)/\{p(y \mid x) + \lambda\}$. If $p(y \mid x) = 0$, its minimum over $[0,1]$ is attained uniquely at $D = 0$. Inverting $D_s = \psi(s)$ yields the optimal score

$$s^\star(x,y) = \beta \log \frac{\lambda}{p(y \mid x)} = \beta(\log \lambda - \log p(y \mid x)),$$

when $p(y \mid x) > 0$, and $s^\star(x,y) = +\infty$ when $p(y \mid x) = 0$. Pointwise comparison establishes global optimality even if the minimum risk is infinite. When $\mathcal{R}_{\lambda,\beta}(s^\star) < \infty$, any other minimizer has zero integrated pointwise excess, and strict pointwise optimality implies equality with $s^\star$ $P_X \otimes \mu$-almost everywhere.

By definition of $t^\star_\alpha(x)$, we have $s^\star(x,Y) \le t^\star_\alpha(x)$ if and only if $p(Y \mid x) \ge c_\alpha(x)$ under $Y \mid X = x$. Hence, we have

$$\mathbb{P}(s^\star(X,Y) \le t^\star_\alpha(x) \mid X = x) = 1 - \alpha,$$

and $t^\star_\alpha(x)$ is a conditional $(1-\alpha)$-quantile of $s^\star(x,Y)$: its left-limit CDF is no larger than $1-\alpha$. The same comparison gives (15).

It remains to establish minimum-volume optimality and uniqueness. Let $c = c_\alpha(x)$ and $C^\star = C^\star_\alpha(x)$. Since $p(y \mid x) \ge c$ for all $y \in C^\star$,

$$c\,\mu(C^\star) \le \int_{C^\star} p(y \mid x)\, d\mu(y) = 1 - \alpha,$$

which implies $\mu(C^\star) \le (1-\alpha)/c < \infty$. Let $C \subseteq \mathcal{Y}$ be measurable and satisfy $\mathbb{P}(Y \in C \mid X = x) \ge 1 - \alpha$. If $\mu(C) = \infty$, the volume comparison is immediate. Otherwise,

$$\begin{aligned} &c\{\mu(C) - \mu(C^\star)\} - \{\mathbb{P}(Y \in C \mid X = x) - (1-\alpha)\} \\ &\quad = \int_C \{c - p(y \mid x)\}\, d\mu(y) - \int_{C^\star} \{c - p(y \mid x)\}\, d\mu(y) \\ &\quad = \int_{C^\star \setminus C} \{p(y \mid x) - c\}\, d\mu(y) + \int_{C \setminus C^\star} \{c - p(y \mid x)\}\, d\mu(y) \ge 0. \end{aligned} \tag{36}$$

Feasibility implies $\mu(C) \ge \mu(C^\star)$.

If the volumes agree, the nonnegative right side of (36) must vanish. Since $p < c$ on $C \setminus C^\star$, this gives $\mu(C \setminus C^\star) = 0$; equality of the finite volumes then gives $\mu(C^\star \setminus C) = 0$. Hence $\mu(C \triangle C^\star) = 0$. □

B.3 Proof of Theorem 1

Proof. Conditional on $\mathcal{F}_n$, the function $\widehat{g}$ is fixed and the scores $R_i := \widehat{g}(X_i, Y_i)$, $i = 1, \dots, n_{\mathrm{cal}}+1$, are exchangeable. Adjoin independent uniform variables U_i and rank the pairs (R_i, U_i) lexicographically. Their conditional exchangeability and almost-sure distinctness make the test pair's rank uniform on $\{1, \dots, n_{\mathrm{cal}} + 1\}$.

Let $k_{\mathrm{cal}} := \lceil (n_{\mathrm{cal}} + 1)(1-\alpha) \rceil$. If $k_{\mathrm{cal}} \le n_{\mathrm{cal}}$, test-pair rank at most k_{cal} implies that the test score is at most the k_{cal}-th calibration score. If $k_{\mathrm{cal}} = n_{\mathrm{cal}} + 1$, the threshold is $+\infty$ and coverage is certain. Thus, in either case,

$$\mathbb{P}(R_{n_{\mathrm{cal}}+1} \le \widehat{\gamma}^{\mathrm{cal}} \mid \mathcal{F}_n) \ge \frac{k_{\mathrm{cal}}}{n_{\mathrm{cal}} + 1} \ge 1 - \alpha.$$

When all scores are almost surely distinct, the coverage event is exactly the event that the test score has rank at most k_{cal}. Its conditional probability then equals $k_{\mathrm{cal}}/(n_{\mathrm{cal}}+1) \le 1 - \alpha + 1/(n_{\mathrm{cal}}+1)$. Taking expectations gives (16).

□

B.4 Consistency lemmas

The proof of Theorem 2 relies on five intermediate results that we state and prove below. These lemmas establish, in order, score and level stability under the data law (Lemma 1), consistency of the learned level (Lemma 2), vanishing of the calibration adjustment (Lemma 3), indicator consistency (Lemma 4), and symmetric-difference convergence (Lemma 5).

We first note that Lebesgue-domain consistency implies data-law consistency: splitting the data-law error at $p(Y \mid X) = K$ and using $(\widehat{D}_n - D^\star)^2 \le 1$ gives, for every $K > 0$,

$$\varepsilon_{D,n} \le K \varepsilon_{\mu,n} + \mathbb{P}(p(Y \mid X) > K)\,. \tag{37}$$

Because $p(Y \mid X) < \infty$ almost surely under the data law, the second term tends to zero as $K \to \infty$. Thus (20) implies (19); the two conditions remain distinct because the geometric conclusion requires Lebesgue-domain control.

Lemma 1 (Fitted-score and quantile stability)**.** *Under Assumptions 1 and 2, suppose (19) holds. Then, for every $\varepsilon > 0$, $\int r_{n,\varepsilon}(x)\, dP_X(x) = o_P(1)$, $P_X(\mathcal{X} \setminus \mathcal{X}_{n,\varepsilon}) = o_P(1)$, and*

$$\mathbb{P}\left(|\widehat{s}_n(X,Y) - s^\star(X,Y)| \geq \varepsilon \,|\, \mathcal{F}_n\right) = o_P(1), \tag{38}$$

$$P_X(\{x : \mathcal{T}_n(x) \not\subset (t^\star_\alpha(x) - \varepsilon, t^\star_\alpha(x) + \varepsilon)\}) = o_P(1). \tag{39}$$

Proof. Under the data law, $0 < p(Y \mid X) < \infty$ almost surely, hence $D^\star(X,Y) \in (0,1)$ almost surely. Fix $\zeta \in (0, 1/2)$. Uniform continuity of ψ^{-1} on $[\zeta/2, 1-\zeta/2]$ gives $\delta_{\zeta,\varepsilon} \in (0, \zeta/2)$ such that $v \in [\zeta, 1-\zeta]$ and $|v' - v| < \delta_{\zeta,\varepsilon}$ imply $|\psi^{-1}(v') - \psi^{-1}(v)| < \varepsilon$. Conditional Markov's inequality therefore yields

$$\mathbb{P}(|\widehat{s}_n(X,Y) - s^\star(X,Y)| \geq \varepsilon \mid \mathcal{F}_n) \leq \mathbb{P}(D^\star(X,Y) \notin [\zeta, 1-\zeta]) + \varepsilon_{D,n}/\delta^2_{\zeta,\varepsilon}.$$

Letting $n \to \infty$ and then $\zeta \downarrow 0$ proves (38); the tower property identifies its left side with $\int r_{n,\varepsilon}\, dP_X$. For any $\zeta > 0$, the definition of $\mathcal{X}_{n,\varepsilon}$ gives

$$P_X(\mathcal{X} \setminus \mathcal{X}_{n,\varepsilon}) \leq P_X(\kappa_\varepsilon(X) \leq \zeta) + \frac{2}{\zeta} \int r_{n,\varepsilon}(x)\, dP_X(x).$$

The second term vanishes in probability for fixed ζ; the first tends to zero as $\zeta \downarrow 0$ because $\kappa_\varepsilon > 0$ almost surely.

Fix $x \in \mathcal{X}_{n,\varepsilon}$. Outside an event of conditional probability $r_{n,\varepsilon}(x)$, $\widehat{s}_n(x,Y) \leq t^\star_\alpha(x) - 2\varepsilon$ implies $s^\star(x,Y) < t^\star_\alpha(x) - \varepsilon$, whereas $s^\star(x,Y) \leq t^\star_\alpha(x) + \varepsilon$ implies $\widehat{s}_n(x,Y) < t^\star_\alpha(x) + 2\varepsilon$. Thus

$$\begin{cases} \widehat{F}_{n,x}(t^\star_\alpha(x) - 2\varepsilon) \leq 1 - \alpha - \kappa_\varepsilon(x)/2, \\ \widehat{F}_{n,x}((t^\star_\alpha(x) + 2\varepsilon)^-) \geq 1 - \alpha + \kappa_\varepsilon(x)/2. \end{cases} \tag{40}$$

The quantile-set definition and monotonicity of the CDF now imply

$$\mathcal{T}_n(x) \subset (t^\star_\alpha(x) - 2\varepsilon, t^\star_\alpha(x) + 2\varepsilon).$$

Applying this conclusion with $\varepsilon/2$, and using $P_X(\mathcal{X} \setminus \mathcal{X}_{n,\varepsilon/2}) = o_P(1)$, proves (39). □

Lemma 2 (Consistency of the learned level and centered score)**.** *Under the assumptions of Lemma 1, suppose additionally that (21) holds. Then, for every $\varepsilon > 0$,*

$$P_X\left(\left\{x : |\widehat{t}_n(x) - t^\star_\alpha(x)| \geq \varepsilon\right\}\right) = o_P(1), \tag{41}$$

$$\mathbb{P}\left(|\widehat{g}_n(X,Y) - g^\star_\alpha(X,Y)| \geq \varepsilon \,|\, \mathcal{F}_n\right) = o_P(1). \tag{42}$$

Proof. If $\mathcal{T}_n(x) \subset (t^\star_\alpha(x) - \varepsilon/2, t^\star_\alpha(x) + \varepsilon/2)$ and $\operatorname{dist}(\widehat{t}_n(x), \mathcal{T}_n(x)) < \varepsilon/2$, then $|\widehat{t}_n(x) - t^\star_\alpha(x)| < \varepsilon$. Moreover,

$$P_X\left(\operatorname{dist}(\widehat{t}_n(X), \mathcal{T}_n(X)) \geq \varepsilon/2\right) \leq \frac{\varepsilon_{t,n}}{\min\{1, \varepsilon^2/4\}} = o_P(1).$$

Combining this with (39) proves (41). Finally,

$$|\widehat{g}_n(X,Y) - g^\star_\alpha(X,Y)| \leq |\widehat{s}_n(X,Y) - s^\star(X,Y)| + |\widehat{t}_n(X) - t^\star_\alpha(X)|.$$

A union bound together with (38) and (41) proves (42). □

Lemma 3 (Calibration adjustment vanishes)**.** *Under Assumptions 1 and 2, together with (19) and (21) in Assumption 3, we have $\widehat{\gamma}^{\mathrm{cal}}_n \xrightarrow{P} 0$.*

Proof. Define $G_n(u) := \mathbb{P}\left(\widehat{g}_n(X,Y) \leq u \,|\, \mathcal{F}_n\right)$ and $G_\star(u) := \mathbb{P}\left(g^\star_\alpha(X,Y) \leq u\right)$. Fix $\varepsilon > 0$. Integrating the strict conditional inequalities in (18) gives

$$G_\star(-\varepsilon/2) < 1 - \alpha < G_\star(\varepsilon/2).$$

By Lemma 2, $\bar{r}_n := \mathbb{P}(|\widehat{g}_n(X,Y) - g^\star_\alpha(X,Y)| \geq \varepsilon/2 \,|\, \mathcal{F}_n) = o_P(1)$. Outside this error event,

$$\widehat{g}_n(X,Y) \leq -\varepsilon \implies g^\star_\alpha(X,Y) < -\varepsilon/2$$
$$g^\star_\alpha(X,Y) \leq \varepsilon/2 \implies \widehat{g}_n(X,Y) < \varepsilon.$$

Therefore

$$G_n(-\varepsilon) \le G_\star(-\varepsilon/2) + \bar{r}_n$$
$$G_n(\varepsilon) \ge G_\star(\varepsilon/2) - \bar{r}_n.$$

Hence, with the deterministic constant

$$\delta_\varepsilon := \tfrac{1}{2}\min\{(1-\alpha) - G_\star(-\varepsilon/2),\ G_\star(\varepsilon/2) - (1-\alpha)\} > 0,$$

the event $\{\bar{r}_n \le \delta_\varepsilon\}$ has probability tending to one, and on it

$$G_n(-\varepsilon) \le 1 - \alpha - \delta_\varepsilon, \quad G_n(\varepsilon) \ge 1 - \alpha + \delta_\varepsilon.$$

Conditional on $\mathcal{F}_n$, the calibration scores are i.i.d. with CDF G_n. Let $\widehat{G}_n$ be their empirical CDF. At each of the two fixed arguments $u \in \{-\varepsilon, \varepsilon\}$,

$$\mathbb{E}\left[\left(\widehat{G}_n(u) - G_n(u)\right)^2 \,\middle|\, \mathcal{F}_n\right] \le \frac{1}{4n_{\mathrm{cal}}},$$

so $\widehat{G}_n(u) - G_n(u) = o_P(1)$. Let $k_{\mathrm{cal}} := \lceil (n_{\mathrm{cal}}+1)(1-\alpha) \rceil$. Since $1-\alpha < 1$ and $n_{\mathrm{cal}} \to \infty$, eventually $k_{\mathrm{cal}} \le n_{\mathrm{cal}}$ and $k_{\mathrm{cal}}/n_{\mathrm{cal}} \to 1-\alpha$. Thus, with probability tending to one,

$$\widehat{G}_n(-\varepsilon) < \frac{k_{\mathrm{cal}}}{n_{\mathrm{cal}}} \le \widehat{G}_n(\varepsilon).$$

The k_{cal}-th calibration order statistic therefore satisfies

$$-\varepsilon < \widehat{\gamma}_n^{\mathrm{cal}} \le \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary, the result follows. □

Lemma 4 (Indicator consistency under the data law)**.** *Under the assumptions of Theorem 2(i),*

$$\mathbb{P}\Big(\mathbf{1}\left\{Y \in \widehat{C}_{\alpha,n}(X)\right\} \ne \mathbf{1}\left\{Y \in C^\star_\alpha(X)\right\} \,\Big|\, \mathcal{H}_n\Big) = o_P(1). \tag{43}$$

Proof. The bound $|\psi'(u)| \le 1/(4\beta)$, together with Lemmas 2 and 3, yields for every $\delta > 0$

$$\begin{aligned} P_X(|\widehat{\vartheta}_n(X) - \vartheta^\star_\alpha(X)| > \delta) &\le P_X(|\widehat{t}_n(X) - t^\star_\alpha(X)| > 2\beta\delta) \\ &\quad + \mathbf{1}\{|\widehat{\gamma}_n^{\mathrm{cal}}| > 2\beta\delta\} = o_P(1). \end{aligned} \tag{44}$$

Because the fresh observation is independent of $\mathcal{H}_n$ and the fitted discriminator is $\mathcal{F}_n$-measurable,

$$\mathbb{E}\left[\left(\widehat{D}_n(X,Y) - D^\star(X,Y)\right)^2 \,\middle|\, \mathcal{H}_n\right] = \varepsilon_{D,n}.$$

The oracle boundary has zero probability under the data law:

$$\begin{aligned} \mathbb{P}\left(D^\star(X,Y) = \vartheta^\star_\alpha(X)\right) &= \mathbb{P}\left(p(Y \mid X) = c_\alpha(X)\right) \\ &= \int c_\alpha(x)\, \mu\{y : p(y \mid x) = c_\alpha(x)\}\, dP_X(x) = 0. \end{aligned}$$

For fixed $\delta > 0$, the deterministic comparison above and Markov's inequality give

$$\begin{aligned} &\mathbb{P}\left(\mathbf{1}\left\{Y \in \widehat{C}_{\alpha,n}(X)\right\} \ne \mathbf{1}\left\{Y \in C^\star_\alpha(X)\right\} \,\middle|\, \mathcal{H}_n\right) \\ &\quad \le \frac{\varepsilon_{D,n}}{\delta^2} + P_X\left(|\widehat{\vartheta}_n(X) - \vartheta^\star_\alpha(X)| > \delta\right) + \mathbb{P}(|D^\star(X,Y) - \vartheta^\star_\alpha(X)| \le 2\delta). \end{aligned}$$

For fixed δ, the first two terms are $o_P(1)$. The last term is deterministic and tends to zero as $\delta \downarrow 0$ by continuity from above. Taking first $n \to \infty$ and then $\delta \downarrow 0$ proves the result. □

Lemma 5 (Symmetric-difference consistency)**.** *Under the assumptions of Theorem 2(ii), for every* $\varepsilon > 0$*,*

$$P_X\left(\left\{x : \mu\Big(\widehat{C}_{\alpha,n}(x) \triangle C^\star_\alpha(x)\Big) > \varepsilon\right\}\right) = o_P(1). \tag{45}$$

Consequently, for a fresh $X \sim P_X$ *independent of* $\mathcal{H}_n$*,*

$$\mu\left(\widehat{C}_{\alpha,n}(X) \triangle C^\star_\alpha(X)\right) \xrightarrow{P} 0.$$

Proof. Let

$$L_n(x) := \int_{\mathcal{Y}} \left(\widehat{D}_n(x,y) - D^\star(x,y)\right)^2 d\mu(y),$$
$$V_\delta(x) := \mu\{y : |D^\star(x,y) - \vartheta^\star_\alpha(x)| \le \delta\}.$$

Then

$$\int L_n(x)\, dP_X(x) = \varepsilon_{\mu,n}.$$

Fix x satisfying Assumption 2 and put $c = c_\alpha(x)$. Since $v \mapsto v/(v+\lambda)$ is strictly increasing,

$$\delta_x := \frac{1}{2}\left\{\frac{c}{c+\lambda} - \frac{c}{c+2\lambda}\right\} > 0.$$

If $|D^\star(x,y) - \vartheta^\star_\alpha(x)| \le \delta_x$, then $p(y \mid x) > c/2$. Therefore

$$V_{\delta_x}(x) \le \mu\{y : p(y \mid x) \ge c/2\} \le \frac{2}{c} < \infty.$$

As $\delta \downarrow 0$, the boundary bands decrease to $\{y : p(y \mid x) = c\}$. Since $V_{\delta_x}(x) < \infty$, continuity from above applies to these bands; the zero-contour condition therefore gives

$$V_\delta(x) \longrightarrow 0 \qquad (\delta \downarrow 0) \tag{46}$$

for P_X-almost every x. Hence, for every $v > 0$,

$$P_X(V_\delta(X) > v) \longrightarrow 0 \qquad (\delta \downarrow 0).$$

On the event $|\widehat{\vartheta}_n(x) - \vartheta^\star_\alpha(x)| \le \delta$, the deterministic comparison gives

$$\widehat{C}_{\alpha,n}(x) \triangle C^\star_\alpha(x) \subseteq \{y : |\widehat{D}_n(x,y) - D^\star(x,y)| > \delta\} \cup \{y : |D^\star(x,y) - \vartheta^\star_\alpha(x)| \le 2\delta\}.$$

Therefore

$$\mu\Big(\widehat{C}_{\alpha,n}(x) \triangle C^\star_\alpha(x)\Big) \le \frac{L_n(x)}{\delta^2} + V_{2\delta}(x).$$

For every $\varepsilon, \delta > 0$,

$$P_X\Big(\mu(\widehat{C}_{\alpha,n}(X) \triangle C^\star_\alpha(X)) > \varepsilon\Big)$$
$$\le P_X\Big(|\widehat{\vartheta}_n(X) - \vartheta^\star_\alpha(X)| > \delta\Big) + \frac{2\varepsilon_{\mu,n}}{\varepsilon\delta^2} + P_X(V_{2\delta}(X) > \varepsilon/2).$$

For fixed δ, the first two terms are $o_P(1)$. The last term tends to zero as $\delta \downarrow 0$ by (46). This proves (45).

The random quantity on the left of (45) takes values in $[0,1]$, so convergence to zero in probability implies convergence of its expectation to zero. Independence of the fresh X from $\mathcal{H}_n$ identifies this expectation with

$$\mathbb{P}\Big(\mu(\widehat{C}_{\alpha,n}(X) \triangle C^\star_\alpha(X)) > \varepsilon\Big),$$

which proves the final assertion. □

B.5 Proof of Theorem 2

Proof. Lemma 3 proves $\widehat{\gamma}^{\mathrm{cal}}_n \xrightarrow{P} 0$.

For conditional coverage, define the $\mathcal{H}_n$-measurable function

$$e^{\mathrm{cov}}_n(x) := \Big|\mathbb{P}\Big(Y \in \widehat{C}_{\alpha,n}(x) \mid X = x, \mathcal{H}_n\Big) - (1-\alpha)\Big|.$$

Since $\mathbb{P}(Y \in C^\star_\alpha(x) \mid X = x) = 1 - \alpha$, Lemma 4 yields

$$\bar{e}^{\mathrm{cov}}_n := \int e^{\mathrm{cov}}_n(x)\, dP_X(x)$$
$$\le \mathbb{P}\Big(\mathbf{1}\{Y \in \widehat{C}_{\alpha,n}(X)\} \ne \mathbf{1}\{Y \in C^\star_\alpha(X)\} \,\Big|\, \mathcal{H}_n\Big) = o_P(1).$$

We can set

$$\xi_n := \sqrt{\bar{e}_n^{\mathrm{cov}}} + n^{-1}, \qquad \Lambda_n := \{x : e_n^{\mathrm{cov}}(x) \le \xi_n\}.$$

Then

$$P_X(\Lambda_n^c) \le \frac{\bar{e}_n^{\mathrm{cov}}}{\xi_n} \le \sqrt{\bar{e}_n^{\mathrm{cov}}} = o_P(1), \qquad \sup_{x \in \Lambda_n} e_n^{\mathrm{cov}}(x) \le \xi_n = o_P(1).$$

Because the fresh X is independent of $\mathcal{H}_n$,

$$\mathbb{P}(X \in \Lambda_n \mid \mathcal{H}_n) = P_X(\Lambda_n) = 1 - o_P(1),$$

which proves asymptotic conditional coverage in the sense of Definition 1.

Under the additional Lebesgue-domain condition, Lemma 5 proves (22). □

B.6 Sufficient conditions in terms of population losses

The consistency conditions above concern the fitted functions. The following propositions give sufficient conditions in terms of contrastive excess risk and conditional pinball regret. For the first, the Bayes risk $\mathcal{R}^\star_{\lambda,\beta} := \mathcal{R}_{\lambda,\beta}(s^\star)$ has the explicit form

$$\beta^{-1}\mathcal{R}^\star_{\lambda,\beta} = \mathbb{E}\left[\log\left(1 + \frac{\lambda}{p(Y \mid X)}\right)\right] + \lambda \mathbb{E}_X \int_{\mathcal{Y}} \log\left(1 + \frac{p(y \mid X)}{\lambda}\right) d\mu(y). \tag{47}$$

The second term is at most one by $\log(1+u) \le u$. Thus $\mathcal{R}^\star_{\lambda,\beta} < \infty$ precisely when the first term is finite. In particular, finite $\mu(\mathcal{Y})$ suffices, since the constant score $s = 0$ then has finite risk.

Proposition 2 (Contrastive excess risk implies discriminator consistency)**.** *Let $\mathcal{R}^\star_{\lambda,\beta}$ denote the integrated pointwise Bayes risk, and assume $\mathcal{R}^\star_{\lambda,\beta} < \infty$. For every measurable real-valued score s,*

$$\begin{aligned}\mathcal{R}_{\lambda,\beta}(s) - \mathcal{R}^\star_{\lambda,\beta} = \beta\, \mathbb{E}_X \int_{\mathcal{Y}} \{p(y \mid X) + \lambda\} \times \\ \mathrm{KL}(\mathrm{Bern}(D^\star(X,y)) \,\|\, \mathrm{Bern}(D_s(X,y)))\; d\mu(y),\end{aligned} \tag{48}$$

with both sides interpreted in $[0,\infty]$. Consequently,

$$\mathbb{E}\big[(D_s(X,Y) - D^\star(X,Y))^2\big] \le \frac{\mathcal{R}_{\lambda,\beta}(s) - \mathcal{R}^\star_{\lambda,\beta}}{2\beta}, \tag{49}$$

$$\mathbb{E}_X \int_{\mathcal{Y}} (D_s(X,y) - D^\star(X,y))^2\, d\mu(y) \le \frac{\mathcal{R}_{\lambda,\beta}(s) - \mathcal{R}^\star_{\lambda,\beta}}{2\beta\lambda}. \tag{50}$$

In particular, writing $\mathcal{R}_{\lambda,\beta}(\widehat{s}_n)$ for the population risk of the fitted score evaluated conditionally on $\mathcal{F}_n$, the condition $\mathcal{R}_{\lambda,\beta}(\widehat{s}_n) - \mathcal{R}^\star_{\lambda,\beta} = o_P(1)$ implies both (19) and (20).

Proof. At a fixed (x,y), write $p = p(y \mid x)$, $D^\star = D^\star(x,y)$, and $D = D_s(x,y)$. Since

$$p = (p+\lambda)D^\star, \qquad \lambda = (p+\lambda)(1 - D^\star),$$

subtracting the pointwise Bayes value from the pointwise loss gives

$$\beta(p+\lambda)\left[D^\star \log\frac{D^\star}{D} + (1-D^\star)\log\frac{1-D^\star}{1-D}\right].$$

This is exactly the integrand in (48), including the case $p = 0$. It is nonnegative, so integration establishes the identity.

Pinsker's inequality for Bernoulli laws with natural logarithms gives

$$\mathrm{KL}(\mathrm{Bern}(u) \| \mathrm{Bern}(v)) \ge 2(u-v)^2.$$

Substituting this bound into (48) yields

$$\mathcal{R}_{\lambda,\beta}(s) - \mathcal{R}^\star_{\lambda,\beta} \ge 2\beta\left[\mathbb{E}(D_s(X,Y) - D^\star(X,Y))^2 + \lambda \mathbb{E}_X \int_{\mathcal{Y}} (D_s - D^\star)^2\, d\mu\right].$$

Dropping either nonnegative term proves (49) and (50). Applying these deterministic inequalities to each realization of the fitted score proves the final statement. □

Proposition 3 (Pinball excess risk implies level consistency)**.** *For fixed x, define the centered conditional pinball risk*

$$\mathcal{Q}^{\circ}_{n,x}(u) := \mathbb{E}[\rho_{1-\alpha}(\widehat{s}_n(x,Y) - u) - \rho_{1-\alpha}(\widehat{s}_n(x,Y)) \,|\, X = x, \mathcal{F}_n], \qquad u \in \mathbb{R}. \tag{51}$$

It is finite for every u and has minimizer set $\mathcal{T}_n(x)$. Using the selector $\underline{t}_n(x) \in \mathcal{T}_n(x)$ defined above, the population pinball regret is

$$\mathcal{E}^{\mathrm{pb}}_n := \int \left[\mathcal{Q}^{\circ}_{n,x}(\widehat{t}_n(x)) - \mathcal{Q}^{\circ}_{n,x}(\underline{t}_n(x))\right] \, dP_X(x). \tag{52}$$

Under Assumptions 1 and 2, if (19) holds and $\mathcal{E}^{\mathrm{pb}}_n = o_P(1)$, then (21) holds.

Proof. The pinball loss is 1-Lipschitz, so the absolute value of the difference inside (51) is at most $|u|$. Thus the centered risk is finite without any moment assumption on $\widehat{s}_n(x,Y)$. It is convex, with subdifferential

$$\partial \mathcal{Q}^{\circ}_{n,x}(u) = \left[\widehat{F}_{n,x}(u^-) - (1-\alpha),\ \widehat{F}_{n,x}(u) - (1-\alpha)\right].$$

Its minimizer set is therefore exactly $\mathcal{T}_n(x)$.

For every $u \in \mathbb{R}$, integration of the pinball-loss increment gives

$$\begin{aligned} \Delta_{n,x}(u) &:= \mathcal{Q}^{\circ}_{n,x}(u) - \mathcal{Q}^{\circ}_{n,x}(\underline{t}_n(x)) \\ &= \int_{\underline{t}_n(x)}^{u} \left[\widehat{F}_{n,x}(v) - (1-\alpha)\right] \, dv \geq 0. \end{aligned} \tag{53}$$

The identity follows by integrating the almost-everywhere derivative over the finite interval between $\underline{t}_n(x)$ and u; the CDF and its left limit differ only at countably many points.

Fix $\varepsilon > 0$. Lemma 1, applied at $\varepsilon/4$, gives $P_X(\mathcal{X} \setminus \mathcal{X}_{n,\varepsilon/4}) = o_P(1)$. For $x \in \mathcal{X}_{n,\varepsilon/4}$, (40) becomes

$$\begin{cases} \widehat{F}_{n,x}(t^\star_\alpha(x) - \varepsilon/2) \leq 1 - \alpha - \kappa_{\varepsilon/4}(x)/2, \\ \widehat{F}_{n,x}((t^\star_\alpha(x) + \varepsilon/2)^-) \geq 1 - \alpha + \kappa_{\varepsilon/4}(x)/2. \end{cases}$$

If $x \in \mathcal{X}_{n,\varepsilon/4}$ and $u \geq t^\star_\alpha(x) + \varepsilon$, the portion of (53) from $t^\star_\alpha(x) + \varepsilon/2$ to u is at least $\varepsilon\kappa_{\varepsilon/4}(x)/4$. Integrating on the opposite side gives the same bound for $u \leq t^\star_\alpha(x) - \varepsilon$. Thus

$$x \in \mathcal{X}_{n,\varepsilon/4}, \quad |u - t^\star_\alpha(x)| \geq \varepsilon \quad \Longrightarrow \quad \Delta_{n,x}(u) \geq \varepsilon\kappa_{\varepsilon/4}(x)/4.$$

For every $\zeta > 0$,

$$P_X(|\widehat{t}_n(X) - t^\star_\alpha(X)| \geq \varepsilon) \leq P_X(\mathcal{X} \setminus \mathcal{X}_{n,\varepsilon/4}) + P_X(\kappa_{\varepsilon/4}(X) \leq \zeta) + \frac{4\mathcal{E}^{\mathrm{pb}}_n}{\varepsilon\zeta}.$$

Taking first $n \to \infty$ and then $\zeta \downarrow 0$ proves

$$P_X\left(|\widehat{t}_n(X) - t^\star_\alpha(X)| \geq \varepsilon\right) = o_P(1).$$

By (39), the quantile selector $\underline{t}_n(X)$ converges to $t^\star_\alpha(X)$ in the same sense. Since

$$\mathrm{dist}(\widehat{t}_n(x), \mathcal{T}_n(x)) \leq |\widehat{t}_n(x) - \underline{t}_n(x)|,$$

the distance to the fitted-score quantile set converges to zero in P_X-probability. Finally, for any $v \in (0,1)$,

$$\varepsilon_{t,n} \leq v^2 + P_X\left(\mathrm{dist}(\widehat{t}_n(X), \mathcal{T}_n(X)) > v\right).$$

Taking $n \to \infty$ and then $v \downarrow 0$ gives (21). □

Together, Propositions 2 and 3 show that consistency of the population contrastive excess risk and the fitted-score population pinball regret is sufficient for the estimation conditions used in Theorem 2, provided the Bayes contrastive risk is finite.

B.7 IMPORTANCE-SAMPLING APPROXIMATION

Proposition 4 (Unbiased importance-sampling estimate)**.** *Let $\mathcal{G}$ be a σ-field with respect to which the training observations, current score s, and proposal densities q_i are measurable. Suppose $0 < q_i(y) < \infty$ for μ-almost every y and*

$$I_i(s) := \int_{\mathcal{Y}} \ell_\beta^-(s(X_i, y))\, d\mu(y) < \infty.$$

Conditional on $\mathcal{G}$, draw $Y_{ij}^- \overset{\text{iid}}{\sim} q_i$, $j = 1, \dots, M$. Then

$$\mathbb{E}\left[\frac{1}{M} \sum_{j=1}^{M} \frac{\ell_\beta^-(s(X_i, Y_{ij}^-))}{q_i(Y_{ij}^-)} \,\middle|\, \mathcal{G} \right] = I_i(s).$$

The proposal may depend on (X_i, Y_i) and on the preceding fitting history.

Proof. Conditionally on $\mathcal{G}$, the score and proposal are fixed, so

$$\mathbb{E}\left[\frac{\ell_\beta^-(s(X_i, Y_{ij}^-))}{q_i(Y_{ij}^-)} \,\middle|\, \mathcal{G} \right] = \int_{\mathcal{Y}} \frac{\ell_\beta^-(s(X_i, y))}{q_i(y)} q_i(y)\, d\mu(y) = I_i(s).$$

Averaging over j proves the claim. □

Remark 1 (Tail behavior of the importance estimator)**.** *Write $\xi_j := \ell_\beta^-(s(X_i, Y_{ij}^-))/q_i(Y_{ij}^-) \geq 0$. By Proposition 4, $\mathbb{E}[\xi_j \mid \mathcal{G}] = I_i(s)$ whenever $I_i(s) < \infty$, but finite variance additionally requires tail overlap between integrand and proposal. At the population optimum,*

$$\mathbb{E}[\xi_j^2 \mid \mathcal{G}] = \beta^2 \int_{\mathcal{Y}} \frac{\log^2\{1 + p(y \mid X_i)/\lambda\}}{q_i(y)}\, d\mu(y) \leq \frac{\beta^2}{\lambda^2} \int_{\mathcal{Y}} \frac{p(y \mid X_i)^2}{q_i(y)}\, d\mu(y),$$

which is finite whenever $\int_{\mathcal{Y}} p(y \mid X_i)^2/q_i(y)\, d\mu(y) < \infty$. However, on $\mathcal{Y} = \mathbb{R}^d$, a globally Lipschitz fitted score (as produced by any feed-forward network with Lipschitz activations) makes $\ell_\beta^-(s(x, y))$ decay at most exponentially in $\|y\|$, whereas $1/q_i(y)$ grows at least like $\exp(c\|y\|^2)$ for a finite Gaussian-mixture proposal, so $\mathbb{E}[\xi_j^r \mid \mathcal{G}] = \infty$ for every $r > 1$. The estimator remains almost surely convergent by the strong law whenever $I_i(s) < \infty$ (e.g., when $s(x, y) \geq a\|y\| - b$ for some $a > 0$), but the usual finite-variance $M^{-1/2}$ rate and central limit theorem are unavailable.

Two modifications may address this issue. First, a polynomially tailed proposal (e.g., a Student-t mixture) yields $\mathbb{E}[\xi_j^2 \mid \mathcal{G}] < \infty$ under the same linear-growth condition, as the exponential decay of $\ell_\beta^-(s)^2$ dominates the polynomial growth of $1/q_i$. Second, one may truncate: for $\tau_M \to \infty$ with $\tau_M/M \to 0$, the estimator

$$\widehat{I}_{i,M}^{\tau} := \frac{1}{M} \sum_{j=1}^{M} (\xi_j \wedge \tau_M).$$

satisfies $\mathbb{E}[(\widehat{I}_{i,M}^{\tau} - I_i(s))^2 \mid \mathcal{G}] \to 0$ with no moment assumption beyond the first, since the bias $\mathbb{E}[(\xi_j - \tau_M)_+ \mid \mathcal{G}] \to 0$ by dominated convergence and $\mathrm{Var}(\widehat{I}_{i,M}^{\tau} \mid \mathcal{G}) \leq \tau_M I_i(s)/M \to 0$.

B.8 THEORETICAL PROPERTIES OF RES-LECP

The results below characterize the population residual target, the reference approximation, and score alignment. For the boundary weight in (25), $w_{\min} \leq w_{b-1} \leq 1$; hence data-weight integrability is automatic, while reference normalizability remains a separate condition.

Proposition 5 (Population target of one residual stage)**.** *Suppose the conditional density exists. Fix jointly measurable functions $s^{[b-1]} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $w_{b-1} : \mathcal{X} \times \mathcal{Y} \to (0, \infty)$. Assume $\mathbb{E}[w_{b-1}(X, Y)] < \infty$ and, for P_X-almost every x,*

$$0 < Z_{b-1}(x) := \int_{\mathcal{Y}} w_{b-1}(x, y) e^{-s^{[b-1]}(x,y)/\beta}\, d\mu(y) < \infty.$$

Define $\pi_{b-1}(y \mid x) = w_{b-1}(x,y)e^{-s^{[b-1]}(x,y)/\beta}/Z_{b-1}(x)$. *For the residual population risk*

$$\mathcal{R}_b(\delta) = \mathbb{E}\Big[w_{b-1}(X,Y)\ell_\beta^+(\delta(X,Y))\Big] + \mathbb{E}_X \int_{\mathcal{Y}} \ell_\beta^-(\delta(X,y))\pi_{b-1}(y \mid X)\, d\mu(y),$$

the extended-valued Bayes correction is

$$\delta_b^\star(x,y) = -s^{[b-1]}(x,y) - \beta \log p(y \mid x) - \beta \log Z_{b-1}(x) \tag{54}$$

on $\{p > 0\}$, *with* $\delta_b^\star(x,y) = +\infty$ *on* $\{p = 0\}$. *For a residual step size* $\eta \in (0,1]$, *the corresponding population update* $s^{[b]} = s^{[b-1]} + \eta\delta_b^\star$ *satisfies*

$$s^{[b]}(x,y) = (1-\eta)s^{[b-1]}(x,y) - \eta\beta \log p(y \mid x) - \eta\beta \log Z_{b-1}(x). \tag{55}$$

In particular, $\eta = 1$ *recovers the conditional negative log-density up to an additive function of* x.

Proof. Condition on $X = x$. The two contrast measures have densities $w_{b-1}(x,y)p(y \mid x)$ and $\pi_{b-1}(y \mid x)$ with respect to μ. The first is integrable for P_X-almost every x by assumption, and the second is a probability density. Moreover,

$$\mathcal{R}_b(0) = \beta \log 2 \left\{\mathbb{E}[w_{b-1}(X,Y)] + 1\right\} < \infty.$$

At points where $p(y \mid x) > 0$, both densities are positive. Pointwise minimization of the logistic loss gives

$$\delta = \beta \log \frac{\pi_{b-1}(y \mid x)}{w_{b-1}(x,y)p(y \mid x)} = -s^{[b-1]}(x,y) - \beta \log p(y \mid x) - \beta \log Z_{b-1}(x).$$

Thus the explicit y-dependence of the common positive weight cancels from the density ratio, while normalization contributes only the x-dependent offset $-\beta \log Z_{b-1}(x)$. If $p(y \mid x) = 0$, the positive term vanishes and the minimum over extended-valued corrections is attained at $\delta = +\infty$. Pointwise comparison and integration establish population optimality. Substituting (54) into $s^{[b]} = s^{[b-1]} + \eta\delta_b^\star$ gives (55). □

Proposition 6 (Self-normalized reference approximation)**.** *Fix* x, *frozen functions* $s^{[b-1]}, w_{b-1}$ *as in Proposition 5, and a proposal density* q *satisfying* $0 < q(y) < \infty$ *for* μ*-almost every* y. *Suppose* $0 < Z_{b-1}(x) < \infty$. *For* $Y_j^- \overset{\text{iid}}{\sim} q$, *define*

$$W_j = \frac{w_{b-1}(x, Y_j^-)e^{-s^{[b-1]}(x,Y_j^-)/\beta}}{q(Y_j^-)}.$$

For every fixed measurable f *satisfying*

$$\int_{\mathcal{Y}} |f(y)|\pi_{b-1}(y \mid x)\, d\mu(y) < \infty,$$

$$\frac{\sum_{j=1}^M W_j f(Y_j^-)}{\sum_{j=1}^M W_j} \to \int_{\mathcal{Y}} f(y)\pi_{b-1}(y \mid x)\, d\mu(y) \qquad \textit{almost surely as } M \to \infty.$$

Proof. The importance weights satisfy $\mathbb{E}_q[W_1] = Z_{b-1}(x) \in (0,\infty)$, and

$$\mathbb{E}_q[W_1|f(Y_1^-)|] = Z_{b-1}(x) \int_{\mathcal{Y}} |f(y)|\pi_{b-1}(y \mid x)\, d\mu(y) < \infty.$$

The strong law of large numbers therefore applies to numerator and denominator separately. Their almost-sure limits are

$$Z_{b-1}(x) \int f(y)\pi_{b-1}(y \mid x)\, d\mu(y)$$

and $Z_{b-1}(x)$, respectively. Taking the ratio proves the result. □

Corollary 1 (Oracle recovery after score alignment)**.** *Let $(\widehat{s}_n, \widehat{t}_n)$ satisfy Assumption 1. Let $a_n : \mathcal{X} \to \mathbb{R}$ be a real-valued random function, jointly measurable in x and the fitting randomness and $\mathcal{F}_n$-measurable, and define*

$$\widetilde{s}_n(x, y) = \widehat{s}_n(x, y) - a_n(x), \qquad \widetilde{t}_n(x) = \widehat{t}_n(x) - a_n(x).$$

Under Assumption 2, the corresponding conclusions of Theorem 2 hold for the original calibrated regions whenever the theorem's consistency conditions hold for the aligned pair $(\widetilde{s}_n, \widetilde{t}_n)$. The level-consistency condition is unchanged by this alignment.

Proof. Conditional on $X = x$ and $\mathcal{F}_n$, the offset $a_n(x)$ is fixed. Hence the conditional quantile set of $\widetilde{s}_n(x, Y)$ is $\widetilde{\mathcal{T}}_n(x) = \mathcal{T}_n(x) - a_n(x)$, and

$$\operatorname{dist}\left(\widetilde{t}_n(x), \widetilde{\mathcal{T}}_n(x)\right) = \operatorname{dist}\left(\widehat{t}_n(x), \mathcal{T}_n(x)\right).$$

Moreover, $\widetilde{s}_n(x, y) - \widetilde{t}_n(x) = \widehat{s}_n(x, y) - \widehat{t}_n(x)$. Thus alignment leaves every calibration score, the calibration quantile, and the final prediction region unchanged. Applying Theorem 2 to the aligned pair proves the claim. □

For the selected Res-LECP prefix, $\widehat{s}_n = s^{[b^\star]},\ \widehat{t}_n = t^{[0]} + u_{b^\star},\ u_0 = 0$. At a unit-step population residual solution,

$$s^{[b]}(x, y) = -\beta \log p(y \mid x) + a_b(x).$$

Subtracting $a_b(x) - \beta \log \lambda$ aligns this score with $s^\star(x, y)$. Equivalently, the corresponding oracle level is $-\beta \log c_\alpha(x) + a_b(x)$, so the centered population score is exactly

$$\beta \log \frac{c_\alpha(x)}{p(y \mid x)} = g^\star_\alpha(x, y).$$

Corollary 1 therefore provides the oracle-recovery statement for the selected Res-LECP fit whenever the aligned final pair satisfies the required consistency conditions.

C Experimental Details

C.1 Datasets and preprocessing

Table 3 summarizes the ten benchmarks. Each synthetic dataset contains 10,000 independent observations with a scalar covariate X. Comb transitions from one to two to five conditional modes as X increases; Rings is a two-component radial mixture with asymmetric X-scaled radial noise; Spiral has one-sided radial noise; and Lemniscate adds heteroskedastic noise around a figure-eight curve. These examples test discontinuous connectivity changes, holes, narrow curved supports, and self-intersecting geometry, respectively.

The four synthetic laws are defined below. All variables are conditionally independent given X unless noted otherwise, and $\mathrm{Exp}(m)$ denotes the exponential distribution with mean m.

- **Comb.** $X \sim \mathrm{Unif}(0, 1)$. Draw Z uniformly from

$$\mathcal{C}(x) = \begin{cases} \{0\}, & x < 0.35, \\ \{-1.5,\ 1.5\}, & 0.35 \le x < 0.70, \\ \{-3,\ -1.5,\ 0,\ 1.5,\ 3\}, & x \ge 0.70. \end{cases} \tag{56}$$

  Given Z, the response is

$$Y = \begin{cases} Z + \varepsilon, & Z = 0, \quad \varepsilon \sim \mathcal{N}(0,\ 0.15^2), \\ Z + \eta, & Z > 0, \quad \eta \sim \mathrm{Exp}(0.2), \\ Z - \eta, & Z < 0, \quad \eta \sim \mathrm{Exp}(0.2). \end{cases} \tag{57}$$

- **Rings.** $X \sim \mathrm{Unif}(0.1, 1)$, $B \sim \mathrm{Bernoulli}(1/2)$, $\Theta \sim \mathrm{Unif}(0, 2\pi)$, $\eta_1 \sim \mathrm{Exp}(1.5X)$, and $\eta_2 \sim \mathrm{Exp}(2X)$. Define

$$R = \begin{cases} \max\{4 - \eta_1,\ 0\}, & B = 0, \\ 6 + \eta_2, & B = 1, \end{cases} \qquad Y = 0.1\, R\,(\cos\Theta,\ \sin\Theta). \tag{58}$$

- **Spiral.** $X \sim \mathrm{Unif}(0.1, 1)$, $U \sim \mathrm{Unif}(0, 1)$, and $\Theta = 4\pi\sqrt{U}$. Define
$$\eta \mid (X, \Theta) \sim \mathrm{Exp}\big\{X(0.2 + 0.4\,\Theta)\big\}, \qquad Y = 0.1\,(1.5\,\Theta + \eta)\,(\cos\Theta, \sin\Theta). \tag{59}$$
- **Lemniscate.** $X \sim \mathrm{Unif}(0.1, 1)$, $T \sim \mathrm{Unif}(-\pi, \pi)$, and $\varepsilon_1, \varepsilon_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. With $\sigma(X, T) = 0.05 + 0.5\,X\,|T|/\pi$,
$$Y = 0.1\left(\frac{3.2\cos T}{1 + \sin^2 T} + \sigma(X, T)\,\varepsilon_1,\ \frac{6.4\sin T\cos T}{1 + \sin^2 T} + \sigma(X, T)\,\varepsilon_2\right). \tag{60}$$

Following the benchmark setup of Dheur et al. (2025), all real datasets use identical source definitions and preprocessing routines. After a random split, separate marginal quantile transforms map the covariates and responses to standard-normal margins. Both transforms are fitted on the proper training sample only and then applied unchanged to validation, calibration, and test samples. Volumes for real data are measured in this transformed response space.

Table 3: Benchmark datasets used in our experiments. Here n, p, and d denote sample size, covariate dimension, and response dimension.

Dataset	n	p	d	Type / source
Comb	10,000	1	1	Synthetic
Rings	10,000	1	2	Synthetic
Spiral	10,000	1	2	Synthetic
Lemniscate	10,000	1	2	Synthetic
Bio	45,730	8	2	Real; Rana (2013); Feldman et al. (2023)
House	21,613	17	2	Real; Harlfoxem (2016); Feldman et al. (2023)
Taxi	61,286	4	2	Real; Wang et al. (2023)
Households	7,207	14	4	Real; Camehl et al. (2025)
RF1	9,005	64	8	Real; Tsoumakas et al. (2011); Dheur et al. (2025)
RF2	9,005	64	8	Real; Tsoumakas et al. (2011); Dheur et al. (2025)

C.2 Splits and training

Data splitting and replication. For every dataset and seed, we allocate 60% of the observations to proper training, 10% to validation, 10% to conformal calibration, and 20% to testing. The same split is shared by all methods. Model selection and all hyperparameter decisions use only the training and validation samples; the calibration sample is reserved for final split-conformal calibration. The main benchmark uses ten seeds (reporting mean and standard deviation); the sensitivity and ablation studies use the first five seeds and report paired 95% t intervals against the same-seed default.

LECP and Res-LECP. Both x and y are encoded by two-layer networks of width 128; the score network and conditional-level network each have two hidden layers of width 256. The score and level are updated in alternation for at most $1{,}000$ epochs, with checkpoints evaluated every 20 epochs; the selected score is then followed by at most 200 epochs of level refinement with the score frozen. Each observed response is paired with $M = 1{,}000$ reference samples from an equally weighted mixture of two local diagonal Gaussians centered at Y_i with relative standard-deviation scales 0.5 and 1. Default hyperparameters are $\beta = 0.1$ and $\lambda = 0.01$; for RF1 and RF2, the reference intensity $\lambda = 1$ was selected from $\{0.01, 0.1, 1\}$ using validation-calibrated volume on a split; the remaining datasets use the default $\lambda = 0.01$. We report sensitivity to λ in Appendix D.2. Res-LECP adds $B = 2$ residual stages of at most 100 epochs each, with validation every 10 epochs and early-stopping patience of 3 checks. We set the residual step size $\eta = 1$, boundary bandwidth $h = 2$, and weight floor $w_{\min} = \alpha/2$. The final level refit uses at most 100 epochs with patience 15. The residual prefix with the smallest validation-calibrated volume is selected.

The validation-calibrated volume is estimated with $4{,}096$ Monte Carlo samples per validation covariate. For $d \le 2$, we use Sobol quasi-Monte Carlo integration over a bounding box whose sides extend 10% beyond the training-response range in each coordinate. For $d > 2$, we directly use importance sampling from the fitted flow shared with the flow-based baselines, matching the procedure described in Appendix C.3 for final test-set volume evaluation.

Flow-based baselines (C-HDR, L-CP, PCP, CP2-PCP, C-PCP). These five methods share a single conditional masked autoregressive flow (MAF) per dataset and seed, fitted by maximum likelihood. The MAF comprises three affine autoregressive transforms, each with two residual blocks of width 256, and a standard-Gaussian latent distribution; training runs for at most 500 epochs with validation-NLL early stopping (patience 50). The five methods differ only in their nonconformity scores: C-HDR ranks an observed response by its fitted density relative to conditional draws; L-CP uses the Euclidean norm of the inverse-flow latent coordinate; PCP measures nearest-neighbor distance to conditional draws; CP2-PCP normalizes this distance by an estimated conditional quantile of draw-to-draw nearest-neighbor distances; and C-PCP applies the empirical conditional CDF of the same distance. Each method uses 100 conditional draws per Monte Carlo pass.

SLS. SLS is fitted independently. Its model combines three conditional volume-preserving flow components (three coupling layers each) with a covariate-dependent frontier and conditional-quantile network. The frontier has hidden width 256 with three hidden layers; the quantile network uses two. The precision head uses full Cholesky parameterization for $d < 5$ and a low-rank parameterization otherwise. SLS trains the frontier for at most 1,000 epochs and refines the quantile network for at most 200 epochs, using the volume objective for $d < 5$ and its log-volume counterpart otherwise. All methods share the same held-out calibration split.

C.3 Evaluation

The reported volume and CC-MAE are defined in Equation (23). For each synthetic test covariate, we estimate $\pi_{\widehat{C}}(X_i)$ from 1,024 fresh draws from its conditional generator. Empirical marginal coverage is $n_{\text{test}}^{-1} \sum_i \mathbf{1}\{Y_i \in \widehat{C}(X_i)\}$. SLS volumes are available analytically. For every other method, we use a common importance-sampling estimator with 5,000 draws per test covariate from the fitted conditional-density proposal (the MAF used by flow-based baselines). For real data, put $Z = \mathbf{1}\{Y \in \widehat{C}(X)\}$ and $\tau = 1 - \alpha$. The population L_1-ERT of Braun et al. (2026b) is

$$\mathrm{ERT} = \mathbb{E}_X \big| \mathbb{E}[Z \mid X] - \tau \big|. \tag{61}$$

We estimate it with the cross-fitted classification excess-risk procedure by `covmetrics` (Braun et al., 2026b), using a LightGBM classifier. To define the empirical WSC, let $\mathcal{V}$ be the evaluated projection directions and $I(v, a, b) = \{i : a \le v^\top X_i \le b\}$. With $\delta \in (0, 1)$,

$$\widehat{\mathrm{WSC}}_\delta = \min_{\substack{v \in \mathcal{V},\, a < b: \\ |I(v,a,b)| \ge \delta n_{\text{test}}}} \frac{1}{|I(v,a,b)|} \sum_{i \in I(v,a,b)} Z_i. \tag{62}$$

This is the lowest observed coverage over admissible covariate slabs. The direction set $\mathcal{V}$ consists of 1,000 unit vectors drawn uniformly from $\mathbb{S}^{p-1}$, shared across all methods within each dataset and split. We use `covmetrics` with minimum slab mass $\delta = 0.10$ (Braun et al., 2026b).

D Additional Experimental Results

D.1 Results across miscoverage levels

Tables 4 and 5 present the complete results at $\alpha = 0.05$ and $\alpha = 0.20$, complementing the main-text results at $\alpha = 0.10$ (Tables 1–2). Since all methods are split-conformalized, marginal validity is expected by construction: indeed, mean marginal coverage falls within $[0.95, 0.97]$, $[0.89, 0.91]$, and $[0.79, 0.81]$ for the three nominal levels, respectively. We therefore focus the remaining discussion on volume and conditional reliability.

Volume performance. The volume rankings observed at $\alpha = 0.10$ persist across levels. On the four synthetic datasets, LECP or Res-LECP achieves the smallest volume in all twelve dataset–level combinations, with particularly large margins on Comb, where the multimodal structure rewards level-set thresholding. On the six real datasets, LECP or Res-LECP attains lower volume than SLS in 29 of 30 dataset–level comparisons, the sole exception being Bio at $\alpha = 0.20$ (difference 0.01), and the reductions relative to the remaining baselines are generally larger still. On House at $\alpha = 0.05$, for example, Res-LECP (1.08) reduces volume by 18% compared with SLS (1.32) and by over 40% compared with L-CP (1.84).

Conditional reliability. Turning to conditional reliability, LECP and Res-LECP achieve the lowest or tied-lowest CC-MAE in all twelve synthetic dataset–level combinations, confirming that the volume gains do not sacrifice conditional accuracy. On the real datasets, ERT for both variants remains within 0.01–0.03 of the best method in all eighteen settings. WSC is occasionally surpassed by baselines that produce substantially larger prediction sets. For instance, C-HDR on RF2 at $\alpha = 0.05$ reaches WSC 0.90 versus 0.84 for LECP, but at roughly 50% greater volume. So when volume and conditional reliability are considered jointly, LECP and Res-LECP offer the most favorable balance. The RF1 and RF2 results at $\alpha = 0.20$ further illustrate why volume alone can be misleading: PCP attains the smallest volume on both datasets (0.27 and 0.26), yet its WSC collapses to 0.23 and 0.24 and its ERT rises to 0.17, revealing severe under-coverage in certain subgroups. Res-LECP, by contrast, achieves only marginally larger volumes (0.28 on both) while maintaining WSC above 0.63 and ERT at 0.08, providing a far more reliable coverage profile.

Table 4: Complete results at $\alpha = 0.05$. Entries are mean (standard deviation) over ten splits; subscripts display standard deviations. Boldface marks the best mean in each row.

Dataset	Metric	LECP	Res-LECP	SLS	L-CP	C-HDR	C-PCP	CP2-PCP	PCP
Synthetic datasets									
Comb	Volume (↓)	$1.61_{(0.04)}$	$\mathbf{1.60}_{(0.04)}$	$2.15_{(0.07)}$	$4.29_{(0.14)}$	$4.50_{(0.19)}$	$5.89_{(0.18)}$	$5.83_{(0.16)}$	$5.45_{(0.10)}$
	CC-MAE (↓)	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$0.02_{(0.00)}$	$0.05_{(0.00)}$	$0.05_{(0.00)}$	$0.04_{(0.00)}$	$0.04_{(0.00)}$	$0.05_{(0.00)}$
Rings	Volume (↓)	$\mathbf{1.32}_{(0.03)}$	$\mathbf{1.32}_{(0.03)}$	$1.56_{(0.03)}$	$1.56_{(0.02)}$	$1.58_{(0.03)}$	$1.77_{(0.04)}$	$1.75_{(0.03)}$	$1.70_{(0.04)}$
	CC-MAE (↓)	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$0.02_{(0.00)}$	$0.02_{(0.00)}$	$0.02_{(0.00)}$	$0.03_{(0.00)}$
Spiral	Volume (↓)	$\mathbf{3.09}_{(0.06)}$	$\mathbf{3.09}_{(0.06)}$	$3.31_{(0.09)}$	$3.84_{(0.07)}$	$3.75_{(0.05)}$	$4.21_{(0.06)}$	$4.17_{(0.06)}$	$4.11_{(0.06)}$
	CC-MAE (↓)	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$0.02_{(0.00)}$	$0.02_{(0.00)}$	$0.02_{(0.00)}$	$0.03_{(0.00)}$
Lemnis-cate	Volume (↓)	$\mathbf{0.40}_{(0.01)}$	$\mathbf{0.40}_{(0.01)}$	$0.55_{(0.02)}$	$0.56_{(0.01)}$	$0.52_{(0.01)}$	$0.62_{(0.01)}$	$0.62_{(0.01)}$	$0.61_{(0.01)}$
	CC-MAE (↓)	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$0.02_{(0.00)}$	$0.02_{(0.00)}$	$0.02_{(0.00)}$	$0.02_{(0.00)}$
Real-world datasets									
Bio	Volume (↓)	$\mathbf{1.07}_{(0.02)}$	$\mathbf{1.07}_{(0.02)}$	$1.16_{(0.02)}$	$1.61_{(0.19)}$	$1.36_{(0.08)}$	$1.49_{(0.08)}$	$1.37_{(0.05)}$	$1.29_{(0.04)}$
	ERT (↓)	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$
	WSC (↑)	$0.90_{(0.01)}$	$0.90_{(0.01)}$	$0.91_{(0.01)}$	$0.91_{(0.01)}$	$\mathbf{0.92}_{(0.01)}$	$\mathbf{0.92}_{(0.01)}$	$0.91_{(0.01)}$	$0.89_{(0.01)}$
House	Volume (↓)	$1.09_{(0.04)}$	$\mathbf{1.08}_{(0.04)}$	$1.32_{(0.06)}$	$1.84_{(0.06)}$	$2.71_{(1.71)}$	$2.05_{(0.08)}$	$1.94_{(0.06)}$	$1.80_{(0.05)}$
	ERT (↓)	$0.01_{(0.00)}$	$0.01_{(0.00)}$	$0.02_{(0.00)}$	$0.01_{(0.00)}$	$\mathbf{0.00}_{(0.03)}$	$0.01_{(0.00)}$	$0.01_{(0.00)}$	$0.02_{(0.00)}$
	WSC (↑)	$0.85_{(0.02)}$	$0.85_{(0.01)}$	$0.85_{(0.02)}$	$0.85_{(0.02)}$	$\mathbf{0.89}_{(0.06)}$	$0.85_{(0.02)}$	$0.84_{(0.02)}$	$0.84_{(0.02)}$
Taxi	Volume (↓)	$2.92_{(0.04)}$	$\mathbf{2.91}_{(0.04)}$	$3.17_{(0.05)}$	$3.48_{(0.08)}$	$3.48_{(0.08)}$	$4.08_{(0.09)}$	$3.95_{(0.06)}$	$3.86_{(0.06)}$
	ERT (↓)	$0.01_{(0.00)}$	$0.01_{(0.00)}$	$\mathbf{0.00}_{(0.00)}$	$0.01_{(0.00)}$	$0.01_{(0.00)}$	$0.01_{(0.00)}$	$\mathbf{0.00}_{(0.00)}$	$\mathbf{0.00}_{(0.00)}$
	WSC (↑)	$0.92_{(0.01)}$	$0.92_{(0.01)}$	$0.92_{(0.01)}$	$0.92_{(0.00)}$	$0.92_{(0.01)}$	$\mathbf{0.93}_{(0.01)}$	$0.92_{(0.00)}$	$0.92_{(0.01)}$
House-holds	Volume (↓)	$2.91_{(0.11)}$	$\mathbf{2.89}_{(0.12)}$	$3.19_{(0.12)}$	$3.12_{(0.10)}$	$3.95_{(1.59)}$	$3.59_{(0.17)}$	$3.36_{(0.06)}$	$3.21_{(0.05)}$
	ERT (↓)	$\mathbf{0.00}_{(0.01)}$	$\mathbf{0.00}_{(0.01)}$	$\mathbf{0.00}_{(0.01)}$	$0.01_{(0.01)}$	$\mathbf{0.00}_{(0.02)}$	$0.01_{(0.01)}$	$\mathbf{0.00}_{(0.01)}$	$\mathbf{0.00}_{(0.00)}$
	WSC (↑)	$0.85_{(0.03)}$	$0.84_{(0.03)}$	$0.85_{(0.02)}$	$0.85_{(0.02)}$	$\mathbf{0.89}_{(0.06)}$	$0.87_{(0.02)}$	$0.84_{(0.02)}$	$0.86_{(0.02)}$
RF1	Volume (↓)	$0.35_{(0.02)}$	$\mathbf{0.34}_{(0.02)}$	$0.65_{(0.35)}$	$0.39_{(0.02)}$	$0.44_{(0.07)}$	$0.53_{(0.04)}$	$0.51_{(0.04)}$	$0.42_{(0.02)}$
	ERT (↓)	$0.02_{(0.01)}$	$0.02_{(0.00)}$	$0.02_{(0.00)}$	$0.02_{(0.01)}$	$\mathbf{0.01}_{(0.03)}$	$0.02_{(0.01)}$	$0.02_{(0.01)}$	$0.04_{(0.01)}$
	WSC (↑)	$0.84_{(0.02)}$	$0.84_{(0.02)}$	$0.84_{(0.03)}$	$0.84_{(0.02)}$	$\mathbf{0.87}_{(0.07)}$	$\mathbf{0.87}_{(0.02)}$	$0.85_{(0.02)}$	$0.74_{(0.05)}$
RF2	Volume (↓)	$0.34_{(0.01)}$	$\mathbf{0.33}_{(0.01)}$	$0.61_{(0.21)}$	$0.38_{(0.02)}$	$0.49_{(0.18)}$	$0.51_{(0.04)}$	$0.50_{(0.05)}$	$0.39_{(0.02)}$
	ERT (↓)	$0.02_{(0.01)}$	$0.02_{(0.00)}$	$0.02_{(0.01)}$	$0.02_{(0.01)}$	$\mathbf{0.00}_{(0.03)}$	$0.02_{(0.01)}$	$0.02_{(0.01)}$	$0.03_{(0.01)}$
	WSC (↑)	$0.84_{(0.03)}$	$0.84_{(0.03)}$	$0.85_{(0.03)}$	$0.85_{(0.02)}$	$\mathbf{0.90}_{(0.07)}$	$0.87_{(0.02)}$	$0.85_{(0.02)}$	$0.75_{(0.04)}$

Table 5: Complete results at $\alpha = 0.20$. Entries are mean (standard deviation) over ten splits; subscripts display standard deviations. Boldface marks the best mean in each row.

Dataset	Metric	LECP	Res-LECP	SLS	L-CP	C-HDR	C-PCP	CP2-PCP	PCP
Synthetic datasets									
Comb	Volume (↓)	$0.90_{(0.03)}$	$\mathbf{0.89}_{(0.03)}$	$1.63_{(0.14)}$	$3.39_{(0.10)}$	$3.40_{(0.08)}$	$4.40_{(0.15)}$	$4.34_{(0.15)}$	$3.83_{(0.15)}$
	CC-MAE (↓)	$\mathbf{0.03}_{(0.01)}$	$\mathbf{0.03}_{(0.01)}$	$0.04_{(0.01)}$	$0.12_{(0.01)}$	$0.12_{(0.00)}$	$0.07_{(0.00)}$	$0.07_{(0.00)}$	$0.15_{(0.01)}$
Rings	Volume (↓)	$\mathbf{0.93}_{(0.02)}$	$\mathbf{0.93}_{(0.02)}$	$1.14_{(0.03)}$	$1.28_{(0.02)}$	$1.25_{(0.02)}$	$1.44_{(0.03)}$	$1.43_{(0.03)}$	$1.41_{(0.02)}$
	CC-MAE (↓)	$\mathbf{0.02}_{(0.00)}$	$\mathbf{0.02}_{(0.00)}$	$\mathbf{0.02}_{(0.00)}$	$\mathbf{0.02}_{(0.00)}$	$0.04_{(0.00)}$	$0.04_{(0.00)}$	$0.04_{(0.00)}$	$0.05_{(0.00)}$
Spiral	Volume (↓)	$\mathbf{2.14}_{(0.03)}$	$\mathbf{2.14}_{(0.03)}$	$2.29_{(0.06)}$	$3.05_{(0.18)}$	$2.83_{(0.12)}$	$3.44_{(0.07)}$	$3.42_{(0.08)}$	$3.39_{(0.08)}$
	CC-MAE (↓)	$\mathbf{0.02}_{(0.01)}$	$\mathbf{0.02}_{(0.00)}$	$\mathbf{0.02}_{(0.00)}$	$\mathbf{0.02}_{(0.01)}$	$0.04_{(0.00)}$	$0.04_{(0.00)}$	$0.04_{(0.00)}$	$0.07_{(0.00)}$
Lemnis-cate	Volume (↓)	$0.32_{(0.01)}$	$\mathbf{0.31}_{(0.01)}$	$0.39_{(0.06)}$	$0.49_{(0.02)}$	$0.43_{(0.02)}$	$0.52_{(0.02)}$	$0.51_{(0.02)}$	$0.51_{(0.02)}$
	CC-MAE (↓)	$\mathbf{0.02}_{(0.01)}$	$\mathbf{0.02}_{(0.01)}$	$\mathbf{0.02}_{(0.01)}$	$\mathbf{0.02}_{(0.01)}$	$0.04_{(0.01)}$	$0.04_{(0.00)}$	$0.04_{(0.00)}$	$0.04_{(0.00)}$
Real-world datasets									
Bio	Volume (↓)	$0.66_{(0.01)}$	$0.66_{(0.02)}$	$\mathbf{0.65}_{(0.01)}$	$0.72_{(0.04)}$	$0.66_{(0.02)}$	$0.78_{(0.02)}$	$0.78_{(0.02)}$	$0.73_{(0.02)}$
	ERT (↓)	$\mathbf{0.03}_{(0.00)}$	$0.04_{(0.00)}$	$0.05_{(0.01)}$	$0.04_{(0.01)}$	$0.04_{(0.01)}$	$\mathbf{0.03}_{(0.01)}$	$\mathbf{0.03}_{(0.01)}$	$0.10_{(0.01)}$
	WSC (↑)	$0.70_{(0.01)}$	$0.71_{(0.01)}$	$0.70_{(0.01)}$	$0.71_{(0.02)}$	$\mathbf{0.72}_{(0.02)}$	$\mathbf{0.72}_{(0.02)}$	$0.71_{(0.02)}$	$0.59_{(0.02)}$
House	Volume (↓)	$\mathbf{0.65}_{(0.02)}$	$\mathbf{0.65}_{(0.02)}$	$0.76_{(0.02)}$	$1.14_{(0.04)}$	$1.12_{(0.05)}$	$1.32_{(0.05)}$	$1.31_{(0.05)}$	$1.20_{(0.06)}$
	ERT (↓)	$0.07_{(0.01)}$	$0.07_{(0.01)}$	$0.07_{(0.01)}$	$0.06_{(0.01)}$	$0.06_{(0.01)}$	$\mathbf{0.04}_{(0.01)}$	$\mathbf{0.04}_{(0.01)}$	$0.08_{(0.00)}$
	WSC (↑)	$0.61_{(0.03)}$	$0.61_{(0.03)}$	$\mathbf{0.64}_{(0.03)}$	$0.61_{(0.03)}$	$0.62_{(0.04)}$	$\mathbf{0.64}_{(0.03)}$	$0.63_{(0.03)}$	$0.62_{(0.02)}$
Taxi	Volume (↓)	$2.03_{(0.03)}$	$\mathbf{2.01}_{(0.03)}$	$2.20_{(0.05)}$	$2.35_{(0.03)}$	$2.30_{(0.04)}$	$2.78_{(0.04)}$	$2.76_{(0.03)}$	$2.72_{(0.03)}$
	ERT (↓)	$\mathbf{0.01}_{(0.01)}$	$\mathbf{0.01}_{(0.01)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$\mathbf{0.01}_{(0.00)}$	$0.03_{(0.00)}$
	WSC (↑)	$0.74_{(0.01)}$	$0.74_{(0.01)}$	$0.74_{(0.01)}$	$0.75_{(0.01)}$	$0.75_{(0.01)}$	$\mathbf{0.76}_{(0.01)}$	$0.75_{(0.01)}$	$0.71_{(0.01)}$
House-holds	Volume (↓)	$2.15_{(0.04)}$	$\mathbf{2.14}_{(0.03)}$	$2.28_{(0.05)}$	$2.20_{(0.05)}$	$2.19_{(0.07)}$	$2.51_{(0.05)}$	$2.49_{(0.04)}$	$2.41_{(0.05)}$
	ERT (↓)	$0.03_{(0.01)}$	$0.03_{(0.01)}$	$\mathbf{0.02}_{(0.01)}$	$0.03_{(0.01)}$	$0.03_{(0.01)}$	$\mathbf{0.02}_{(0.01)}$	$\mathbf{0.02}_{(0.01)}$	$0.03_{(0.01)}$
	WSC (↑)	$0.61_{(0.03)}$	$0.60_{(0.03)}$	$\mathbf{0.63}_{(0.02)}$	$0.62_{(0.03)}$	$\mathbf{0.63}_{(0.03)}$	$\mathbf{0.63}_{(0.01)}$	$\mathbf{0.63}_{(0.01)}$	$\mathbf{0.63}_{(0.02)}$
RF1	Volume (↓)	$0.30_{(0.02)}$	$0.28_{(0.01)}$	$0.42_{(0.06)}$	$0.28_{(0.03)}$	$0.28_{(0.03)}$	$0.38_{(0.04)}$	$0.38_{(0.04)}$	$\mathbf{0.27}_{(0.04)}$
	ERT (↓)	$\mathbf{0.07}_{(0.01)}$	$0.08_{(0.01)}$	$0.10_{(0.01)}$	$0.10_{(0.02)}$	$0.10_{(0.02)}$	$\mathbf{0.07}_{(0.01)}$	$\mathbf{0.07}_{(0.02)}$	$0.17_{(0.01)}$
	WSC (↑)	$0.63_{(0.03)}$	$\mathbf{0.64}_{(0.03)}$	$0.62_{(0.02)}$	$0.54_{(0.06)}$	$0.55_{(0.06)}$	$0.61_{(0.03)}$	$0.60_{(0.03)}$	$0.23_{(0.06)}$
RF2	Volume (↓)	$0.30_{(0.01)}$	$0.28_{(0.01)}$	$0.41_{(0.04)}$	$0.27_{(0.01)}$	$0.27_{(0.01)}$	$0.37_{(0.02)}$	$0.37_{(0.02)}$	$\mathbf{0.26}_{(0.01)}$
	ERT (↓)	$0.07_{(0.01)}$	$0.08_{(0.01)}$	$0.10_{(0.01)}$	$0.09_{(0.02)}$	$0.09_{(0.01)}$	$\mathbf{0.06}_{(0.01)}$	$\mathbf{0.06}_{(0.01)}$	$0.17_{(0.01)}$
	WSC (↑)	$0.62_{(0.03)}$	$\mathbf{0.64}_{(0.03)}$	$0.62_{(0.03)}$	$0.56_{(0.06)}$	$0.57_{(0.05)}$	$0.61_{(0.03)}$	$0.61_{(0.03)}$	$0.24_{(0.04)}$

D.2 Sensitivity analysis

We vary one LECP hyperparameter at a time at $\alpha = 0.10$, fixing the rest at their registered values: the Lebesgue reference intensity λ by factors $\{0.1, 1, 10, 100, 1000\}$, temperature β by factors $\{0.25, 0.5, 1, 2, 4\}$, both local proposal standard deviations by factors $\{0.1, 0.25, 0.5, 1, 2\}$, and the Monte Carlo reference count over $\{250, 500, 1000, 2000\}$. Figure 3 plots the volume ratio to the default (blue, left axis) alongside CC-MAE or ERT (orange, right axis) across two synthetic and three real datasets.

The reference intensity λ and temperature β are the most forgiving: sweeping λ over four orders of magnitude or β over a 16-fold range leaves volume nearly unchanged, with conditional-reliability metrics fluctuating within seed-to-seed error bars. The proposal scale is the only hyperparameter that produces a clear failure mode: shrinking both local standard deviations to $0.1\times$ inflates mean volume by 89%–257% across datasets, as the overly narrow proposals under-explore density tails and force the conformal threshold upward to maintain marginal coverage; widening to $2\times$ incurs no comparable penalty. Increasing the Monte Carlo count from $1{,}000$ to $2{,}000$ yields volume reductions that are small and often within noise, while roughly doubling the sampling cost. Overall, LECP is robust to moderate perturbations of these four hyperparameters.

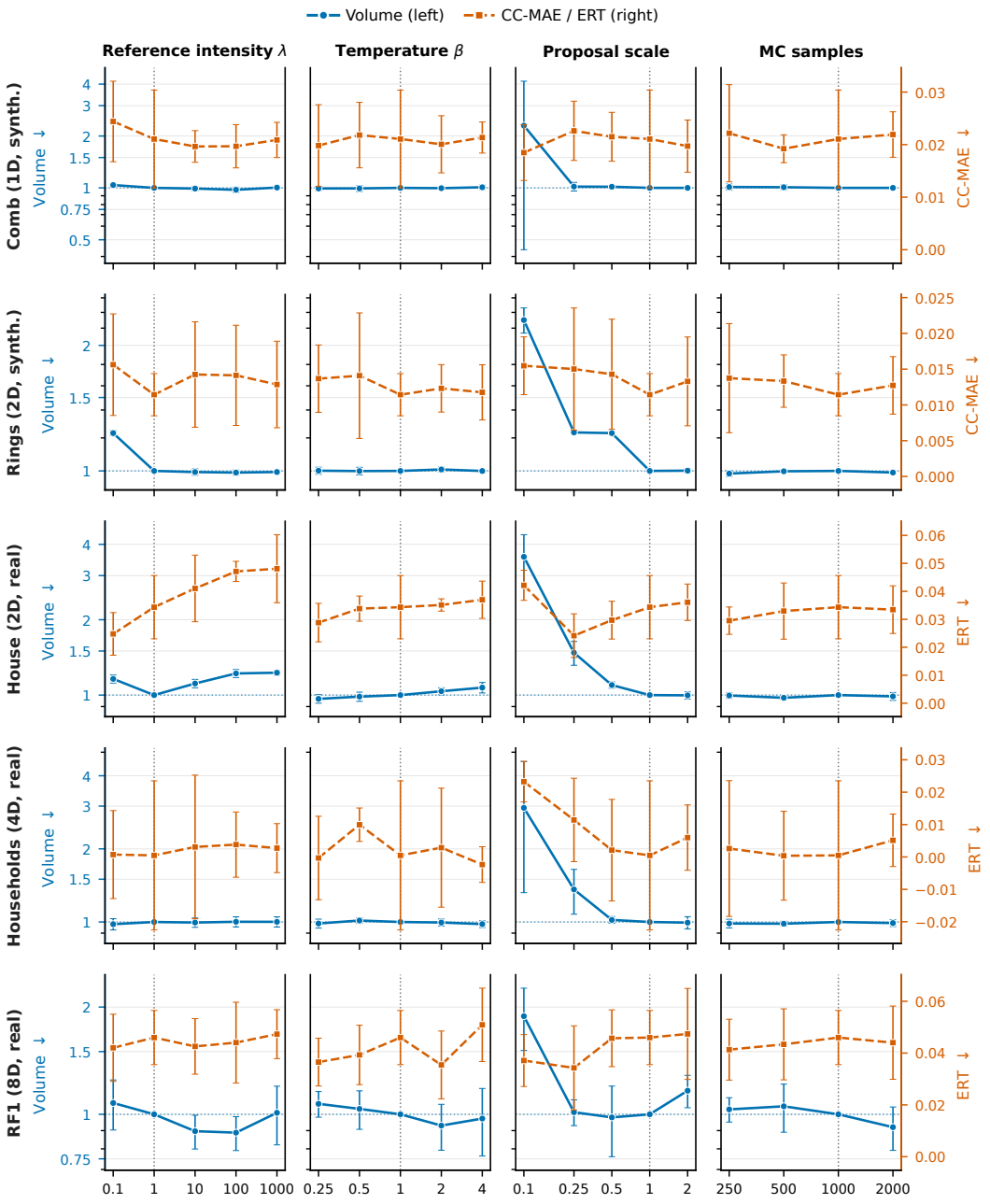


Figure 3: One-factor sensitivity of LECP at $\alpha = 0.10$. Rows correspond to datasets; each column varies one of λ, β, local proposal standard deviation, and Monte Carlo reference count. Blue curves show the volume ratio to the used default (left axis); orange curves show CC-MAE (synthetic) or ERT (real) on the right axis. Points and error bars are five-seed means and 95% t-intervals; dotted vertical lines mark the default values.

D.3 Mechanism ablations

The Monte Carlo estimator (5) couples two mechanisms: *proposal locality* (centering reference draws at each training response Y_i rather than at a fixed location) and *importance correction* (reweighting each draw by $\lambda/q_i(y)$ to ensure a Lebesgue effective reference measure). We dis-

entangle their contributions through a structured ablation on five datasets spanning $d \in \{1, 2, 4, 8\}$ at $\alpha = 0.10$, keeping the rest of the pipeline (Algorithm 1) unchanged.

Proposals. Let $\widehat{\Sigma} = \mathrm{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_d^2)$ denote the diagonal training-response covariance and $\bar{Y} = n^{-1}\sum_i Y_i$ the training-response mean. The *local* and *global* proposals are the two-component Gaussian mixtures

$$q_i^{\mathrm{loc}}(y) = \tfrac{1}{2}\,\mathcal{N}\big(y;\, Y_i,\, \tfrac{1}{4}\widehat{\Sigma}\big) + \tfrac{1}{2}\,\mathcal{N}\big(y;\, Y_i,\, \widehat{\Sigma}\big), \tag{63}$$

$$q^{\mathrm{glob}}(y) = \tfrac{1}{2}\,\mathcal{N}\big(y;\, \bar{Y},\, \tfrac{1}{4}\widehat{\Sigma}\big) + \tfrac{1}{2}\,\mathcal{N}\big(y;\, \bar{Y},\, \widehat{\Sigma}\big), \tag{64}$$

where $\mathcal{N}(\cdot\,;\,\mu, \Sigma)$ is the Gaussian density with mean μ and covariance Σ. Both have full $\mathbb{R}^d$ support; they differ only in whether the center adapts to each training observation. We construct four ablation variants by removing one or both mechanisms from the default, as summarized in Table 6.

Table 6: Ablation variants. Each row specifies the proposal q, per-draw weight ω, post-hoc score correction h, and the Bayes-optimal score target (up to an additive function of x); $p \equiv p(y \mid x)$. Rows whose target equals $-\beta \log p$ preserve the correct density level-set geometry.

Variant	Proposal q	Weight $\omega(y)$	Correction $h(y)$	Population target
LECP	q_i^{loc}	$\lambda / q_i^{\mathrm{loc}}(y)$	—	$-\beta \log p$
$-$IS	q_i^{loc}	1	—	$-\beta \log p + \beta \log \bar{q}^{\mathrm{loc}\,\dagger}$
$-$Local	q^{glob}	$\lambda / q^{\mathrm{glob}}(y)$	—	$-\beta \log p$
$-$Local $-$IS	q^{glob}	1	—	$-\beta \log p + \beta \log q^{\mathrm{glob}}$
$-$Local $-$IS $+$Corr	q^{glob}	1	$-\beta \log q^{\mathrm{glob}}(y)$	$-\beta \log p$

† $\bar{q}^{\mathrm{loc}}(y \mid x) := \mathbb{E}\big[q_i^{\mathrm{loc}}(y) \mid X_i = x\big]$ depends on the unknown conditional law and admits no tractable post-hoc correction.

Results. This ablation study highlights three key properties of the proposed architecture (Figure 4).

IS correction is the dominant mechanism. Dropping both mechanisms ($-$Local $-$IS) inflates volume by 60–95% across all five datasets (all t-intervals strictly excluding unity). The post-hoc score correction ($-$Local $-$IS $+$Corr) recovers much of this gap for $d \leq 2$ but retains 40–50% excess on House and RF1, confirming that a distorted training objective is only partially repaired at inference.

Locality becomes critical in higher dimensions. Removing locality alone ($-$Local) matches the default for $d \leq 4$ but incurs 40–50% volume inflation on RF1 ($d = 8$), where fixed-center draws grow sparse relative to the conditional support. Removing only the IS correction ($-$IS) imposes a milder but steadily increasing penalty of 5–15% excess volume on the real benchmarks.

Conditional reliability remains stable. CC-MAE and ERT stay within seed-to-seed uncertainty for nearly all variants. Cases where an ablation achieves marginally lower ERT coincide with substantially larger volume, reflecting conservative over-coverage rather than improved calibration.

D.4 Computational cost

Figure 5 reports mean training time and full-test inference time at $\alpha = 0.10$ over ten splits. Training time excludes the final conformal calibration step; inference time measures prediction-set membership evaluation on the complete test split.

Training time. The five baselines built on top of a conditional normalizing flow—L-CP, C-HDR, C-PCP, CP2-PCP, and PCP—share a single MAF fit, so their identical columns reflect one training run. LECP is 2.4–17.7$\times$ slower than this shared MAF because it alternates between two models (a score network and a level network) and, at each alternation, draws a fresh batch of Monte Carlo reference samples for every training pair; additional overhead comes from estimating the self-calibrated mean volume on the validation set at numerous preset checkpoints via Monte Carlo integration. SLS is the most expensive method overall (3.3–4.4$\times$ slower than LECP), as it likewise alternates training but fits a frontier model comprising three normalizing flows.

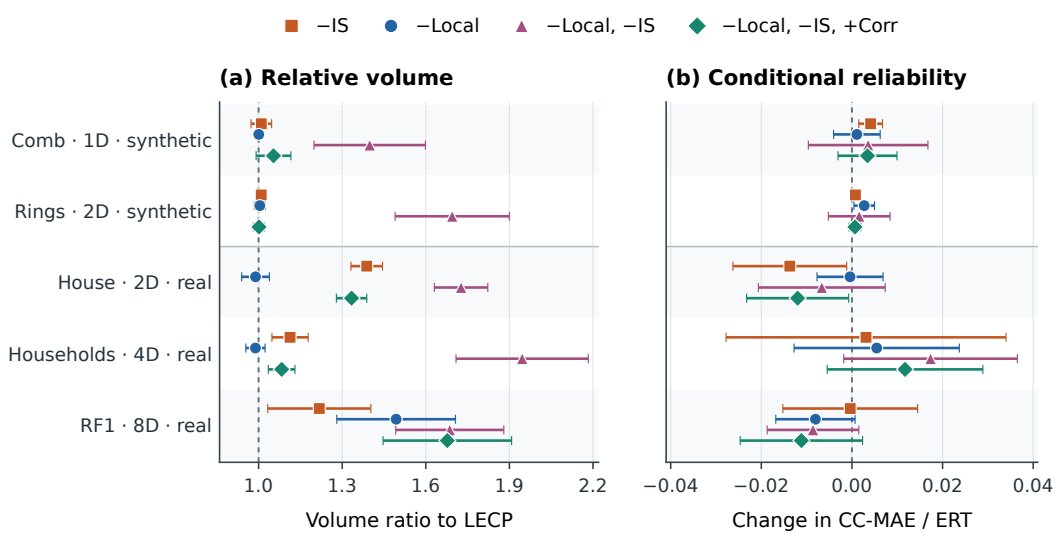


Figure 4: Mechanism ablations at $\alpha = 0.10$. **(a)** Paired volume ratio to default LECP (dotted line at 1); values above 1 indicate efficiency loss. **(b)** Paired change in CC-MAE (synthetic) or ERT (real); values near zero indicate preserved conditional reliability. Points and bars are five-seed means and 95% t-intervals. Row labels indicate dataset name, response dimension, and type.

Inference time. At test time the cost ranking reverses. Flow-based baselines require forward or inverse passes through the normalizing flow to map between the latent and original response spaces; PCP-family methods additionally draw large sets of conditional samples for each test point, compounding this cost. SLS incurs both flow evaluations and frontier-set computations, making it 12.6–33.1× slower than LECP. By contrast, LECP evaluates only two scalar-output networks (score and level) in a single forward pass, achieving 5.7–11.6× faster inference than the next-cheapest baseline (L-CP) across all ten datasets.

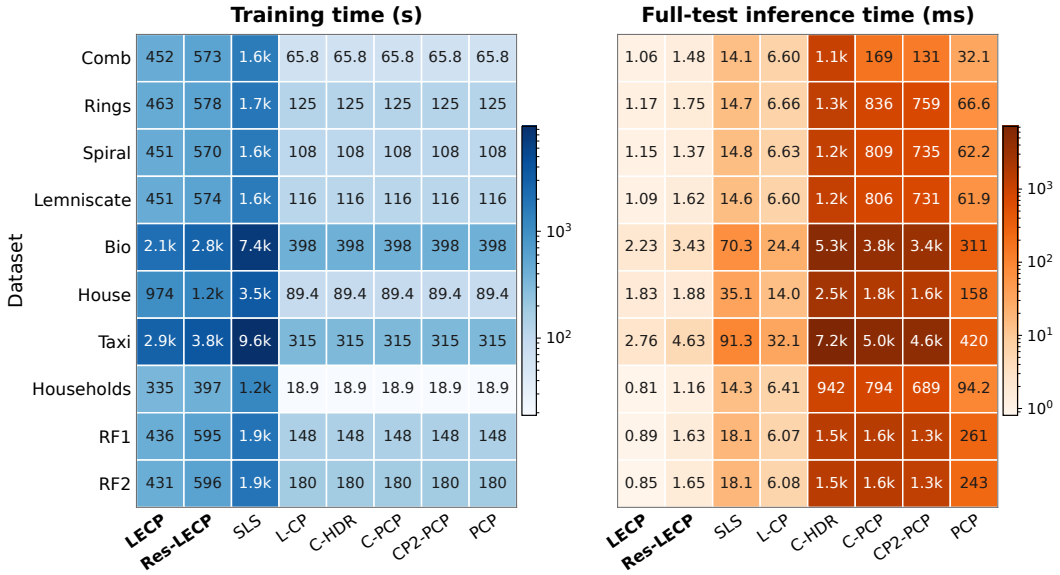


Figure 5: Mean training time and full-test inference time at $\alpha = 0.10$ over ten splits. Colors use a logarithmic scale; cell annotations report mean seconds for training and milliseconds for inference.

D.5 Residual refinement under conditional scale stress

The main benchmarks show only modest differences between LECP and Res-LECP because the base contrastive score already captures most of the conditional geometry. To isolate the regime where residual refinement becomes essential, we construct a stress test that amplifies covariate-dependent scale while holding the latent conditional shape fixed.

Setup. Let $Z \in \mathbb{R}^2$ follow the heteroskedastic spiral of the main Spiral benchmark (Appendix C), conditionally on a rescaled covariate $U(X) = 0.1 + 0.9(X - 0.25)/0.75$ with $X \sim \mathrm{Unif}[0.25, 1]$. The observed response is

$$Y = X^\gamma Z, \qquad \gamma \in \{0, 1, 2, 3, 4\}. \tag{65}$$

Increasing γ amplifies the covariate-dependent scale without altering the latent conditional shape. For each of ten seeds the same (X, Z) draws are reused across all γ values, so only the deterministic scaling differs. We generate $n = 20{,}000$ observations per seed, apply the standard $60/10/10/20\%$ split, and fit LECP and Res-LECP ($B = 2$) at $\alpha = 0.10$ with default hyperparameters; the LECP result is the frozen base model inside the corresponding Res-LECP fit. The analytic conditional density provides the oracle 90% HDR $H(x)$, enabling evaluation of HDR IoU:

$$\mathrm{IoU} = \frac{\mu(C \cap H)}{\mu(C \cup H)},$$

relative volume $\mu(C)/\mu(H)$ (optimum one), and CC-MAE. All set operations use the invertible latent coordinates $z = y/x^\gamma$.

Results. At $\gamma = 0$ the two methods are indistinguishable (IoU ≈ 0.87, relative volume ≈ 1.08; Figure 6). As γ grows, LECP's IoU drops to 0.48 and its relative volume inflates to $3.2\times$ at $\gamma = 4$, whereas Res-LECP maintains IoU above 0.71 and relative volume below $1.7\times$ throughout, with a peak paired IoU gain of $+0.28$ at $\gamma = 3$. CC-MAE remains comparable, confirming that the improvement is in geometric efficiency rather than at the expense of conditional reliability. The conditional slices of Figure 7 illustrate the mechanism: at $\gamma \geq 2$, LECP's regions extend well beyond the conditional support, while Res-LECP's boundaries remain close to the oracle contour. As described in Section 3.4, boundary-focused weighting concentrates estimation effort on the level-set contour that determines volume, and the x-dependent offset from the reference normalization is absorbed by the subsequent level refit.

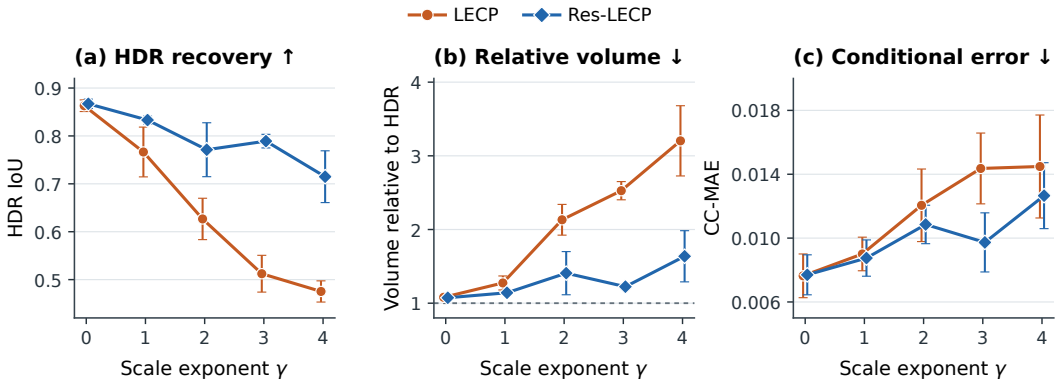


Figure 6: Performance under increasing conditional-scale stress ($Y = X^\gamma Z$, $\alpha = 0.10$). **(a)** Calibrated IoU with the analytic 90% conditional HDR (higher is better). **(b)** Prediction-set volume relative to the HDR volume; one is optimal (lower is better). **(c)** CC-MAE (lower is better). Points and error bars are means and 95% t-intervals over ten paired seeds. Res-LECP increasingly improves HDR recovery and efficiency as γ grows, while conditional reliability remains comparable.

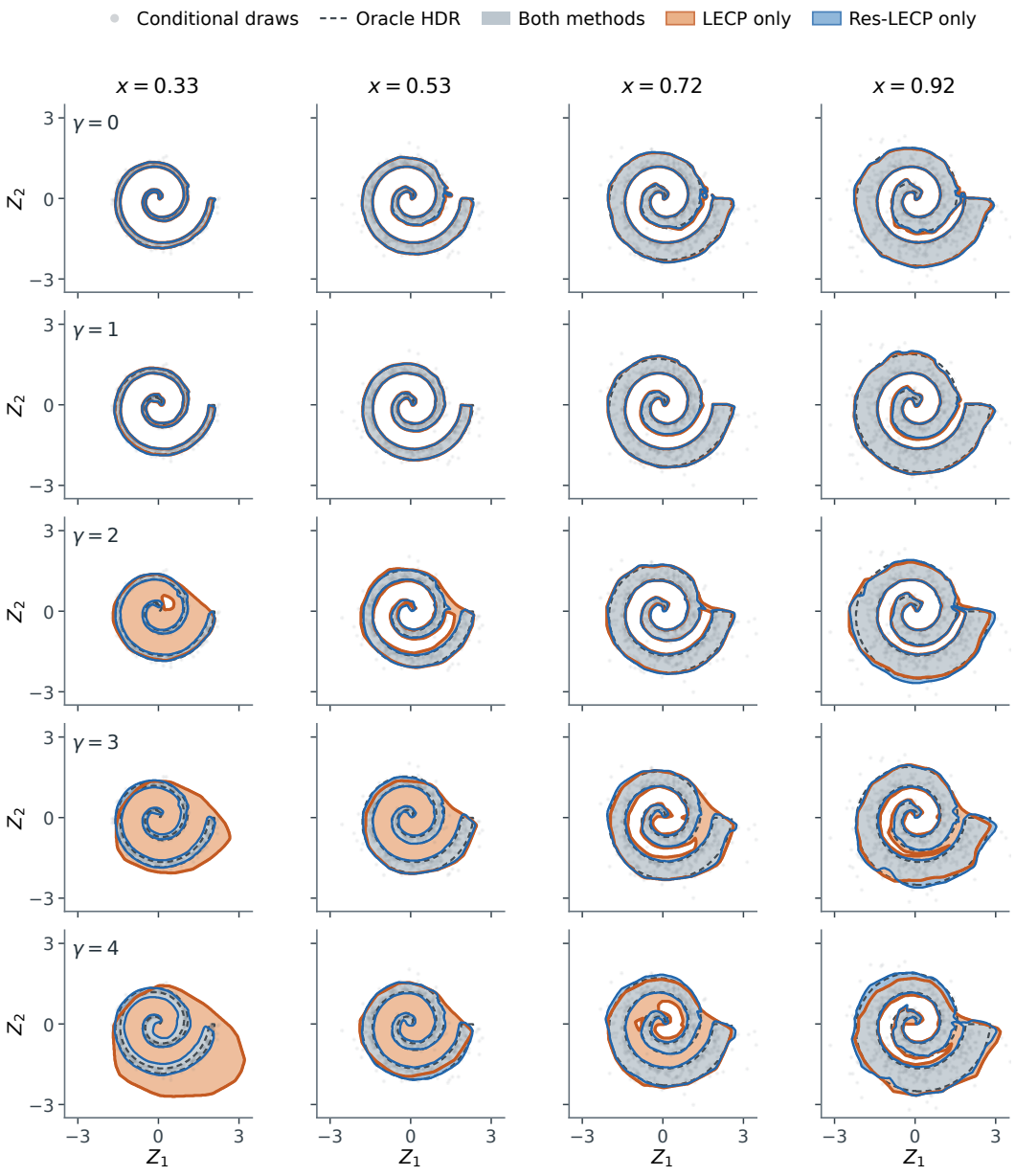


Figure 7: Representative calibrated prediction regions under conditional-scale stress, displayed in latent coordinates $z = y/x^{\gamma}$. Rows index γ; columns index the conditioning value x. Gray points are conditional draws; the dashed contour is the analytic 90% conditional HDR. Gray-blue fill denotes regions included by both methods; orange and blue fills denote LECP-only and Res-LECP-only regions, respectively. At $\gamma = 0$ both methods track the oracle closely; at larger γ, LECP over-covers substantially while Res-LECP remains compact. One representative seed is shown.