Co-optimization for Adaptive Conformal Prediction

Xiaoyi Su[1] **Zhixin Zhou**[2] **Rui Luo**[1]
[1]Department of Systems Engineering, City University of Hong Kong
[2]Alpha Benito Research
xiaoyisu3-c@my.cityu.edu.hk zzhou@alphabenito.com ruiluo@cityu.edu.hk

Abstract

In regression, conformal prediction often suffers from inefficiency under heteroscedasticity and skewness due to fixed, non-adaptive interval centering. We propose CoCP, a new method that parameterizes prediction intervals by a center $m(x)$ and a radius $h(x)$, and alternates between learning $h(x)$ via smooth quantile regression on folded residuals and refining $m(x)$ using a smooth interval loss. This corrects mis-centering and drives the interval toward high-density regions. Beyond finite-sample marginal validity via split-conformal calibration, we prove that CoCP also attains asymptotic conditional coverage and optimal interval length if the base estimators are consistent. Experiments demonstrate that CoCP yields tighter intervals and achieves improved empirical conditional-coverage reliability. Code is available at `https://github.com/jjoeysu/cocp`.

1 Introduction

Even though conformal prediction (CP) [47, 43, 30, 2] provides a distribution-free recipe for constructing prediction sets with finite-sample marginal coverage under exchangeability, valid prediction intervals should further be *as short as the conditional distribution permits* while adapting to heteroscedasticity and shape changes. Shorter intervals carry direct downstream value, e.g., tightening the pessimism gap in robust predict-then-optimize [34] and sharpening escalation decisions in selective regression [26]. In this context, Conformalized Quantile Regression (CQR) [39, 41] is a popular instantiation combining quantile estimation with CP to produce adaptive, marginally valid intervals.

The theoretical benchmark for interval efficiency is the highest density interval (HDI). By the Neyman–Pearson lemma, the shortest $(1-\alpha)$-region corresponds to a density level set $\{y : f(y) \geq \lambda\}$, which reduces to the contiguous HDI under unimodality [7, 24]. For interior solutions, its endpoints satisfy an equal-density boundary condition. However, standard methods like CQR enforce equal-tailed error rates (e.g., $\alpha/2$ on each side). Under skewed distributions, this rigid constraint systematically displaces the interval from high-density regions, yielding unnecessarily long intervals even when conditional coverage is maintained (Figure 1a). Thus, an optimally efficient interval must not only scale to accommodate local noise but also translate its center toward concentrated probability mass.

The core insight of this paper is a geometric pathway from any interval to a HDI-like interval, conceptualized via the folded distribution visualization in Figure 1. Consider an interval parameterized by a center $m(x)$ and a radius $h(x)$. For a fixed center m, the radius required to capture $1-\alpha$ conditional mass is determined by a single boundary quantity: the $(1-\alpha)$-quantile of the folded residual $|Y - m(X)| \mid X = x$. If the conditional densities at the two corresponding endpoints are imbalanced, translating m slightly toward the higher-density side shifts the data distribution relative to the center. Visually, this relative motion *pushes* denser probability mass into the interval while *pulling* sparser mass out (Figure 1c). Because a net positive mass enters the interval, the boundary quantile—and thus the radius h—must contract to maintain the exact $1-\alpha$ mass, thereby shortening the interval. This "push–pull" dynamic continues until the endpoint densities balance, which, under

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

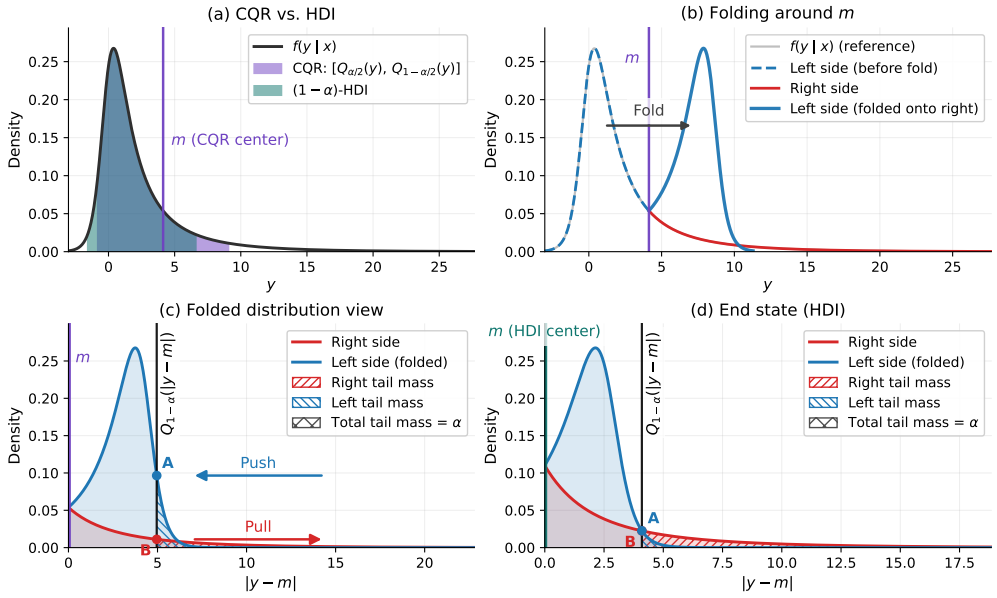

Figure 1: From any interval to HDI via folding and boundary balancing. (a) Under skewness, equal-tailed intervals can be displaced relative to the HDI and are typically longer. (b) Folding around a candidate center m maps y to $|y - m|$, producing a two-layer view. (c) The folded $(1 - \alpha)$-boundary corresponds to two endpoints; translating m shortens the interval when the endpoint densities are imbalanced (push–pull). (d) At equilibrium the endpoint densities match, recovering the HDI.

unimodality, perfectly recovers the HDI (Figure 1d). Thus, the discrepancy between any interval and HDI-like interval can be elegantly framed as a *boundary-balancing* problem in a folded geometry.

We operationalize this principle with **CoCP** (**Co**-optimization for adaptive **C**onformal **P**rediction). CoCP explicitly couples the learning of $m(x)$ and $h(x)$, enabling the interval to simultaneously translate and scale. Specifically, h tracks the folded $(1 - \alpha)$-boundary via a smooth pinball loss, while m is refined using a differentiable smooth interval loss. The surrogate concentrates the learning signal near the interval endpoints (Figure 2), directly capturing the boundary imbalance. Finally, we apply a standard conformal calibration step to guarantee finite-sample marginal coverage. Therefore, the co-optimization stage can be viewed as an interval-oriented conformalized training, learning adaptive interval geometry, while calibration ensures marginal validity.

Our main contributions are summarized as follows: (1) **Folded geometry perspective.** We introduce a boundary-balancing viewpoint that explains the inefficiency of center-invariant intervals under skewness, motivating the co-optimization of translation and scaling. (2) **CoCP framework.** We propose a differentiable alternating optimization method. Beyond distribution-free finite-sample marginal validity, we show that CoCP asymptotically recovers the optimal interval length and achieves asymptotic conditional coverage under the stated regularity and consistency conditions. (3) **Empirical performance.** Across diverse benchmarks, CoCP yields tighter intervals and consistently improves empirical conditional reliability over the evaluated baselines.

2 Related Work

Standard CQR [39, 41] yields heteroscedastic but equal-tailed intervals. To better adapt to skewed or multimodal data, distributional methods [10, 25] extract high-density sets from learned conditional distributions. For instance, Conformal Histogram Regression (CHR) [42] and Conformal Interquantile Regression (CIR) [20] conformalize histograms or interquantile partitions, while generative models like C-HDR [25, 13] capture multimodality. Other approaches optimize restricted set classes [16] or Conformal Thresholded Intervals (CTI) [32]. However, as noted in [4], these methods often require estimating the full conditional density or CDF, rely on discretization, and may produce non-contiguous or unstable sets. Direct neural interval construction using soft-coverage heuristics

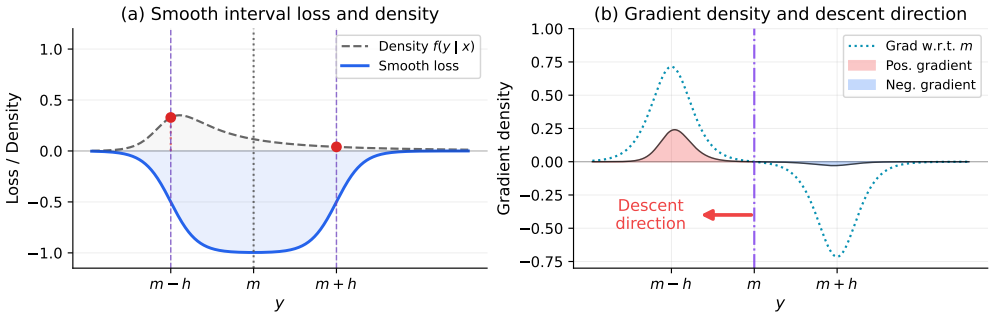

Figure 2: Gradient dynamics of the smooth interval loss. (a) The smooth interval loss serves as a differentiable surrogate for the negative interval indicator. The underlying data density (dashed line) is imbalanced at the two boundaries. (b) The gradient of the smooth loss with respect to the center m acts as a signed sampling kernel, being positive on the left boundary and negative on the right. The resulting gradient density (the product of the conditional density and the derivative kernel) is asymmetric under skewness; the higher-density boundary generates a dominant directional signal for gradient descent, shifting the center m toward the concentration of probability mass.

[35] avoids full density estimation but relies on brittle Lagrangian penalties. In contrast, CoCP bypasses full density estimation by embedding soft-coverage within a principled folded geometry, using boundary-local gradients to autonomously center the interval and minimize the required radius.

Beyond marginal validity, achieving exact conditional coverage is fundamentally limited without strong assumptions [29, 15]. Consequently, recent work focuses on rigorous diagnostics, such as evaluating the Excess Risk of Target coverage (ERT) [5], and principled relaxations via localization [45, 19, 23] or group fairness [3]. Other approaches calibrate against specific function classes like reproducing kernel Hilbert spaces (RKHS) [18], traced efficiently by SpeedCP [31], and extended by Conformal Prediction with Length-Optimization (CPL) [27]. Alternatively, methods like Colorful Pinball Conformal Prediction (CPCP) [9] and others [28] directly minimize conditional coverage errors. CoCP provides a distinct geometric solution: it improves conditional reliability not through complex reweighting or functional constraints, but by explicitly co-optimizing the interval's translation and scaling to adapt to local density imbalances.

In addition, several methods improve local adaptivity by transforming conformity scores. Techniques like Rectified Conformal Prediction (RCP) [36], adaptive split-Jackknife+ [12], and locally trained predictors [11] refine the scale of uncertainty. Other works improve statistical efficiency via cross-conformal data reuse [46, 17]. However, these score-adaptation approaches typically inherit a fixed interval center from a base predictor, adjusting only the scale and failing to correct mis-centering under skewness. CoCP fundamentally differs by abandoning the fixed-center paradigm, instead alternating between radius scaling and center translation to achieve boundary-balancing optimality.

3 The Geometry of Efficient Intervals and Lyapunov Dynamics

This section establishes the geometric foundation of CoCP. We demonstrate how actively translating an interval's center minimizes its required radius to achieve the HDI lower bound, and show that this dynamic forms a monotonically descending Lyapunov system.

Let $(X, Y) \sim P$ with $X \sim P_X$, $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $Y \in \mathbb{R}$. Let $f(y \mid x)$ denote the conditional density of $Y \mid X = x$. For a unimodal and regular $f(y \mid x)$, the shortest $(1-\alpha)$-coverage interval is the unique HDI. In center-radius form, $C^*(x) = [m^*(x) - h^*(x), m^*(x) + h^*(x)]$, which is uniquely characterized by two simultaneous conditions:

1. **Coverage condition:** $\int_{m^*-h^*}^{m^*+h^*} f(y \mid x)\, dy = 1 - \alpha$.
2. **Density-equality condition:** $f(m^* - h^* \mid x) = f(m^* + h^* \mid x)$.

Logistic smoothing. Directly optimizing this geometric objective via gradient descent is intractable due to the zero gradients produced by the hard indicator $\mathbf{1}\{|Y - m| \le h\}$. To overcome this, we

introduce a logistic smoothing. Define the symmetric logistic kernel $K_\beta(u) \coloneqq \frac{1}{\beta}\sigma'(u/\beta)$, where $\sigma(z) \coloneqq (1+e^{-z})^{-1}$ and $\beta > 0$. The smoothed conditional density is defined as the convolution:

$$f_\beta(y \mid x) \coloneqq (f(\cdot \mid x) * K_\beta)(y) = \int_{\mathbb{R}} f(u \mid x)\, K_\beta(y-u)\, du. \tag{1}$$

Proposition 1. *Fix $x \in \mathcal{X}$. Suppose $f(\cdot \mid x)$ is a bounded unimodal probability density on $\mathbb{R}$. Then $f_\beta(\cdot \mid x) \in C^1(\mathbb{R})$, and $f'_\beta(\cdot \mid x)$ has exactly one real zero $\nu_\beta(x)$. Moreover, $f'_\beta(y \mid x) > 0$ for $y < \nu_\beta(x)$, and $f'_\beta(y \mid x) < 0$ for $y > \nu_\beta(x)$. In particular, $f_\beta(\cdot \mid x)$ is strictly unimodal.*

For a candidate (m, h), define the smoothed coverage function $\Gamma(m, h; x) \coloneqq \int_{m-h}^{m+h} f_\beta(y \mid x)\, dy$. Since $f_\beta > 0$, Γ is strictly monotonically increasing with respect to h. Thus, the minimal feasible radius $\phi_\beta(m; x) \coloneqq \inf\{h > 0 : \Gamma(m, h; x) \geq 1-\alpha\}$ is achieved at: $\Gamma(m, \phi_\beta; x) = 1-\alpha$.

Translation dynamics and the Lyapunov property. Applying the implicit function theorem to $\Gamma(m, \phi_\beta; x) = 1-\alpha$, we derive the exact gradient of the required radius with respect to the center:

$$\frac{\partial \phi_\beta(m; x)}{\partial m} = -\frac{\frac{\partial \Gamma}{\partial m}}{\frac{\partial \Gamma}{\partial \phi_\beta}} = -\frac{f_\beta(m+\phi_\beta \mid x) - f_\beta(m-\phi_\beta \mid x)}{f_\beta(m+\phi_\beta \mid x) + f_\beta(m-\phi_\beta \mid x)}. \tag{2}$$

Note that it depends only on the density at the interval endpoints. This reveals an elegant *push–pull* dynamic: shifting the center m toward the denser endpoint strictly decreases the required radius.

4 Co-optimizing Interval Geometry via Alternating Learning

CoCP implements this principle through differentiable alternating optimization between the center $m(x)$ and the radius $h(x)$ in Algorithm 1. This creates a duality between HDI statistical properties and the functional optimization landscape.

Radius update (Coverage). Given a fixed center m, we update h to estimate the smoothed $(1-\alpha)$-quantile of the absolute residual by optimizing the *smooth pinball loss*:

$$\mathcal{L}_{\mathrm{H}}(h; m, \beta) \coloneqq \mathbb{E}\left[\beta \sum_{s\in\{1,-1\}} \zeta\left(\frac{s\cdot h(X) - (Y - m(X))}{\beta}\right) - (1-\alpha)h(X)\right], \tag{3}$$

where $\zeta(z) = \log(1+e^z)$. In the unrestricted population limit, the functional first-order condition (FOC) $\partial\mathcal{L}_{\mathrm{H}}/\partial h = 0$ exactly recovers $\Gamma(m, h; X) = 1-\alpha$, fulfilling the first defining property of the HDI. Furthermore, because $\zeta(z)$ is globally strictly convex, $\mathcal{L}_{\mathrm{H}}$ is globally strictly convex in h.

Center update (Density-equality). Let $\tilde{\mathbf{1}}_\beta(y; m, h) \coloneqq \sigma((h-(y-m))/\beta) - \sigma((-h-(y-m))/\beta)$ be the logistic-smoothed interval indicator. To correct mis-centering, we fix h and optimize the *smooth interval loss* (the negative expected smoothed coverage, unified with $\mathcal{L}_{\mathrm{H}}$ via $\zeta' = \sigma$):

$$\mathcal{L}_{\mathrm{M}}(m; h, \beta) \coloneqq -\mathbb{E}\left[\tilde{\mathbf{1}}_\beta(Y; m(X), h(X))\right]. \tag{4}$$

The functional gradient evaluates exactly to the boundary density difference: $\partial\mathcal{L}_{\mathrm{M}}/\partial m = -\big(f_\beta(m+h \mid X) - f_\beta(m-h \mid X)\big)$. Setting $\partial\mathcal{L}_{\mathrm{M}}/\partial m = 0$ perfectly recovers the second defining property of the HDI. Furthermore, $\partial^2\mathcal{L}_{\mathrm{M}}/\partial m^2 = f'_\beta(m-h \mid X) - f'_\beta(m+h \mid X)$. By Proposition 1, if the interval contains the mode in its interior, then $f'_\beta(m-h \mid X) > 0$, $f'_\beta(m+h \mid X) < 0$, making $\partial^2\mathcal{L}_{\mathrm{M}}/\partial m^2 > 0$ and $\mathcal{L}_{\mathrm{M}}$ locally strictly convex in the population limit.

Idealized Lyapunov dynamics. Equations (3) and (4) form a tightly coupled dynamical system. Comparing $\partial\mathcal{L}_{\mathrm{M}}/\partial m$ with the implicit derivative in (2), we can formalize a population dynamic.

Proposition 2 (Idealized Lyapunov descent)**.** *Fix $x \in \mathcal{X}$. Suppose the radius is instantaneously equilibrated to $h_t = \phi_\beta(m_t; x)$, and the center follows the population gradient flow $\frac{dm_t}{dt} = -\frac{\partial\mathcal{L}_{\mathrm{M}}}{\partial m}(m_t; h_t, \beta)$. Then the required radius is a Lyapunov function:*

$$\frac{d}{dt}\phi_\beta(m_t; x) = -\frac{\big[f_\beta(m_t+h_t \mid x) - f_\beta(m_t-h_t \mid x)\big]^2}{f_\beta(m_t+h_t \mid x) + f_\beta(m_t-h_t \mid x)} \leq 0, \tag{5}$$

with equality if and only if the density-equality condition is met.

This property provides the geometric intuition for Algorithm 1: minimizing the center loss autonomously drives the interval toward density balance, intrinsically minimizing the required radius.

Final radius fine-tuning and calibration. After convergence, we perform a final fine-tuning of h to remove the logistic smoothing bias by minimizing the standard hard pinball loss $\rho_\tau(u) := u(\tau - \mathbf{1}\{u < 0\})$ on the exact folded residual:

$$\mathcal{L}_{\mathrm{PB}}(h; m) \coloneqq \mathbb{E}\left[\rho_{1-\alpha}\left(|Y - m(X)| - h(X)\right)\right]. \tag{6}$$

Crucially, this step decouples the two objectives: the alternating optimization dictates the center m to minimize interval length, while this final fine-tuning accurately targets the true conditional quantile to ensure reliable conditional coverage. Finally, we aggregate the K-fold ensembled predictors $\widehat{m}$ and $\widehat{h}$, and calibrate them on an independent set $\mathcal{D}_{\mathrm{cal}}$ to guarantee finite-sample marginal validity.

Algorithm 1 CoCP: Co-optimization for Adaptive Conformal Prediction

Require: Training data $\mathcal{D}_{\mathrm{train}}$, calibration data $\mathcal{D}_{\mathrm{cal}}$, miscoverage $\alpha \in (0, 1)$, folds K, iterations T, temperature $\beta > 0$

Ensure: Calibrated interval function $\widehat{C}(\cdot)$

Partition $\mathcal{D}_{\mathrm{train}}$ into K disjoint folds $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$
for $k = 1, \ldots, K$ **do**
  Set $\mathcal{D}_h^{(k)} = \mathcal{D}_k$ and $\mathcal{D}_m^{(k)} = \mathcal{D}_{\mathrm{train}} \setminus \mathcal{D}_k$
  Warm-start $\widehat{m}^{(k)}$ on $\mathcal{D}_m^{(k)}$ by minimizing Mean Squared Error (MSE)
  for $t = 1, \ldots, T$ **do**
    Radius update: fix $\widehat{m}^{(k)}$, train $\widehat{h}^{(k)}$ on $\mathcal{D}_h^{(k)}$ via smooth pinball loss (3)
    Center update: fix $\widehat{h}^{(k)}$, train $\widehat{m}^{(k)}$ on $\mathcal{D}_m^{(k)}$ via smooth interval loss (4)
  end for
  Fine-tune $\widehat{h}^{(k)}$ on $\mathcal{D}_h^{(k)}$ by pinball loss (6) with $\widehat{m}^{(k)}$ fixed
end for
Form ensemble predictors $\widehat{m}, \widehat{h}$ via averaging: $\widehat{m} = \frac{1}{K}\sum_{k=1}^{K} \widehat{m}^{(k)}$, $\widehat{h} = \frac{1}{K}\sum_{k=1}^{K} \widehat{h}^{(k)}$
Compute nonconformity scores $S_i = |Y_i - \widehat{m}(X_i)|/\widehat{h}(X_i)$ for all $(X_i, Y_i) \in \mathcal{D}_{\mathrm{cal}}$
Set $\widehat{q}$ as the $\lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$-th smallest score in $\{S_i\}_{i=1}^{n_{\mathrm{cal}}} \cup \{+\infty\}$
return $\widehat{C}(x) = [\widehat{m}(x) - \widehat{q}\widehat{h}(x), \widehat{m}(x) + \widehat{q}\widehat{h}(x)]$

5 Theoretical Properties

We establish three layers of guarantees for CoCP. First, we show that the algorithm inherits distribution-free *finite-sample marginal coverage*. Second, we characterize the exact population fixed point of the alternating updates. Finally, we prove that under standard regularity and consistency conditions, CoCP achieves *asymptotic conditional coverage* [30, 41], together with *asymptotic optimal interval length* for unimodal distributions. Full proofs are deferred to Appendix A.

Theorem 1 (Finite-sample marginal coverage)**.** *Let $\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{cal}}+1}$ be exchangeable, where the first n_{cal} pairs form $\mathcal{D}_{\mathrm{cal}}$ and $(X_{n_{\mathrm{cal}}+1}, Y_{n_{\mathrm{cal}}+1})$ is a new test point. Suppose $\widehat{m}$ and $\widehat{h}$ are constructed via Algorithm 1 with $\widehat{h}(x) > 0$ a.s.. Then, for every $\alpha \in (0, 1)$,*

$$1 - \alpha \;\le\; \mathbb{P}\big(Y_{n_{\mathrm{cal}}+1} \in \widehat{C}(X_{n_{\mathrm{cal}}+1})\big) \;\le\; 1 - \alpha + \frac{1}{n_{\mathrm{cal}} + 1}, \tag{7}$$

where the upper bound additionally requires almost-sure distinctness of the $n_{\mathrm{cal}} + 1$ scores.

5.1 The Smoothed HDI as Population Target

The efficiency and conditional coverage depend on the fixed point of the CoCP updates. The first-order conditions of CoCP's losses admit an exact interpretation: the radius FOC $\nabla_h \mathcal{L}_H = 0$ implies $h = \phi_\beta(m; x)$, and the center FOC $\nabla_m \mathcal{L}_M = 0$ enforces $f_\beta(m + h \mid x) = f_\beta(m - h \mid x)$. Their joint solution defines the *smoothed HDI*. Crucially, as the temperature β decays, this smoothed target asymptotically recovers the true, non-differentiable HDI.

Definition 1 (Smoothed HDI)**.** Fix $\beta > 0$ and $x \in \mathcal{X}$. The smoothed HDI of $Y \mid X = x$ at level $1-\alpha$ is any pair $(m^\dagger_\beta(x), h^\dagger_\beta(x)) \in \mathbb{R} \times (0,\infty)$ satisfying

$$\Gamma\big(m^\dagger_\beta(x), h^\dagger_\beta(x); x\big) = 1-\alpha, \quad \text{and} \quad f_\beta\big(m^\dagger_\beta(x) + h^\dagger_\beta(x) \mid x\big) = f_\beta\big(m^\dagger_\beta(x) - h^\dagger_\beta(x) \mid x\big). \tag{8}$$

Lemma 1 (Existence, uniqueness, and optimality)**.** *Suppose $f(\cdot \mid x)$ is bounded, integrable, and unimodal. Then for every $\beta > 0$, the smoothed HDI $(m^\dagger_\beta(x), h^\dagger_\beta(x))$ exists, is unique, and coincides with the shortest $(1-\alpha)$-coverage interval of the smoothed density $f_\beta(\cdot \mid x)$.*

Lemma 2 (Vanishing smoothing bias)**.** *Under Assumption 1 (introduced below), for P_X-a.e. $x \in \mathcal{X}$,*

$$m^\dagger_\beta(x) \;\to\; m^*(x), \qquad h^\dagger_\beta(x) \;\to\; h^*(x), \qquad \text{as } \beta \to 0^+.$$

5.2 Asymptotic Conditional Coverage and Optimal Length

Throughout the asymptotic analysis, all observations are assumed i.i.d. from P. Let n be the total sample size, and let $n_{\text{train}} \asymp n$ and $n_{\text{cal}} \asymp n$ denote the sizes of the training and calibration sets, respectively. To achieve asymptotic optimality as $n \to \infty$, the temperature β_n must decay to zero to eliminate smoothing bias. Furthermore, the number of alternating iterations T in Algorithm 1 should be sufficiently large to reach the optimization fixed point. We formalize this through standard regularity conditions on the data distribution (Assumption 1), the consistency of the base estimators (Assumption 2), and the stability of the calibration scores (Assumption 3).

Assumption 1 (Density regularity)**.** For P_X-a.e. $x \in \mathcal{X}$, the conditional law $Y \mid X = x$ admits a density $f(\cdot \mid x)$ such that: (D1) $f(\cdot \mid x)$ is uniformly bounded by $f_{\max} < \infty$, integrable, and unimodal, and $\mathbb{E}[|Y|] < \infty$; (D2) the shortest $(1-\alpha)$-coverage interval is unique, denoted as $C^*(x) = [m^*(x) - h^*(x), m^*(x) + h^*(x)]$; (D3) $f(\cdot \mid x)$ is continuous on a neighborhood of the endpoints, and $f(m^* - h^* \mid x) = f(m^* + h^* \mid x) > 0$; (D4) the conditional CDF $F(\cdot \mid x)$ is continuous and strictly increasing on the support.

Assumption 2 (Estimation consistency)**.** Let $\phi_0(m; X)$ denote the $(1-\alpha)$-quantile of $|Y - m| \mid X$. There exist a deterministic schedule $\beta_n \to 0^+$ and deterministic sequences $\eta_{m,n}, \rho_{m,n}, \eta_{h,n}, \rho_{h,n} \to 0$ such that, for a fresh test input $X \sim P_X$ independent of $\mathcal{D}_{\text{train}}$,

$$\mathbb{P}\Big(\mathbb{E}\left[(\widehat{m}(X) - m^\dagger_{\beta_n}(X))^2 \,\middle|\, \mathcal{D}_{\text{train}} \right] \le \eta_{m,n} \Big) \;\ge\; 1 - \rho_{m,n}, \tag{9}$$

$$\mathbb{P}\Big(\mathbb{E}\left[(\widehat{h}(X) - \phi_0(\widehat{m}(X); X))^2 \,\middle|\, \mathcal{D}_{\text{train}} \right] \le \eta_{h,n} \Big) \;\ge\; 1 - \rho_{h,n}. \tag{10}$$

Assumption 3 (Calibration regularity)**.** With probability tending to 1 (over $\mathcal{D}_{\text{train}}$), $\widehat{h}(X) \in [h_-, h_+]$ for some deterministic constants $0 < h_- \le h_+ < \infty$. Moreover, the score $S := |Y - \widehat{m}(X)|/\widehat{h}(X)$ has a Lebesgue density f_S bounded below by some $s_0 > 0$ on a neighborhood $(1-\delta_0, 1+\delta_0)$ of 1.

Under these conditions, CoCP not only guarantees marginal coverage but asymptotically recovers the oracle $(1-\alpha)$-HDI. Note that conditional coverage comes primarily from radius consistency around the learned center, while optimal length is achieved via center consistency toward the true HDI center.

Theorem 2 (Asymptotic conditional coverage and optimal length)**.** *Under Assumptions 1, 2, and 3, the CoCP prediction interval $\widehat{C}(\cdot)$ from Algorithm 1 satisfies the following as $n \to \infty$:*

(i) Calibration consistency: $\widehat{q} \xrightarrow{\mathbb{P}} 1$.

(ii) Asymptotic conditional coverage: there exist measurable sets $\Lambda_n = \Lambda_n(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}}) \subseteq \mathcal{X}$ with $\mathbb{P}(X \in \Lambda_n \mid \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}}) \to 1$ such that

$$\sup_{x \in \Lambda_n} \Big| \mathbb{P}(Y \in \widehat{C}(X) \mid X = x, \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}}) - (1-\alpha) \Big| \;\xrightarrow{\mathbb{P}}\; 0. \tag{11}$$

(iii) Asymptotic optimal length: $\mathrm{Len}(\widehat{C}(X)) \big/ \mathrm{Len}(C^*(X)) \xrightarrow{\mathbb{P}} 1$.

Remark 1 (Boundary HDIs). Assumption 1(D3) excludes distributions whose true HDI touches a support boundary (e.g., the exponential law, with HDI $[0, q_{1-\alpha}]$). In such cases the equal-density condition fails for f itself; however, by Proposition 1, f_β remains strictly unimodal on $\mathbb{R}$ for every $\beta > 0$, so the smoothed HDI is still interior and well-targeted by CoCP. Appendix B establishes an $O(\beta)$ boundary shift, compared with the $O(\beta^2)$ bias obtained for interior HDIs under stronger local smoothness conditions. The conclusions of Theorem 2 extend to this boundary regime when $\beta_n \to 0$.

Local sensitivity of conditional coverage. The density-equality condition also has implications for the sensitivity of conditional coverage to estimation errors. For an interior HDI, expanding the conditional coverage of $\widehat{C}(x)$ around (m^*, h^*) at $X = x$ yields

$$\begin{aligned}
&F(\widehat{m} + \widehat{qh} \mid x) - F(\widehat{m} - \widehat{qh} \mid x) - (1-\alpha) \\
&= F(\widehat{m} + \widehat{qh} \mid x) - F(\widehat{m} - \widehat{qh} \mid x) - [F(m^* + h^* \mid x) - F(m^* - h^* \mid x)] \\
&= [f(m^* + h^* \mid x) - f(m^* - h^* \mid x)]\,(\widehat{m} - m^*) \\
&\quad + [f(m^* + h^* \mid x) + f(m^* - h^* \mid x)]\,(\widehat{qh} - h^*) \\
&\quad + o(|\widehat{m} - m^*| + |\widehat{qh} - h^*|).
\end{aligned} \tag{12}$$

At the HDI, the endpoint densities are equal, so the first-order contribution of center-estimation error vanishes. Thus, conditional coverage is locally insensitive to small center errors when the radius is held fixed, with the leading error determined by the calibrated radius, which is estimated through a single conditional quantile of the folded residuals. This local robustness may contribute to improved conditional reliability in finite samples.

Remark 2 (Scope of guarantees). Note that for CQR, estimation errors in both conditional quantiles generally contribute to the first-order coverage error. In particular, under skewness, the endpoint densities need not agree, so even a common shift of both endpoints affects conditional coverage at first order. Nevertheless, consistently estimated CQR quantiles can also yield asymptotic conditional coverage under suitable regularity conditions [41]. The distinction therefore lies in the local sensitivity to estimation errors and, under unimodality, CoCP's additional recovery of the shortest conditional interval. Theorem 1 provides the standard finite-sample marginal guarantee under exchangeability, while Theorem 2(ii) establishes asymptotic conditional coverage uniformly on data-dependent sets Λ_n whose covariate probability tends to one, subject to Assumptions 1–3.

Consistency of the ensemble. Proposition 2 characterizes the idealized population dynamics, whereas Algorithm 1 averages fold-specific predictors before calibration. The consistency conditions in Theorem 2 therefore concern the final ensemble. For fixed K, if each fold-specific center $\widehat{m}^{(k)}$ is L^2-consistent for the common target $m^\dagger_{\beta_n}$, Jensen's inequality gives

$$\mathbb{E}\Big[\big(\widehat{m}(X) - m^\dagger_{\beta_n}(X)\big)^2\Big] \le \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}\Big[\big(\widehat{m}^{(k)}(X) - m^\dagger_{\beta_n}(X)\big)^2\Big], \tag{13}$$

so averaging preserves center consistency. The same conclusion holds for the radii when each fold-specific radius consistently estimates its corresponding folded-residual quantile, using the Lipschitz continuity of ϕ_0 (Lemma 3) and the consistency of the fold-specific centers.

6 Experiments

Protocol and implementation. We evaluate CoCP on (i) controlled synthetic benchmarks with known conditional distributions and (ii) real-world regression datasets. We target $\alpha = 0.1$ and report the mean (standard deviation) over 10 random splits of the data into $50\%/10\%/20\%/20\%$ (train/val/cal/test). The validation and calibration set are used for early stopping and conformal calibration, respectively. All learning-based components share the same backbone (a 2-layer, 64-unit MLP) for fair comparison. More implementation details are provided in Appendix C.

Baselines. We compare against standard residual-based split CP (Split [33]), CQR [39], distributional-set baselines (C-HDR [25, 13], CHR [42]), and conditional-coverage-oriented methods (CPL [27], CPCP [9], RCP [36]). Due to the prohibitive computational cost of RKHS-based conditional conformal methods [18], we include a simplified version without test imputation, denoted as Naive CC, exclusively on synthetic data.

Metrics. We always report marginal Coverage and mean interval Length. For conditional reliability, we report the exact mean absolute coverage error (ConMAE) on synthetic data where the true conditional CDF is known. On real data, we report diagnostic surrogates: Mean Squared Conditional Error (MSCE [28]), Worst-Slice Coverage (WSC [8, 40]), and Excess Risk of Target coverage (ERT [5]), including ℓ_1-ERT and ℓ_2-ERT. Formal definitions are in Appendix C.4.

6.1 Synthetic Data

Setup. We generate $n = 20{,}000$ samples with $X \sim \mathrm{Unif}(-2, 2)$ and $Y = \theta(X) + \eta(X, \epsilon)$, where location $\theta(x) = \frac{1}{2}\sin(1.5x)$ and scale $s(x) = 0.15 + 0.25x^2$. To cover symmetric, moderately skewed, and highly skewed conditional distributions, we instantiate three noise families for the noise term: **Normal**: $\eta = s(X)\epsilon$ with $\epsilon \sim \mathcal{N}(0, 1)$; **LogNormal**: $\eta = s(X)(\epsilon - e^{-0.6^2})$ with $\epsilon \sim \mathrm{LogNormal}(0, 0.6^2)$; and **Exponential**: $\eta \sim \mathrm{Exp}(\lambda = 1/s(X))$, shifting the support to $[\theta(X), \infty)$. For reference, we include the **Oracle** interval, defined as the true $(1-\alpha)$-HDI; note it is an absolute length lower bound only among intervals attaining exact conditional coverage.

Main results. As shown in Table 1, CoCP consistently achieves the lowest conditional coverage error (ConMAE) among all practical baselines across the three synthetic benchmarks, demonstrating superior conditional reliability while maintaining nominal marginal coverage. The benefits of co-optimizing interval translation and scaling are particularly pronounced under skewed noise. On the LogNormal benchmark, CoCP substantially reduces average interval length (e.g., a $\approx 14\%$ reduction compared to CQR) while nearly halving the ConMAE. On the highly skewed Exponential benchmark, CoCP yields the tightest non-oracle intervals, reducing length by $\approx 21\%$ and ConMAE by $\approx 53\%$ relative to CQR. As corroborated by Figure 3 and the $T = 0$ ablation in Appendix D.3, the K-fold ensemble inherently stabilizes intervals, but it is the alternating center updates that correct skew-induced mis-centering and recover the optimal HDI geometry.

Table 1: Experimental results on synthetic datasets (mean $\pm$ std over 10 splits, $\alpha = 0.1$). Lower Length and ConMAE are better. The Oracle row reports the true $(1-\alpha)$-HDI for reference.

Method	Normal			LogNormal			Exponential		
	Coverage	Length (↓)	ConMAE (↓)	Coverage	Length (↓)	ConMAE (↓)	Coverage	Length (↓)	ConMAE (↓)
Oracle	0.9000 (0.0000)	1.5922 (0.0101)	0.0000 (0.0000)	0.9000 (0.0000)	0.9591 (0.0061)	0.0000 (0.0000)	0.9000 (0.0000)	1.1145 (0.0070)	0.0000 (0.0000)
Split	0.8995 (0.0050)	1.8735 (0.0404)	0.1119 (0.0030)	0.9000 (0.0051)	1.2377 (0.0301)	0.0964 (0.0036)	0.8989 (0.0031)	1.5679 (0.0217)	0.0937 (0.0036)
CPL	0.9010 (0.0091)	**1.5065** (0.0382)	0.0764 (0.0061)	0.9009 (0.0047)	0.9741 (0.0229)	0.0509 (0.0039)	0.8968 (0.0076)	1.1524 (0.0416)	0.0346 (0.0061)
Naive CC	0.9001 (0.0033)	1.5902 (0.0159)	0.0093 (0.0021)	0.9001 (0.0028)	0.9694 (0.0171)	0.0078 (0.0025)	0.8962 (0.0029)	1.2270 (0.0431)	0.0092 (0.0026)
CQR	0.9013 (0.0043)	1.5996 (0.0167)	0.0091 (0.0020)	0.9006 (0.0050)	1.1123 (0.0317)	0.0127 (0.0043)	0.8973 (0.0029)	1.4251 (0.0186)	0.0164 (0.0045)
CPCP	0.9027 (0.0130)	1.6046 (0.0362)	0.0189 (0.0052)	0.9014 (0.0151)	0.9813 (0.0471)	0.0169 (0.0048)	0.8933 (0.0090)	1.2276 (0.0485)	0.0149 (0.0071)
RCP	0.9004 (0.0045)	1.5946 (0.0349)	0.0136 (0.0037)	0.9003 (0.0047)	0.9759 (0.0293)	0.0123 (0.0046)	0.8995 (0.0043)	1.2496 (0.0345)	0.0101 (0.0029)
C-HDR	0.9005 (0.0038)	1.5988 (0.0217)	0.0086 (0.0021)	0.9008 (0.0030)	0.9797 (0.0220)	0.0096 (0.0021)	0.8971 (0.0024)	1.2514 (0.0328)	0.0080 (0.0024)
CHR	0.9025 (0.0043)	1.6207 (0.0195)	0.0097 (0.0019)	0.9031 (0.0055)	0.9894 (0.0142)	0.0111 (0.0018)	0.9019 (0.0025)	1.1662 (0.0157)	0.0113 (0.0023)
CoCP	0.9007 (0.0037)	1.5977 (0.0153)	**0.0066** (0.0013)	0.9002 (0.0029)	**0.9599** (0.0132)	**0.0066** (0.0019)	0.8975 (0.0026)	**1.1277** (0.0121)	**0.0077** (0.0022)

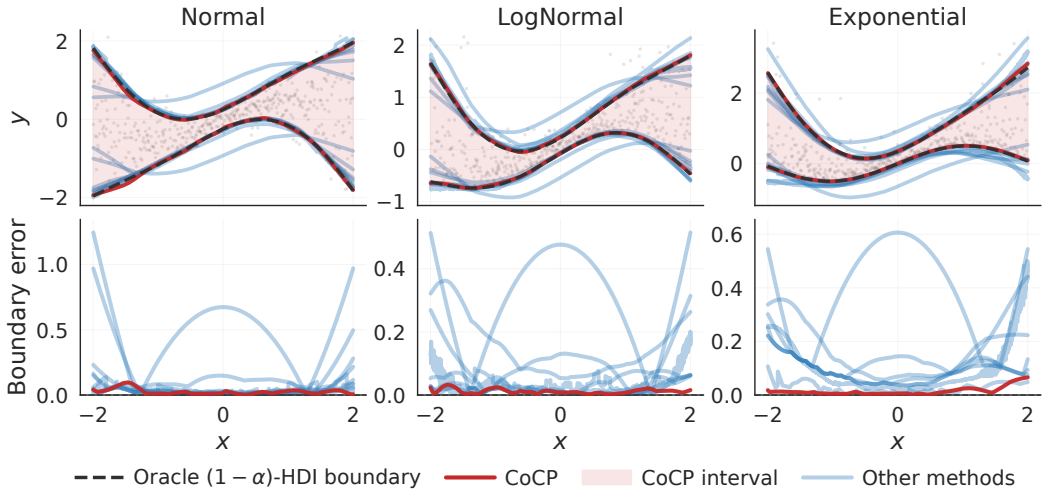

Figure 3: Interval geometry on synthetic benchmarks. **Top:** CoCP and baseline intervals vs. the oracle $(1-\alpha)$-HDI. **Bottom:** Boundary error (mean absolute error between predicted and oracle endpoints). While most baselines perform comparably under symmetric noise, under skewness they suffer from either severe mis-centering or full-density estimation artifacts. CoCP avoids these issues, maintaining tight geometric alignment with the oracle HDI across all settings.

6.2 Real Data

Setup. We evaluate on seven commonly used real-world regression datasets spanning different feature dimensions and noise characteristics: *Bike* [14], *Bio* [37], *Blog* [6], *Facebook-1* [44], *Facebook-2* [44], *Homes* [22], and *Supercon.* [21] (summarized in Appendix C.1). Because true conditional coverage is unavailable, we report efficiency measures (Length) and conditional-reliability diagnostics (MSCE/WSC/ERT), all computed on the preprocessed response scale (Appendix C.2).

Table 2: Experimental results on real-world datasets (mean $\pm$ std over 10 splits, $\alpha = 0.1$). Best values per dataset/metric (excluding Coverage) are highlighted in bold.

Dataset	Method	Coverage	Length (↓)	MSCE (↓)	WSC (↑)	ℓ_1-ERT (↓)	ℓ_2-ERT (↓)
Bike	Split	0.8980 (0.0111)	0.7700 (0.0360)	0.0041 (0.0016)	0.6974 (0.0259)	0.0663 (0.0063)	0.0080 (0.0022)
	CPL	0.9146 (0.0100)	0.7938 (0.0373)	0.0021 (0.0015)	0.7800 (0.0515)	0.0369 (0.0087)	0.0022 (0.0023)
	CQR	0.8993 (0.0040)	0.7856 (0.1255)	0.0018 (0.0006)	0.7573 (0.0180)	0.0350 (0.0127)	0.0020 (0.0015)
	CPCP	0.9030 (0.0162)	0.7475 (0.0515)	0.0010 (0.0003)	0.7933 (0.0251)	0.0206 (0.0090)	0.0003 (0.0007)
	RCP	0.9014 (0.0096)	0.7497 (0.0474)	0.0013 (0.0006)	0.7787 (0.0158)	0.0250 (0.0125)	0.0008 (0.0009)
	C-HDR	0.8991 (0.0089)	0.7617 (0.0327)	0.0015 (0.0005)	0.7740 (0.0123)	0.0270 (0.0133)	0.0013 (0.0014)
	CHR	0.9067 (0.0122)	0.7018 (0.0349)	0.0013 (0.0004)	0.7919 (0.0218)	0.0279 (0.0095)	0.0011 (0.0007)
	CoCP	0.8988 (0.0059)	**0.6053** (0.0190)	**0.0006** (0.0004)	**0.7982** (0.0096)	**0.0131** (0.0078)	**-0.0001** (0.0004)
Bio	Split	0.9032 (0.0049)	1.7584 (0.0301)	0.0006 (0.0002)	0.8258 (0.0070)	0.0241 (0.0044)	0.0009 (0.0003)
	CPL	0.9057 (0.0081)	1.6358 (0.0449)	0.0005 (0.0002)	0.8438 (0.0130)	0.0161 (0.0042)	0.0004 (0.0003)
	CQR	0.8986 (0.0052)	1.5251 (0.0164)	0.0003 (0.0001)	0.8455 (0.0079)	0.0088 (0.0028)	0.0001 (0.0001)
	CPCP	0.9048 (0.0072)	1.6452 (0.0413)	0.0004 (0.0002)	0.8463 (0.0119)	0.0134 (0.0049)	0.0002 (0.0002)
	RCP	0.9013 (0.0042)	1.6355 (0.0265)	0.0003 (0.0003)	0.8403 (0.0090)	0.0074 (0.0063)	0.0003 (0.0004)
	C-HDR	0.9019 (0.0046)	1.5601 (0.0292)	0.0006 (0.0004)	0.8328 (0.0153)	0.0179 (0.0067)	0.0006 (0.0004)
	CHR	0.9017 (0.0072)	1.5145 (0.0350)	0.0003 (0.0001)	0.8460 (0.0097)	0.0106 (0.0044)	0.0001 (0.0001)
	CoCP	0.9013 (0.0056)	**1.4231** (0.0142)	**0.0002** (0.0001)	**0.8478** (0.0091)	**0.0059** (0.0038)	**-0.0000** (0.0001)
Blog	Split	0.8998 (0.0040)	3.7416 (0.0604)	0.0048 (0.0007)	0.7349 (0.0140)	0.0726 (0.0043)	0.0068 (0.0015)
	CPL	0.9179 (0.0069)	3.7549 (0.1810)	0.0012 (0.0004)	**0.8503** (0.0093)	0.0377 (0.0060)	-0.0000 (0.0008)
	CQR	0.8973 (0.0027)	2.7947 (0.1018)	0.0014 (0.0004)	0.8087 (0.0101)	0.0300 (0.0048)	-0.0007 (0.0007)
	CPCP	0.8992 (0.0061)	3.3285 (0.0708)	0.0005 (0.0003)	0.8328 (0.0137)	0.0353 (0.0075)	0.0003 (0.0013)
	RCP	0.8980 (0.0043)	3.3837 (0.0917)	0.0008 (0.0004)	0.8154 (0.0130)	0.0359 (0.0037)	0.0009 (0.0009)
	C-HDR	0.8978 (0.0023)	3.1153 (0.2141)	0.0011 (0.0008)	0.8230 (0.0216)	0.0268 (0.0063)	-0.0005 (0.0012)
	CHR	0.9093 (0.0049)	**2.5531** (0.0744)	0.0006 (0.0003)	0.8479 (0.0117)	0.0208 (0.0047)	**-0.0018** (0.0006)
	CoCP	0.8994 (0.0046)	2.7025 (0.0897)	**0.0004** (0.0001)	0.8416 (0.0069)	**0.0177** (0.0036)	-0.0016 (0.0007)
Facebook-1	Split	0.9012 (0.0053)	2.1838 (0.0341)	0.0021 (0.0004)	0.7814 (0.0154)	0.0547 (0.0051)	0.0037 (0.0009)
	CPL	0.8993 (0.0069)	1.8935 (0.0756)	0.0004 (0.0002)	0.8425 (0.0090)	0.0172 (0.0061)	0.0002 (0.0005)
	CQR	0.8984 (0.0060)	1.7500 (0.0462)	0.0008 (0.0003)	0.8144 (0.0099)	0.0249 (0.0074)	0.0006 (0.0007)
	CPCP	0.8991 (0.0053)	1.9626 (0.0567)	0.0005 (0.0003)	0.8252 (0.0197)	0.0235 (0.0108)	0.0005 (0.0011)
	RCP	0.8998 (0.0049)	1.9151 (0.0394)	0.0003 (0.0001)	0.8395 (0.0117)	0.0146 (0.0049)	0.0000 (0.0004)
	C-HDR	0.9006 (0.0027)	1.8299 (0.0381)	0.0007 (0.0005)	0.8287 (0.0148)	0.0158 (0.0074)	0.0002 (0.0008)
	CHR	0.9059 (0.0052)	1.6920 (0.0344)	0.0006 (0.0002)	0.8297 (0.0074)	0.0188 (0.0080)	-0.0001 (0.0003)
	CoCP	0.8998 (0.0047)	**1.6380** (0.0232)	**0.0002** (0.0001)	**0.8470** (0.0074)	**0.0056** (0.0051)	**-0.0004** (0.0003)
Facebook-2	Split	0.8991 (0.0026)	2.0821 (0.0192)	0.0015 (0.0004)	0.7964 (0.0080)	0.0560 (0.0026)	0.0035 (0.0006)
	CPL	0.8973 (0.0034)	1.7975 (0.0381)	0.0002 (0.0001)	0.8482 (0.0093)	0.0155 (0.0037)	0.0003 (0.0003)
	CQR	0.8977 (0.0036)	1.6684 (0.0315)	0.0006 (0.0004)	0.8269 (0.0124)	0.0246 (0.0057)	0.0007 (0.0005)
	CPCP	0.8994 (0.0026)	1.8508 (0.0430)	0.0003 (0.0002)	0.8433 (0.0148)	0.0208 (0.0098)	0.0006 (0.0008)
	RCP	0.8996 (0.0031)	1.8083 (0.0380)	0.0003 (0.0001)	0.8477 (0.0089)	0.0160 (0.0028)	0.0004 (0.0002)
	C-HDR	0.8986 (0.0017)	1.7364 (0.0254)	0.0004 (0.0002)	0.8402 (0.0096)	0.0206 (0.0045)	0.0005 (0.0003)
	CHR	0.9020 (0.0032)	**1.4582** (0.0400)	0.0005 (0.0002)	0.8361 (0.0089)	0.0230 (0.0033)	0.0004 (0.0002)
	CoCP	0.8988 (0.0024)	1.5376 (0.0141)	**0.0001** (0.0001)	**0.8547** (0.0072)	**0.0103** (0.0023)	**-0.0001** (0.0001)
Homes	Split	0.9000 (0.0087)	0.5964 (0.0287)	0.0112 (0.0016)	0.5788 (0.0302)	0.0931 (0.0044)	0.0191 (0.0022)
	CPL	0.9087 (0.0119)	0.6331 (0.0377)	0.0009 (0.0004)	0.7909 (0.0335)	0.0357 (0.0093)	0.0021 (0.0015)
	CQR	0.9034 (0.0059)	0.5664 (0.0189)	0.0007 (0.0005)	0.8192 (0.0158)	0.0118 (0.0078)	0.0003 (0.0006)
	CPCP	0.8978 (0.0094)	0.5888 (0.0212)	0.0009 (0.0007)	0.7984 (0.0413)	0.0201 (0.0117)	0.0007 (0.0012)
	RCP	0.8992 (0.0047)	0.5845 (0.0204)	0.0006 (0.0004)	0.8020 (0.0207)	0.0169 (0.0082)	0.0008 (0.0009)
	C-HDR	0.9003 (0.0046)	0.5373 (0.0154)	0.0007 (0.0003)	0.8202 (0.0088)	0.0080 (0.0067)	-0.0001 (0.0003)
	CHR	0.9013 (0.0091)	0.5428 (0.0165)	0.0007 (0.0006)	0.8199 (0.0152)	0.0149 (0.0085)	0.0002 (0.0007)
	CoCP	0.9017 (0.0065)	**0.5242** (0.0087)	**0.0003** (0.0002)	**0.8283** (0.0145)	**0.0052** (0.0070)	**-0.0003** (0.0003)
Supercon.	Split	0.9036 (0.0065)	1.0186 (0.0442)	0.0066 (0.0010)	0.7061 (0.0246)	0.0875 (0.0035)	0.0088 (0.0009)
	CPL	0.9068 (0.0134)	0.9371 (0.0325)	0.0012 (0.0005)	0.8074 (0.0298)	0.0333 (0.0093)	0.0015 (0.0014)
	CQR	0.9013 (0.0072)	0.9059 (0.0345)	0.0006 (0.0002)	0.8094 (0.0104)	0.0211 (0.0069)	0.0000 (0.0007)
	CPCP	0.8971 (0.0124)	0.9186 (0.0426)	0.0007 (0.0007)	0.7998 (0.0371)	0.0269 (0.0084)	0.0007 (0.0015)
	RCP	0.9049 (0.0069)	0.9695 (0.0393)	0.0007 (0.0004)	0.8125 (0.0187)	0.0236 (0.0113)	0.0004 (0.0014)
	C-HDR	0.9050 (0.0068)	1.0642 (0.0455)	**0.0003** (0.0002)	0.8221 (0.0128)	0.0207 (0.0062)	-0.0001 (0.0005)
	CHR	0.9131 (0.0063)	0.9875 (0.0396)	0.0013 (0.0004)	0.8111 (0.0140)	0.0287 (0.0053)	0.0003 (0.0007)
	CoCP	0.9041 (0.0044)	**0.8167** (0.0156)	**0.0003** (0.0002)	**0.8245** (0.0069)	**0.0127** (0.0091)	**-0.0006** (0.0006)

Main results. Table 2 summarizes the performance across seven real-world datasets. CoCP achieves the shortest average interval lengths on five datasets while satisfying the nominal marginal coverage. On the remaining two (*Facebook-2* and *Blog*), CoCP remains highly competitive as the runner-up to CHR. This slight trade-off in length is offset by CoCP's superior conditional reliability: on *Facebook-2*, it outperforms CHR across all four diagnostics (MSCE, WSC, ℓ_1-ERT, and ℓ_2-ERT), while on *Blog*, it maintains lower MSCE and ℓ_1-ERT. Overall, CoCP consistently demonstrates robust empirical conditional reliability. It achieves the lowest MSCE and ℓ_1-ERT across all datasets, and leads in WSC and ℓ_2-ERT on six. These results confirm that CoCP effectively enhances local reliability while providing improved efficiency in prediction intervals.

Efficiency and Robustness. CoCP maintains the inference speed of standard CQR (typically < 0.05 seconds) despite its alternating optimization (Appendix D.1). Its offline cost is substantially higher: for example, on *Facebook-2*, the base implementation takes about 1844 seconds versus 66 seconds for CQR. The Fast variant (Appendix D.2) reduces the CoCP time to about 463 seconds via fold parallelism and strict early stopping, but CoCP remains best suited to settings where retraining is infrequent and inference is repeated. Appendix D.4 gives a practical adaptive-β strategy based on an initial local-radius estimate; the asymptotic theory applies to vanishing smoothing schedules.

Further analysis. Appendix D reports several supplementary investigations. A $T = 0$ ablation (Appendix D.5) isolates the contribution of center refinement: raising T from 0 to 5 shortens intervals on nine of ten datasets, while the unrefined center remains competitive with the Split predictor in point-prediction error. A separate predictor can thus coexist with CoCP for point decisions. Fold-disagreement analysis (Appendix D.6) suggests that variance reduction contributes to the stronger benefit of ensembling on real data, while local dip tests provide little evidence of widespread residual multimodality. A validation-tuned asymmetric CQR baseline (Appendix D.7) substantially improves over equal-tailed CQR under skewness, yet CoCP remains shorter on nine of ten datasets and achieves lower conditional error on all ten, confirming that its gains extend beyond relaxing the equal-tailed constraint. The ranking is stable across coverage levels $\alpha \in \{0.05, 0.1, 0.2\}$ even without per-level tuning (Appendix D.8), and persists for training sets as small as $n = 2{,}000$, calibration sets as small as $n_{\text{cal}} = 500$, and three additional real datasets with $n \leq 2{,}200$ (Appendix D.9).

7 Conclusions

This work presented CoCP, a novel conformal regression framework that overcomes the inefficiency of traditional center-invariant intervals. By framing interval construction as a boundary-balancing problem within a folded geometry, CoCP co-optimizes interval translation and scaling. The resulting dynamic push-pull mechanism allows the interval to autonomously seek high-density regions, recovering the optimal HDI under unimodality. CoCP guarantees distribution-free finite-sample marginal coverage from split conformal prediction, while achieving asymptotic conditional coverage and optimal interval length under standard regularity and consistency conditions. Empirically, it produces tighter prediction intervals with stronger conditional reliability across the evaluated benchmarks.

While highly effective, CoCP indeed involves certain trade-offs. Theoretically, its optimality assumes unimodal target distributions, leaving multimodal regions as an open challenge. It also relies on the consistency of base estimators, which is not strictly guaranteed for non-convex neural networks. Practically, the alternating K-fold optimization increases offline training costs (though the inference time is negligible), and performance depends on carefully tuning the smoothing temperature β. Furthermore, extending this geometric push-pull mechanism to multivariate or structured outputs with generative or representation-based construction pipelines [13] remains nontrivial. We provide a more comprehensive discussion of these limitations and exciting avenues for future work in Appendix E.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62506315), Hong Kong Research Grant Council (CityU 21213626), Guangdong Provincial Natural Science Foundation (Grant No. 2026A1515011937), and City University of Hong Kong (Grant No. 9610639). We thank the reviewers and the Area Chair for their valuable comments.

The authors declare no competing interests.

References

[1] C.M. Achilles, Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. Tennessee's Student Teacher Achievement Ratio (STAR) project, 2008. URL `https://doi.org/10.7910/DVN/SIWH9F`.

[2] Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 03 2023. ISSN 1935-8237. doi: 10.1561/2200000101.

[3] Konstantina Bairaktari, Jiayun Wu, and Steven Wu. Kandinsky conformal prediction: Beyond class- and covariate-conditional coverage. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 2581–2602. PMLR, 13–19 Jul 2025.

[4] Jie Bao, Nicolo Colombo, Valery Manokhin, Suqun Cao, and Rui Luo. A review and comparative analysis of univariate conformal regression methods. In Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson, and Henrik Boström, editors, *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pages 282–304. PMLR, 10–12 Sep 2025.

[5] Sacha Braun, David Holzmüller, Michael I. Jordan, and Francis Bach. Conditional coverage diagnostics for conformal prediction. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 9757–9794. PMLR, 06–11 Jul 2026.

[6] Krisztian Buza. BlogFeedback. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C58S3F.

[7] George Casella and Roger Berger. *Statistical inference*. Chapman and Hall/CRC, 2024.

[8] Maxime Cauchois, Suyash Gupta, and John C. Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

[9] Qianyi Chen and Bo Li. Colorful pinball: Density-weighted quantile regression for conditional guarantee of conformal prediction. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 15118–15145. PMLR, 06–11 Jul 2026.

[10] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

[11] Nicolo Colombo. On training locally adaptive cp. In Harris Papadopoulos, Khuong An Nguyen, Henrik Boström, and Lars Carlsson, editors, *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pages 384–398. PMLR, 13–15 Sep 2023.

[12] Nicolas Deutschmann, Mattia Rigotti, and Maria Rodriguez Martinez. Adaptive conformal regression with split-jackknife+ scores. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=1fbTGC3BUD`.

[13] 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, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 13444–13485. PMLR, 13–19 Jul 2025.

[14] Hadi Fanaee-T. Bike Sharing. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5W894.

[15] Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 06 2021. ISSN 2049-8772. doi: 10.1093/imaiai/iaaa017.

[16] 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, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18495–18527. PMLR, 13–19 Jul 2025.

[17] Matteo Gasparin and Aaditya Ramdas. Improving the statistical efficiency of cross-conformal prediction. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18848–18867. PMLR, 13–19 Jul 2025.

[18] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1100–1126, 03 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf008.

[19] Leying Guan. Localized conformal prediction: a generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 03 2023. ISSN 1464-3510. doi: 10.1093/biomet/asac040.

[20] Naixin Guo, Rui Luo, and Zhixin Zhou. Fast conformal prediction using conditional interquantile intervals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 21468–21476, 2026.

[21] Kam Hamidieh. Superconductivity Data. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C53P47.

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

[23] Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: randomization enables robust guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):549–578, 04 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae103.

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

[25] 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.

[26] Ulf Johansson, Cecilia Sönströd, and Henrik Boström. Conformal regression with reject option. In Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström, and Lars Carlsson, editors, *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*, pages 277–294. PMLR, 09–11 Sep 2024.

[27] Shayan Kiyani, George Pappas, and Hamed Hassani. Length optimization in conformal prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 99519–99563. Curran Associates, Inc., 2024. doi: 10.52202/079017-3158.

[28] Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction with learned features. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 24749–24769. PMLR, 21–27 Jul 2024.

[29] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

[30] 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.

[31] Yating Liu, Yeo Jin Jung, Zixuan Wu, Sowon Jeong, and Claire Donnat. SpeedCP: Fast kernel-based conditional conformal prediction. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 76980–77027. PMLR, 06–11 Jul 2026.

[32] Rui Luo and Zhixin Zhou. Conformal thresholded intervals for efficient regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19216–19223, 2025.

[33] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European conference on machine learning*, pages 345–356. Springer, 2002.

[34] Yash P. Patel, Sahana Rayan, and Ambuj Tewari. Conformal contextual robust optimization. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2485–2493. PMLR, 02–04 May 2024.

[35] Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4075–4084. PMLR, 10–15 Jul 2018.

[36] 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, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 49459–49492. PMLR, 13–19 Jul 2025.

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

[38] Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2002. DOI: https://doi.org/10.24432/C53W3X.

[39] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[40] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

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

[42] Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6304–6315. Curran Associates, Inc., 2021.

[43] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.*, 9: 371–421, June 2008. ISSN 1532-4435.

[44] Kamaljot Singh. Facebook Comment Volume. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5Q886.

[45] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[46] Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1):9–28, 2015.

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

[48] I-Cheng Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5PK67.

Appendix Contents

A Theoretical Proofs of Main Results **15**
- A.1 Standing Notation . . . 15
- A.2 Auxiliary Lemmas . . . 15
- A.3 Proof of Proposition 1 . . . 17
- A.4 Proof of Proposition 2 . . . 18
- A.5 Proof of Theorem 1 . . . 18
- A.6 Proof of Lemma 1 . . . 19
- A.7 Proof of Lemma 2 . . . 19
- A.8 Calibration Consistency . . . 20
- A.9 Proof of Theorem 2 . . . 21

B Sharper Bias Rates and Boundary HDIs **22**
- B.1 Second-Order β-Bias in the Interior HDI Regime . . . 22
- B.2 Active Support Boundaries: First-Order β-Shift . . . 24

C Experimental Details **25**
- C.1 Real-world Datasets Details . . . 25
- C.2 Preprocessing . . . 26
- C.3 Implementation Details . . . 26
- C.4 Metrics and Estimation . . . 27

D Additional Results **28**
- D.1 Runtime Comparison . . . 28
- D.2 Accelerating Training: The Fast Variant . . . 29
- D.3 Sensitivity Analysis and Ablation . . . 30
- D.4 Sample-Adaptive Smoothing Strategy . . . 34
- D.5 Center Refinement and Point Prediction . . . 35
- D.6 Fold Variability and Local Residual Modality . . . 36
- D.7 Comparison with Asymmetric Tail Allocation . . . 37
- D.8 Robustness Across Coverage Levels . . . 37
- D.9 Sensitivity to Data Availability . . . 38

E Limitations and Future Work **40**

A Theoretical Proofs of Main Results

A.1 Standing Notation

We briefly define the shorthand notations used throughout the appendix to distinguish between *hard* and *smoothed* objects. The conditional density and CDF of $Y \mid X = x$ are denoted by $f(\cdot \mid x)$ and $F(\cdot \mid x)$. The true HDI endpoints are $\ell^*(x)$ and $u^*(x)$, with center $m^*(x)$ and radius $h^*(x)$.

For any candidate center $m \in \mathbb{R}$, we define the hard radius map $\phi_0(m; x)$ as the conditional $(1-\alpha)$-quantile of $|Y - m| \mid X = x$, with corresponding endpoints

$$\ell_0(m; x) \coloneqq m - \phi_0(m; x) \qquad \text{and} \qquad u_0(m; x) \coloneqq m + \phi_0(m; x).$$

The smoothed radius map is $\phi_\beta(m; x) \coloneqq \inf\{h > 0 : \Gamma(m, h; x) \geq 1 - \alpha\}$. The smoothed HDI endpoints are $\ell^\dagger_\beta(x) \coloneqq m^\dagger_\beta(x) - h^\dagger_\beta(x)$ and $u^\dagger_\beta(x) \coloneqq m^\dagger_\beta(x) + h^\dagger_\beta(x)$. We emphasize the crucial distinction that $h^\dagger_\beta(x) = \phi_\beta(m^\dagger_\beta(x); x) \neq \phi_0(m^\dagger_\beta(x); x)$ in general.

To analyze the density equality condition, we define the endpoint-imbalance functions:

$$G_0(m; x) \coloneqq f(u_0(m; x) \mid x) - f(\ell_0(m; x) \mid x),$$
$$G_\beta(m; x) \coloneqq f_\beta(m + \phi_\beta(m; x) \mid x) - f_\beta(m - \phi_\beta(m; x) \mid x).$$

By Definition 1, $G_\beta(m^\dagger_\beta(x); x) = 0$, and under Assumption 1(D3), $G_0(m^*(x); x) = 0$.

A.2 Auxiliary Lemmas

Lemma 3 (Lipschitz dependence on the center)**.** *For any x and any $m, m' \in \mathbb{R}$, $|\phi_0(m; x) - \phi_0(m'; x)| \leq |m - m'|$ and $|\phi_\beta(m; x) - \phi_\beta(m'; x)| \leq |m - m'|$.*

Proof. Let $U = |Y - m|$ and $V = |Y - m'|$ conditional on $X = x$. By the reverse triangle inequality, $|U - V| = \big||Y - m| - |Y - m'|\big| \le |m - m'|$ almost surely. Hence $U \le V + |m - m'|$ and $V \le U + |m - m'|$ almost surely. Taking the $(1-\alpha)$-quantile on both sides preserves the inequality, yielding $Q_{1-\alpha}(U) \le Q_{1-\alpha}(V) + |m - m'|$ and vice versa, which proves the bound for ϕ_0.

For the smoothed radius ϕ_β, observe that the logistic convolution is equivalent to adding independent noise $\varepsilon_\beta \sim K_\beta$. Let $Y_\beta = Y + \varepsilon_\beta$. Then $\Gamma(m, h; x) = \mathbb{P}(|Y_\beta - m| \le h \mid X = x)$, which means $\phi_\beta(m; x) = Q_{1-\alpha}(|Y_\beta - m| \mid X = x)$. Applying the same reverse triangle inequality to Y_β yields $|\phi_\beta(m; x) - \phi_\beta(m'; x)| \le |m - m'|$. □

Lemma 4 (Coverage is Lipschitz in the radius)**.** *Under Assumption 1, for any x, $m \in \mathbb{R}$ and $h_1, h_2 \ge 0$,*

$$\big|\mathbb{P}(|Y - m| \le h_1 \mid X = x) - \mathbb{P}(|Y - m| \le h_2 \mid X = x)\big| \le 2 f_{\max} |h_1 - h_2|.$$

Proof. Observe that $\mathbb{P}(|Y - m| \le h \mid X = x) = F(m + h \mid x) - F(m - h \mid x)$. Since the density is bounded by $f_{\max}$, the CDF $F(\cdot \mid x)$ is $f_{\max}$-Lipschitz. Thus, each endpoint contributes at most $f_{\max}|h_1 - h_2|$ to the difference. □

Lemma 5 (Lower bound on hard and smoothed radii)**.** *For all x and $m \in \mathbb{R}$,*

$$\phi_0(m; x) \ge h_{\min} \coloneqq (1 - \alpha)/(2 f_{\max}).$$

The same lower bound holds for $\phi_\beta(m; x)$. In particular, $h^*(x), h^\dagger_\beta(x) \ge h_{\min}$.

Proof. Since $f(\cdot \mid x) \le f_{\max}$, the maximum probability mass any interval of length $2h$ can capture is $2h f_{\max}$. To achieve $1 - \alpha$ coverage, we must have $2h f_{\max} \ge 1 - \alpha$, meaning any $h < h_{\min}$ is infeasible. Convolution does not increase the L^∞ norm, so $\|f_\beta(\cdot \mid x)\|_\infty \le f_{\max}$, and the same lower bound holds for ϕ_β. □

Lemma 6 (Uniform L^∞ control of $F_\beta - F$)**.** *For every x and $\beta > 0$, $\sup_y |F_\beta(y \mid x) - F(y \mid x)| \le \|f_\beta(\cdot \mid x) - f(\cdot \mid x)\|_{L^1}$, and the right-hand side tends to 0 as $\beta \to 0^+$.*

Proof. By definition, $F_\beta(y \mid x) - F(y \mid x) = \int_{-\infty}^{y} (f_\beta - f)(z \mid x)\, dz$. Taking the absolute value yields $|F_\beta(y \mid x) - F(y \mid x)| \le \int_{-\infty}^{\infty} |f_\beta(z \mid x) - f(z \mid x)|\, dz = \|f_\beta - f\|_{L^1}$ uniformly in y. The vanishing of $\|K_\beta * f - f\|_{L^1}$ for any integrable f is a standard property of mollifiers. □

Lemma 7 (Local uniform convergence of f_β to f)**.** *Fix x and a compact set $\mathcal{K} \subset \mathbb{R}$. Suppose there exists $\delta_0 > 0$ such that $f(\cdot \mid x)$ is continuous on $\mathcal{K}^{\delta_0} \coloneqq \{y \in \mathbb{R} : \inf_{z \in \mathcal{K}} |y - z| \le \delta_0\}$. Then $\sup_{y \in \mathcal{K}} |f_\beta(y \mid x) - f(y \mid x)| \to 0$ as $\beta \to 0^+$.*

Proof. By continuity on the compact set $\mathcal{K}^{\delta_0}$, $f(\cdot \mid x)$ is uniformly continuous there. Fix $\varepsilon > 0$ and choose $0 < \delta \le \delta_0$ such that $|f(y - u \mid x) - f(y \mid x)| \le \varepsilon$ for $y \in \mathcal{K}$ and $|u| \le \delta$. Then for $y \in \mathcal{K}$,

$$\begin{aligned}
|f_\beta(y \mid x) - f(y \mid x)| &\le \int_{\mathbb{R}} |f(y - u \mid x) - f(y \mid x)| K_\beta(u)\, du \\
&\le \int_{|u| \le \delta} \varepsilon K_\beta(u)\, du + \int_{|u| > \delta} 2 f_{\max} K_\beta(u)\, du \\
&\le \varepsilon + 2 f_{\max} \int_{|t| > \delta/\beta} \sigma'(t)\, dt.
\end{aligned}$$

The second term vanishes as $\beta \to 0^+$ due to the exponential tails of σ'. □

Lemma 8. *Let Z_n be real-valued and $\mathcal{G}_n$ a sigma-field. If there exist deterministic sequences $\eta_n \to 0$ and $\rho_n \to 0$ such that $\mathbb{P}\big(\mathbb{E}[Z_n^2 \mid \mathcal{G}_n] \le \eta_n\big) \ge 1 - \rho_n$, then $Z_n \xrightarrow{\mathbb{P}} 0$. In particular, for every $\varepsilon > 0$,*

$$\mathbb{P}(|Z_n| > \varepsilon) \le \frac{\eta_n}{\varepsilon^2} + \rho_n.$$

Proof. Let $A_n = \{\mathbb{E}[Z_n^2 \mid \mathcal{G}_n] \le \eta_n\}$. By the law of total probability and Markov's inequality conditional on $\mathcal{G}_n$:

$$\begin{aligned}\mathbb{P}(|Z_n| > \varepsilon) &\le \mathbb{P}(|Z_n| > \varepsilon \mid A_n)\mathbb{P}(A_n) + \mathbb{P}(A_n^c) \\ &\le \mathbb{E}\left[\frac{\mathbb{E}[Z_n^2 \mid \mathcal{G}_n]}{\varepsilon^2}\mathbf{1}_{A_n}\right] + \rho_n \\ &\le \frac{\eta_n}{\varepsilon^2} + \rho_n.\end{aligned}$$

Since $\eta_n \to 0$ and $\rho_n \to 0$, it follows that $Z_n \xrightarrow{\mathbb{P}} 0$. □

A.3 Proof of Proposition 1

Proof. Fix x and suppress it from the notation. By the properties $K_\beta \in C^1(\mathbb{R})$, $K_\beta' \in L^\infty(\mathbb{R})$, and $f \in L^1(\mathbb{R})$, we obtain $f_\beta \in C^1(\mathbb{R})$ and $f_\beta' = K_\beta' * f$.

By the unimodality, there exists an (x-dependent) interval $[a, b]$ with $a \le b$ such that f is nondecreasing on $(-\infty, a]$, constant on $[a, b]$, and nonincreasing on $[b, \infty)$. Because f is bounded, integrable, and follows this monotonic structure, it is of bounded variation on $\mathbb{R}$. Hence its distributional derivative Df is a finite signed measure. Write the Jordan decomposition as $Df = \mu - \nu$, where μ and ν are positive measures. By the structural property of f, the positive variation μ is supported on $(-\infty, a]$ and the negative variation ν is supported on $[b, \infty)$. Additionally, the vanishing tails $\lim_{|y|\to\infty} f(y) = 0$ imply that $\mu(\mathbb{R}) = \nu(\mathbb{R}) > 0$.

In the sense of distributions, and hence pointwise by continuity, $f_\beta' = K_\beta * Df$. Therefore, we can write $f_\beta'(y) = A(y) - B(y)$, where

$$A(y) = \int K_\beta(y - u)\, d\mu(u), \qquad B(y) = \int K_\beta(y - v)\, d\nu(v).$$

Given that K_β is strictly positive and both μ and ν are non-trivial finite positive measures, it follows that $A(y) > 0$ and $B(y) > 0$ for all y. Recall that $K_\beta(t) = \frac{1}{4\beta}\operatorname{sech}^2\left(\frac{t}{2\beta}\right)$, which implies

$$(\log K_\beta)''(t) = -\frac{1}{2\beta^2}\operatorname{sech}^2\left(\frac{t}{2\beta}\right) < 0.$$

Thus K_β is strictly log-concave. Consequently, for every $u < v$, the function $y \mapsto K_\beta(y-u)/K_\beta(y-v)$ is strictly decreasing. Consider the ratio $R(y) = A(y)/B(y)$. For any $y_1 < y_2$, we examine the cross-multiplication difference:

$$\begin{aligned}&A(y_2)B(y_1) - A(y_1)B(y_2) \\ &= \iint [K_\beta(y_2 - u)K_\beta(y_1 - v) - K_\beta(y_1 - u)K_\beta(y_2 - v)]\, d\mu(u)d\nu(v).\end{aligned}$$

Since $\operatorname{supp}(\mu) \subseteq (-\infty, a]$ and $\operatorname{supp}(\nu) \subseteq [b, \infty)$ with $a \le b$, we have $u \le v$ for $\mu \times \nu$-almost all (u, v). Furthermore, by the fundamental properties of the Jordan decomposition, the measures μ and ν are mutually singular ($\mu \perp \nu$), meaning they share no point masses. Thus, the strict inequality $u < v$ holds $\mu \times \nu$-almost everywhere. By the strict log-concavity of K_β, the term in the bracket is strictly negative almost everywhere with respect to $\mu \times \nu$. This implies $A(y_2)B(y_1) - A(y_1)B(y_2) < 0$, meaning the ratio $R(y)$ is strictly decreasing on $\mathbb{R}$.

The condition $f_\beta'(y) = 0$ is equivalent to $R(y) = 1$, which guarantees that f_β' has at most one real zero. To establish existence, observe that $K_\beta \in C_0(\mathbb{R})$ and $f \in L^1(\mathbb{R})$ together yield $\lim_{|y|\to\infty} f_\beta(y) = 0$. Furthermore, $f_\beta(y) > 0$ everywhere because $K_\beta > 0$ and f is a probability density. Therefore, f_β attains a positive maximum at some finite point ν_β, and this maximum satisfies $f_\beta'(\nu_\beta) = 0$, proving that f_β' has exactly one real zero.

Finally, the strict monotonicity of R combined with $R(\nu_\beta) = 1$ ensures that $R(y) > 1$ for $y < \nu_\beta$ and $R(y) < 1$ for $y > \nu_\beta$. As $f_\beta'(y) = B(y)[R(y) - 1]$ with $B(y)$ strictly positive, this yields

$$f_\beta'(y) > 0 \quad \text{for } y < \nu_\beta, \qquad f_\beta'(y) < 0 \quad \text{for } y > \nu_\beta.$$

The proof is complete. □

A.4 Proof of Proposition 2

Proof. Fix $x \in \mathcal{X}$ and suppress it from the notation where clear. By definition, the radius is instantaneously equilibrated to $h_t = \phi_\beta(m_t; x)$. The center follows the negative population gradient flow of the smooth interval loss:

$$\frac{dm_t}{dt} = -\frac{\partial \mathcal{L}_{\mathrm{M}}}{\partial m}(m_t; h_t, \beta) = f_\beta(m_t + h_t \mid x) - f_\beta(m_t - h_t \mid x).$$

By the chain rule, the time derivative of the required radius along this trajectory is:

$$\frac{d}{dt}\phi_\beta(m_t; x) = \frac{\partial \phi_\beta(m_t; x)}{\partial m} \cdot \frac{dm_t}{dt}.$$

Recall from Equation (2) that the implicit derivative of the smoothed radius map with respect to the center is:

$$\frac{\partial \phi_\beta(m_t; x)}{\partial m} = -\frac{f_\beta(m_t + h_t \mid x) - f_\beta(m_t - h_t \mid x)}{f_\beta(m_t + h_t \mid x) + f_\beta(m_t - h_t \mid x)}.$$

Substituting both expressions into the chain rule yields:

$$\begin{aligned}\frac{d}{dt}\phi_\beta(m_t; x) &= \left(-\frac{f_\beta(m_t + h_t \mid x) - f_\beta(m_t - h_t \mid x)}{f_\beta(m_t + h_t \mid x) + f_\beta(m_t - h_t \mid x)}\right) \cdot \big(f_\beta(m_t + h_t \mid x) - f_\beta(m_t - h_t \mid x)\big) \\ &= -\frac{\big[f_\beta(m_t + h_t \mid x) - f_\beta(m_t - h_t \mid x)\big]^2}{f_\beta(m_t + h_t \mid x) + f_\beta(m_t - h_t \mid x)}.\end{aligned}$$

Since $f_\beta(\cdot \mid x)$ is a strictly positive smoothed density (as shown in Proposition 1), the denominator is strictly positive. The numerator is a squared real number, hence non-negative. Therefore, we conclude that:

$$\frac{d}{dt}\phi_\beta(m_t; x) \le 0.$$

Equality holds if and only if the numerator is zero, which means $f_\beta(m_t + h_t \mid x) = f_\beta(m_t - h_t \mid x)$. This is precisely the smoothed density-equality condition (the second defining property of the smoothed HDI). Thus, the continuous-time dynamic strictly and monotonically decreases the required radius until the interval is perfectly balanced. □

A.5 Proof of Theorem 1

Proof. By construction, the predictors $(\widehat{m}, \widehat{h})$ are trained on $\mathcal{D}_{\mathrm{train}}$ and are thus independent of $\mathcal{D}_{\mathrm{cal}} \cup \{(X_{n_{\mathrm{cal}}+1}, Y_{n_{\mathrm{cal}}+1})\}$. Define the nonconformity scores $S_i := |Y_i - \widehat{m}(X_i)|/\widehat{h}(X_i)$ for $i = 1, \ldots, n_{\mathrm{cal}} + 1$. Conditional on $(\widehat{m}, \widehat{h})$, the exchangeability of the input pairs implies the exchangeability of the scores S_i.

The covering event for the test point is:

$$\{Y_{n_{\mathrm{cal}}+1} \in \widehat{C}(X_{n_{\mathrm{cal}}+1})\} = \{S_{n_{\mathrm{cal}}+1} \le \widehat{q}\},$$

where $\widehat{q}$ is the $\lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$-th order statistic of $\{S_i\}_{i \le n_{\mathrm{cal}}}$ (and $+\infty$ in the degenerate case). By the fundamental property of exchangeable random variables, the probability that $S_{n_{\mathrm{cal}}+1}$ is among the smallest $\lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil$ values of the $n_{\mathrm{cal}} + 1$ scores is at least the proportional fraction, regardless of ties. Therefore,

$$\mathbb{P}(S_{n_{\mathrm{cal}}+1} \le \widehat{q}) \ge \frac{\lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil}{n_{\mathrm{cal}} + 1} \ge 1 - \alpha.$$

If all $n_{\mathrm{cal}} + 1$ scores are almost-surely distinct (i.e., there are no ties, which is guaranteed if the conditional score distribution is continuous), the rank of $S_{n_{\mathrm{cal}}+1}$ is exactly uniformly distributed on $\{1, \ldots, n_{\mathrm{cal}} + 1\}$, yielding the exact upper bound:

$$\mathbb{P}(S_{n_{\mathrm{cal}}+1} \le \widehat{q}) \le \frac{\lceil (n_{\mathrm{cal}} + 1)(1 - \alpha) \rceil}{n_{\mathrm{cal}} + 1} \le 1 - \alpha + \frac{1}{n_{\mathrm{cal}} + 1}.$$

□

A.6 Proof of Lemma 1

Proof. Fix $x \in \mathcal{X}$ and abbreviate $f_\beta \coloneqq f_\beta(\cdot \mid x)$. By Proposition 1, $f_\beta \in C^1(\mathbb{R})$, $f_\beta > 0$, f_β is strictly unimodal with unique mode ν_β, and $f_\beta(y) \to 0$ as $|y| \to \infty$.

Existence and Uniqueness. The level sets $L_c \coloneqq \{f_\beta \geq c\}$ are closed bounded intervals $[\ell(c), u(c)] \ni \nu_\beta$ for every $c \in (0, \sup f_\beta)$, with ℓ strictly increasing in c, u strictly decreasing, and $f_\beta(\ell(c)) = f_\beta(u(c)) = c$. Define the mass function

$$\Phi(c) \coloneqq \int_{\ell(c)}^{u(c)} f_\beta(y)\, dy.$$

By dominated convergence and the strict monotonicity of ℓ and u, Φ is continuous and strictly decreasing on $(0, \sup f_\beta)$, with $\Phi(0^+) = 1$ and $\Phi((\sup f_\beta)^-) = 0$. Hence there exists a unique $c^\dagger \in (0, \sup f_\beta)$ with $\Phi(c^\dagger) = 1 - \alpha$. Set

$$m_\beta^\dagger \coloneqq \frac{\ell(c^\dagger) + u(c^\dagger)}{2} \quad \text{and} \quad h_\beta^\dagger \coloneqq \frac{u(c^\dagger) - \ell(c^\dagger)}{2} > 0$$

Then $\Gamma(m_\beta^\dagger, h_\beta^\dagger; x) = \Phi(c^\dagger) = 1 - \alpha$ and $f_\beta(m_\beta^\dagger \pm h_\beta^\dagger) = c^\dagger$, satisfying the definition of the smoothed HDI. If any other (m, h) satisfies the conditions with $h > 0$, then $f_\beta(m+h) = f_\beta(m-h)$, and strict unimodality forces $\{m-h, m+h\} = \{\ell(c), u(c)\}$ for some $c \in (0, \sup f_\beta)$. The coverage condition $\Phi(c) = 1 - \alpha$ then forces $c = c^\dagger$ by uniqueness, hence $(m, h) = (m_\beta^\dagger, h_\beta^\dagger)$.

Optimality. Take any interval $[a, b]$ with length $b - a \leq u(c^\dagger) - \ell(c^\dagger) = 2h_\beta^\dagger$, assuming $[a, b] \neq L_{c^\dagger}$. We show $[a, b]$ cannot satisfy $\int_a^b f_\beta \geq 1 - \alpha$. Compute the difference in probability mass using $L_{c^\dagger} = [\ell(c^\dagger), u(c^\dagger)]$:

$$\begin{aligned}
\int_{L_{c^\dagger}} f_\beta - \int_{[a,b]} f_\beta &= \int_{L_{c^\dagger} \setminus [a,b]} f_\beta - \int_{[a,b] \setminus L_{c^\dagger}} f_\beta \\
&> c^\dagger \big(|L_{c^\dagger} \setminus [a,b]| - |[a,b] \setminus L_{c^\dagger}| \big) \\
&= c^\dagger \big(|L_{c^\dagger}| - |[a,b]| \big) \geq 0,
\end{aligned}$$

where the strict inequality uses $f_\beta > c^\dagger$ on the interior of $L_{c^\dagger}$ and $f_\beta < c^\dagger$ on $(L_{c^\dagger})^c$, and the fact that the symmetric difference is non-empty since $[a, b] \neq L_{c^\dagger}$. The equality uses the fact that $|L_{c^\dagger} \cap [a, b]|$ is identical in both expressions. Since $\int_{L_{c^\dagger}} f_\beta = 1 - \alpha$, it follows that $\int_{[a,b]} f_\beta < 1 - \alpha$, contradicting feasibility. Hence $L_{c^\dagger}$ is the unique shortest feasible interval. □

A.7 Proof of Lemma 2

Proof. Fix any x at which Assumption 1 holds, and abbreviate $m_\beta^\dagger \coloneqq m_\beta^\dagger(x)$, $h_\beta^\dagger \coloneqq h_\beta^\dagger(x)$, $\ell_\beta^\dagger \coloneqq m_\beta^\dagger - h_\beta^\dagger$, $u_\beta^\dagger \coloneqq m_\beta^\dagger + h_\beta^\dagger$, and similarly m^*, h^*, ℓ^*, u^* for the true HDI. By Lemma 6,

$$\|F_\beta(\cdot \mid x) - F(\cdot \mid x)\|_\infty \to 0, \quad \|f_\beta(\cdot \mid x) - f(\cdot \mid x)\|_{L^1} \to 0.$$

We argue by combining tightness, optimality, and HDI uniqueness.

By Lemma 5, $h_\beta^\dagger$ is lower bounded by

$$h_\beta^\dagger \geq h_{\min} = (1 - \alpha)/(2 f_{\max}) > 0.$$

Fix any $\varepsilon > 0$. By Assumption 1(D4),

$$F(u^* + \varepsilon \mid x) - F(\ell^* - \varepsilon \mid x) > F(u^* \mid x) - F(\ell^* \mid x) = 1 - \alpha.$$

Thus, by Lemma 6, $F_\beta(u^* + \varepsilon \mid x) - F_\beta(\ell^* - \varepsilon \mid x) > 1 - \alpha$ for β sufficiently small. Hence $[\ell^* - \varepsilon, u^* + \varepsilon]$ is feasible for the smoothed HDI, and Lemma 1 (optimality) gives $2h_\beta^\dagger \leq 2(h^* + \varepsilon)$ for β small. Letting $\varepsilon \to 0$, we obtain:

$$\limsup_{\beta \to 0^+} h_\beta^\dagger \leq h^*. \tag{A.1}$$

From $F_\beta(u_\beta^\dagger \mid x) - F_\beta(\ell_\beta^\dagger \mid x) = 1-\alpha$ and $\|F_\beta - F\|_\infty \to 0$, we have

$$F(u_\beta^\dagger \mid x) - F(\ell_\beta^\dagger \mid x) = 1 - \alpha + o(1).$$

If $\ell_\beta \to -\infty$ along some subsequence, then $u_\beta^\dagger = \ell_\beta^\dagger + 2h_\beta \le \ell_\beta^\dagger + 2(h^*+1) \to -\infty$ for β small (using (A.1)), so $F(u_\beta^\dagger \mid x) \to 0$ and $F(u_\beta^\dagger \mid x) - F(\ell_\beta^\dagger \mid x) \to 0$, contradicting $1-\alpha > 0$. Symmetrically $\ell_\beta^\dagger \to +\infty$ is excluded. Hence $\ell_\beta^\dagger$, $u_\beta^\dagger$, and $m_\beta^\dagger = (\ell_\beta^\dagger + u_\beta^\dagger)/2$ are bounded.

Take any subsequence $\beta_k \to 0^+$ along which $(m_{\beta_k}^\dagger, h_{\beta_k}^\dagger) \to (\bar{m}, \bar{h})$ with $\bar{h} \ge h_{\min} > 0$. By $\|F_\beta - F\|_\infty \to 0$ and continuity of $F(\cdot \mid x)$,

$$F(\bar{m}+\bar{h} \mid x) - F(\bar{m}-\bar{h} \mid x) = \lim_{k\to\infty} \left[F_{\beta_k}(m_{\beta_k}^\dagger + h_{\beta_k}^\dagger \mid x) - F_{\beta_k}(m_{\beta_k}^\dagger - h_{\beta_k}^\dagger \mid x) \right] = 1-\alpha.$$

So $[\bar{m}-\bar{h}, \bar{m}+\bar{h}]$ is a valid $(1-\alpha)$-coverage interval.

From (A.1), $\bar{h} \le h^*$. By Assumption 1, $[\ell^*, u^*]$ is the unique shortest $(1-\alpha)$-coverage interval, of length $2h^*$. Hence the $(1-\alpha)$-coverage interval $[\bar{m}-\bar{h}, \bar{m}+\bar{h}]$ of length $2\bar{h} \le 2h^*$ must satisfy $2\bar{h} \ge 2h^*$ (HDI optimality) and equal $[\ell^*, u^*]$ (HDI uniqueness). Therefore $\bar{h} = h^*$ and $\bar{m} = m^*$.

By tightness, every subsequence of $\{(m_\beta^\dagger, h_\beta^\dagger)\}_{\beta\to 0^+}$ has a further subsequence converging to some limit, which must be (m^*, h^*). Hence the full sequence converges to (m^*, h^*). □

A.8 Calibration Consistency

Proposition 3 (Calibration consistency)**.** *Under Assumptions 1, 2, 3 and $n_{\text{cal}} \to \infty$, $\widehat{q} \xrightarrow{\mathbb{P}} 1$ with rate $|\widehat{q} - 1| = O_{\mathbb{P}}\left(\sqrt{\eta_{h,n}} + n_{\text{cal}}^{-1/2}\right)$.*

Proof. Condition on $\mathcal{D}_{\text{train}}$. Let $S = |Y - \widehat{m}(X)|/\widehat{h}(X)$ with conditional CDF F_S, and let q denote the population conditional $(1-\alpha)$-quantile of S. We bound

$$|\widehat{q} - 1| \le |\widehat{q} - q| + |q - 1|.$$

For each $x \in \mathcal{X}$, by definition of ϕ_0,

$$\mathbb{P}(|Y - \widehat{m}(X)| \le \phi_0(\widehat{m}(X); X) \mid X = x, \mathcal{D}_{\text{train}}) = 1-\alpha.$$

Hence by Lemma 4,

$$\begin{aligned}
&|F_S(1) - (1-\alpha)| \\
&\le \mathbb{E}_X\left[\left| \mathbb{P}\big(|Y - \widehat{m}(X)| \le \widehat{h}(X) \mid X\big) - \mathbb{P}\big(|Y - \widehat{m}(X)| \le \phi_0(\widehat{m}(X); X) \mid X\big) \right| \right] \\
&\le 2 f_{\max}\, \mathbb{E}_X\left[\left| \widehat{h}(X) - \phi_0(\widehat{m}(X); X) \right| \right],
\end{aligned}$$

where all expectations are conditional on $\mathcal{D}_{\text{train}}$. By Cauchy–Schwarz and Assumption 2,

$$\mathbb{E}_X\left[\left| \widehat{h}(X) - \phi_0(\widehat{m}(X); X) \right| \right] \le \sqrt{\eta_{h,n}}$$

on the high-probability event of (10).

On the high-probability event $B_n = \{f_S \ge s_0 \text{ on } (1-\delta_0, 1+\delta_0)\}$ from Assumption 3, F_S has slope $\ge s_0$ on this neighborhood. Thus, the quantile map of S is locally $1/s_0$-Lipschitz. When $|F_S(1) - (1-\alpha)| \le s_0\delta_0/2$ (which happens w.p. tending to one), we have $|q-1| \le |F_S(1) - (1-\alpha)|/s_0 \le \delta_0/2$, ensuring $q \in (1-\delta_0/2, 1+\delta_0/2)$.

Conditional on $\mathcal{D}_{\text{train}}$, the calibration scores are i.i.d. with CDF F_S. Note that $\widehat{q}$ is the $\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil$-th order statistic. The target level is $1-\alpha+O(1/n_{\text{cal}})$. By the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, the empirical CDF satisfies

$$\sup_t |F_{S,n_{\text{cal}}}(t) - F_S(t)| = O_{\mathbb{P}}\left(n_{\text{cal}}^{-1/2}\right).$$

On B_n and for n_{cal} large, the $O(1/n_{\text{cal}})$ shift is absorbed, and $\widehat{q}$ will also lie in $(1-\delta_0, 1+\delta_0)$ with high probability. Within this neighborhood, the quantile map is $1/s_0$-Lipschitz, yielding the same rate for $|\widehat{q} - q|$.

Combining these yields $|\widehat{q} - 1| = O_{\mathbb{P}}\left(\sqrt{\eta_{h,n}} + n_{\text{cal}}^{-1/2}\right) \xrightarrow{\mathbb{P}} 0$. □

A.9 Proof of Theorem 2

Proof. **(i) Calibration consistency** is established in Proposition 3.

(ii) Asymptotic conditional coverage. Choose any deterministic $\delta_n \to 0^+$ such that $\eta_{h,n}/\delta_n^2 \to 0$ and $n_{\text{cal}}^{-1/2}/\delta_n \to 0$; e.g. $\delta_n = \eta_{h,n}^{1/4} + n_{\text{cal}}^{-1/4}$. Define the random sets (depending on $\mathcal{D}_{\text{train}}$):

$$\Lambda_n \coloneqq \big\{x \in \mathcal{X} : \big|\widehat{h}(x) - \phi_0(\widehat{m}(x); x)\big| \le \delta_n \text{ and } \widehat{h}(x) \le h_+\big\}.$$

Let

$$E_{h,n} \coloneqq \big\{\mathbb{E}[(\widehat{h}(X) - \phi_0(\widehat{m}(X); X))^2 \mid \mathcal{D}_{\text{train}}] \le \eta_{h,n}\big\}.$$

On $E_{h,n}$, conditional Markov's inequality gives

$$\mathbb{P}\big(|\widehat{h}(X) - \phi_0(\widehat{m}(X); X)| > \delta_n \mid \mathcal{D}_{\text{train}}\big) \le \eta_{h,n}/\delta_n^2.$$

Since $\mathbb{P}(E_{h,n}^c) \le \rho_{h,n} \to 0$, and $\mathbb{P}(\widehat{h}(X) \le h_+ \mid \mathcal{D}_{\text{train}}) \xrightarrow{\mathbb{P}} 1$ from Assumption 3, we obtain $\mathbb{P}(X \in \Lambda_n \mid \mathcal{D}_{\text{train}}) \xrightarrow{\mathbb{P}} 1$. Define the event $A_n \coloneqq \big\{|\widehat{q} - 1| \le \delta_n\big\}$. Proposition 3 gives $\mathbb{P}(A_n) \to 1$. For any $x \in \Lambda_n$ on the event A_n, since

$$\mathbb{P}(|Y - \widehat{m}(X)| \le \phi_0(\widehat{m}(X); X) \mid X = x) = 1 - \alpha,$$

Lemma 4 gives

$$\begin{aligned}
&\Big|\mathbb{P}(Y \in \widehat{C}(X) \mid X = x, \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}}) - (1 - \alpha)\Big| \\
&= \Big|\mathbb{P}(|Y - \widehat{m}(X)| \le \widehat{q}\,\widehat{h}(X) \mid X = x, \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}}) \\
&\qquad\qquad - \mathbb{P}(|Y - \widehat{m}(X)| \le \phi_0(\widehat{m}(X); X) \mid X = x, \mathcal{D}_{\text{train}})\Big| \\
&\le 2f_{\max}\big|\widehat{q}\,\widehat{h}(x) - \phi_0(\widehat{m}(x); x)\big| \\
&\le 2f_{\max}\Big[\big|\widehat{q} - 1\big|\,\widehat{h}(x) + \big|\widehat{h}(x) - \phi_0(\widehat{m}(x); x)\big|\Big] \\
&\le 2f_{\max}(h_+\delta_n + \delta_n) = 2f_{\max}(h_+ + 1)\delta_n \to 0.
\end{aligned}$$

Since this bound is uniform in $x \in \Lambda_n$ on A_n and $\mathbb{P}(A_n) \to 1$, (11) follows.

(iii) Asymptotic optimal length. We use the decomposition:

$$\begin{aligned}
&\tfrac{1}{2}\big|\operatorname{Len}(\widehat{C}(X)) - \operatorname{Len}(C^*(X))\,\big| \\
&= \big|\widehat{q}\,\widehat{h}(X) - \phi_0(m^*(X); X)\big| \\
&\le \underbrace{|\widehat{q} - 1|\widehat{h}(X)}_{\text{(a) calibration error}} + \underbrace{|\widehat{h}(X) - \phi_0(\widehat{m}(X); X)|}_{\text{(b) radius-fitting error}} + \underbrace{|\widehat{m}(X) - m^*(X)|}_{\text{(c) center error}},
\end{aligned} \tag{A.2}$$

where the bound in (c) uses the Lipschitz property from Lemma 3. We further decompose (c) into a smoothing-bias and an estimation-error part:

$$|\widehat{m}(X) - m^*(X)| \le |\widehat{m}(X) - m^\dagger_{\beta_n}(X)| + |m^\dagger_{\beta_n}(X) - m^*(X)|. \tag{A.3}$$

All four resulting terms vanish in probability:

- (a): $|\widehat{q} - 1| \xrightarrow{\mathbb{P}} 0$ (Proposition 3); $\widehat{h}(X) \le h_+$ w.h.p. by Assumption 3; product is $o_{\mathbb{P}}(1)$.
- (b): $|\widehat{h}(X) - \phi_0(\widehat{m}(X); X)| \xrightarrow{\mathbb{P}} 0$ by (10) and Lemma 8.
- (c.1): $|\widehat{m}(X) - m^\dagger_{\beta_n}(X)| \xrightarrow{\mathbb{P}} 0$ by (9) and Lemma 8.
- (c.2): $|m^\dagger_{\beta_n}(X) - m^*(X)| \to 0$ P_X-a.s. by Lemma 2.

Hence $\operatorname{Len}(\widehat{C}(X)) - \operatorname{Len}(C^*(X)) \xrightarrow{\mathbb{P}} 0$. Since $\operatorname{Len}(C^*(X)) = 2h^*(X) \ge 2h_{\min} > 0$ a.s. by Lemma 5, dividing yields $\operatorname{Len}(\widehat{C}(X))/\operatorname{Len}(C^*(X)) \xrightarrow{\mathbb{P}} 1$. □

B Sharper Bias Rates and Boundary HDIs

We sharpen Lemma 2 in two regimes: (§B.1) a second-order $O(\beta^2)$ bias when the true HDI is interior with C^2 regularity; (§B.2) a first-order $O(\beta)$ shift when the HDI is pinned at an active support boundary. All statements are pointwise in $X = x$. Throughout, we abbreviate $m^\dagger_\beta := m^\dagger_\beta(x)$, $h^\dagger_\beta := h^\dagger_\beta(x)$, $\ell^\dagger_\beta := m^\dagger_\beta - h^\dagger_\beta$, $u^\dagger_\beta := m^\dagger_\beta + h^\dagger_\beta$. Set $\sigma_K^2 := \int_{\mathbb{R}} t^2\sigma'(t)\,dt$ (which is finite, since σ' has exponential tails).

B.1 Second-Order β-Bias in the Interior HDI Regime

Assumption 4 (Local C^2 smoothness with non-flat endpoints)**.** Fix $x \in \mathcal{X}$. There exists $\delta, M > 0$ such that $f(\cdot \mid x)$ is C^2 on $U_*(\delta) := [\ell^* - \delta, \ell^* + \delta] \cup [u^* - \delta, u^* + \delta]$ with

$$\sup_{y \in U_*(\delta)} |f''(y \mid x)| \le M < \infty.$$

Moreover, $f'(\ell^* \mid x) > 0$ and $f'(u^* \mid x) < 0$.

Lemma 9 (Second-order local error of f_β)**.** *Fix $x \in \mathcal{X}$. Under Assumption 4, there exists $C < \infty$ such that for all sufficiently small β,*

$$\sup_{|y-\ell^*| \le \delta/2} |f_\beta(y \mid x) - f(y \mid x)| + \sup_{|y-u^*| \le \delta/2} |f_\beta(y \mid x) - f(y \mid x)| \le C\beta^2.$$

Proof. Fix z with $|z - \ell^*| \le \delta/2$ (the case near u^* is identical). Decompose

$$f_\beta(z \mid x) - f(z \mid x) = \int (f(z-u \mid x) - f(z \mid x))K_\beta(u)\,du.$$

On $|u| \le \delta/2$, $z - u \in U_*(\delta)$, so we can apply Taylor expansion:

$$f(z-u \mid x) = f(z \mid x) - uf'(z \mid x) + \tfrac{u^2}{2}f''(\xi_{z,u} \mid x)$$

for some $\xi_{z,u}$ between z and $z - u$. By the symmetry of the kernel $\int uK_\beta(u)du = 0$,

$$\begin{aligned}\left|\int_{|u|\le\delta/2} \Big(f(z-u \mid x) - f(z \mid x)\Big) K_\beta(u)\,du\right| &\le \frac{M}{2}\int u^2 K_\beta(u)\,du \\ &= \frac{M}{2}\beta^2\sigma_K^2 \\ &= O(\beta^2).\end{aligned}$$

For $|u| > \delta/2$, use boundedness of $f(y \mid x)$:

$$\begin{aligned}\left|\int_{|u|>\delta/2} \Big(f(z-u \mid x) - f(z \mid x)\Big) K_\beta(u)\,du\right| &\le 2f_{\max}\int_{|u|>\delta/2} K_\beta(u)\,du \\ &= 2f_{\max}\int_{|t|>\delta/(2\beta)} \sigma'(t)\,dt \\ &= o(\beta^2),\end{aligned}$$

since σ' has exponentially decaying tails. Combining the two parts proves the first bound. □

Lemma 10 (Second-order local error of F_β)**.** *Fix $x \in \mathcal{X}$. Under Assumption 4, for all sufficiently small β and uniformly for $|y - \ell^*| \le \delta/2$ or $|y - u^*| \le \delta/2$,*

$$F_\beta(y \mid x) = F(y \mid x) + \tfrac{\beta^2\sigma_K^2}{2}f'(y \mid x) + O(\beta^3).$$

Proof. We apply the same Taylor expansion as in Lemma 9, but to F instead of f, using $F' = f$ and $F'' = f'$ on $U_*(\delta)$. The cubic remainder is $O(\beta^3)$ due to the exponential tails of σ'. □

Proposition 4 (Second-order interior bias)**.** *Fix $x \in \mathcal{X}$. Under Assumptions 1 and 4, $|m^\dagger_\beta - m^*| = O(\beta^2)$ and $|h^\dagger_\beta - h^*| = O(\beta^2)$ as $\beta \to 0^+$.*

Proof. By Lemma 2, $\ell_\beta^\dagger \to \ell^*$ and $u_\beta^\dagger \to u^*$, so for β small both $\ell_\beta^\dagger$ and $u_\beta^\dagger$ lie in the C^2 neighborhoods $U_*(\delta/2)$ of Assumption 4.

Step 1. The smoothed coverage equation reads $F_\beta(u_\beta^\dagger \mid x) - F_\beta(\ell_\beta^\dagger \mid x) = 1-\alpha$, and by definition of the hard center-to-radius map, $F(m_\beta^\dagger + \phi_0(m_\beta^\dagger; x) \mid x) - F(m_\beta^\dagger - \phi_0(m_\beta^\dagger; x) \mid x) = 1-\alpha$. Subtracting the two and applying Lemma 10 to the smoothed terms yields:

$$\begin{aligned}&\big[F(u_\beta^\dagger \mid x) - F(m_\beta^\dagger + \phi_0(m_\beta^\dagger; x) \mid x)\big] - \big[F(\ell_\beta^\dagger \mid x) - F(m_\beta^\dagger - \phi_0(m_\beta^\dagger; x) \mid x)\big] \\ &\quad = -\tfrac{\beta^2\sigma_K^2}{2}\big[f'(u_\beta^\dagger \mid x) - f'(\ell_\beta^\dagger \mid x)\big] + O(\beta^3).\end{aligned}$$

Using $u_\beta^\dagger - (m_\beta^\dagger + \phi_0(m_\beta^\dagger; x)) = h_\beta^\dagger - \phi_0(m_\beta^\dagger; x)$ and similarly for the lower endpoint, Taylor expansion on the LHS gives:

$$\begin{aligned}\text{LHS} = (h_\beta^\dagger - \phi_0(m_\beta^\dagger; x))\big[f(m_\beta^\dagger + \phi_0(m_\beta^\dagger; x) \mid x) + f(m_\beta^\dagger - \phi_0(m_\beta^\dagger; x) \mid x)\big] \\ + O\left((h_\beta^\dagger - \phi_0(m_\beta^\dagger; x))^2\right).\end{aligned}$$

Define

$$A_\beta \coloneqq f(m_\beta^\dagger + \phi_0(m_\beta^\dagger; x) \mid x) + f(m_\beta^\dagger - \phi_0(m_\beta^\dagger; x) \mid x).$$

Since $m_\beta^\dagger \to m^*$ and ϕ_0 is continuous,

$$A_\beta \to 2f(u^* \mid x) > 0,$$

so $A_\beta \geq c_0 > 0$ for some c_0 and β small. The RHS is $O(\beta^2)$. Solving this (and using $h_\beta^\dagger - \phi_0(m_\beta^\dagger; x) \to 0$ to absorb the quadratic correction) yields $|h_\beta^\dagger - \phi_0(m_\beta^\dagger; x)| = O(\beta^2)$.

Step 2. By definition,

$$G_0(m_\beta^\dagger; x) = f(m_\beta^\dagger + \phi_0(m_\beta^\dagger; x) \mid x) - f(m_\beta^\dagger - \phi_0(m_\beta^\dagger; x) \mid x).$$

The smoothed HDI density-equality gives $f_\beta(u_\beta^\dagger \mid x) = f_\beta(\ell_\beta^\dagger \mid x)$. Using the local Lipschitzness of f and

$$|u_\beta^\dagger - (m_\beta^\dagger + \phi_0(m_\beta^\dagger; x))| = |h_\beta^\dagger - \phi_0(m_\beta^\dagger; x)| = O(\beta^2)$$

from Step 1, we decompose:

$$\begin{aligned}G_0(m_\beta^\dagger; x) = &\underbrace{\big[f(m_\beta^\dagger + \phi_0(m_\beta^\dagger; x) \mid x) - f(u_\beta^\dagger \mid x)\big]}_{O(\beta^2)\text{ (Lipschitz + Step 1)}} + \underbrace{\big[f(u_\beta^\dagger \mid x) - f_\beta(u_\beta^\dagger \mid x)\big]}_{O(\beta^2)\text{ (Lemma 9)}} \\ &+ \underbrace{\big[f_\beta(u_\beta^\dagger \mid x) - f_\beta(\ell_\beta^\dagger \mid x)\big]}_{=0} + \underbrace{\big[f_\beta(\ell_\beta^\dagger \mid x) - f(\ell_\beta^\dagger \mid x)\big]}_{O(\beta^2)} \\ &+ \underbrace{\big[f(\ell_\beta^\dagger \mid x) - f(m_\beta^\dagger - \phi_0(m_\beta^\dagger; x) \mid x)\big]}_{O(\beta^2)} = O(\beta^2).\end{aligned}$$

Step 3. For m near m^*, the implicit function theorem applied to

$$H_0(m, h; x) \coloneqq F(m + h \mid x) - F(m - h \mid x) - (1-\alpha)$$

at (m^*, h^*) gives that ϕ_0 is C^1 near m^* with derivative:

$$\phi_0'(m; x) = -\frac{G_0(m; x)}{f(u_0(m; x) \mid x) + f(\ell_0(m; x) \mid x)}.$$

At $m = m^*$, $G_0(m^*; x) = 0$, so $\phi_0'(m^*; x) = 0$. Differentiating G_0 via the chain rule (using $u_0'(m) = 1 + \phi_0'(m)$, $\ell_0'(m) = 1 - \phi_0'(m)$) at $m = m^*$ yields:

$$G_0'(m^*; x) = f'(u^* \mid x) \cdot 1 - f'(\ell^* \mid x) \cdot 1 = f'(u^* \mid x) - f'(\ell^* \mid x) < 0$$

by Assumption 4. Hence there exist $c_G > 0$ and $\eta > 0$ such that

$$|G_0(m; x)| \geq c_G |m - m^*| \quad \text{for} \quad |m - m^*| \leq \eta.$$

Since $m_\beta^\dagger \to m^*$, eventually $|m_\beta^\dagger - m^*| \le \eta$. Combining Steps 2 and 3 gives

$$|m_\beta^\dagger - m^*| \le c_G^{-1}|G_0(m_\beta^\dagger; x)| = O(\beta^2).$$

Then by Step 1 and Lemma 3,

$$|h_\beta^\dagger - h^*| \le |h_\beta^\dagger - \phi_0(m_\beta^\dagger; x)| + |\phi_0(m_\beta^\dagger; x) - \phi_0(m^*; x)| \le O(\beta^2) + |m_\beta^\dagger - m^*| = O(\beta^2).$$

□

B.2 Active Support Boundaries: First-Order β-Shift

Assumption 5 (Active lower-boundary truncation)**.** Fix $x \in \mathcal{X}$. $f(\cdot \mid x)$ has support $[a, b]$, with $b \in (a, \infty]$. On (a, b), $f(\cdot \mid x)$ is continuous, bounded by $f_{\max}$, and unimodal; extended by zero outside. The right-limit $f(a^+ \mid x) \coloneqq \lim_{y \to a^+} f(y \mid x)$ exists and is finite-positive. The unique $(1-\alpha)$-HDI is boundary-pinned: $C^*(x) = [a, u^*(x)]$ with $u^*(x) \in (a, b)$ and $f(a^+ \mid x) > f(u^*(x) \mid x) > 0$.

Proposition 5 (First-order boundary shift)**.** *Fix $x \in \mathcal{X}$. Under Assumption 5, define*

$$\kappa \coloneqq \log\left(\frac{f(u^*(x) \mid x)}{f(a^+ \mid x) - f(u^*(x) \mid x)}\right) \in \mathbb{R}.$$

Then as $\beta \to 0^+$,

$$\ell_\beta^\dagger = a + \kappa\beta + o(\beta), \qquad u_\beta^\dagger = u^*(x) + O(\beta), \qquad \mathrm{Len}(C_\beta^\dagger(x)) - \mathrm{Len}(C^*(x)) = O(\beta),$$

and consequently $|m_\beta^\dagger - m^*| = O(\beta)$, $|h_\beta^\dagger - h^*| = O(\beta)$.

Proof. Drop x from the notation. Mass-binding $\Gamma(m_\beta^\dagger, h_\beta^\dagger) = 1 - \alpha$ together with $F_\beta - F = O(\|f_\beta - f\|_{L^1})$ (Lemma 6) gives

$$F(u_\beta^\dagger) - F(\ell_\beta^\dagger) = 1 - \alpha + O(\|f_\beta - f\|_{L^1}), \tag{B.1}$$

and $\|f_\beta - f\|_{L^1} = O(\beta)$ for any bounded unimodal f, since f has bounded total variation $\le 2f_{\max}$,

$$\|f_\beta - f\|_{L^1} \le 2f_{\max} \int |u| K_\beta(u)\, du = O(\beta).$$

The smoothed endpoint balance is

$$f_\beta(u_\beta^\dagger \mid x) = f_\beta(\ell_\beta^\dagger \mid x). \tag{B.2}$$

We first prove that $\ell_\beta^\dagger \to a$ and $u_\beta^\dagger \to u^*$. By Lemma 5, $h_\beta^\dagger \ge h_{\min} > 0$. Using a similar tightness argument as in Lemma 2, one can show $\ell_\beta^\dagger$ and $u_\beta^\dagger$ remain bounded. Suppose along a subsequence $\ell_\beta^\dagger \to \ell_0 \ge a + \delta$ for some $\delta > 0$. Then $F(\ell_\beta^\dagger) \to F(\ell_0) > 0$, and (B.1) forces $u_\beta^\dagger \to u_0$ with $\int_{\ell_0}^{u_0} f = 1 - \alpha$. Lemma 7 on a compact subset of (a, b) combined with (B.2) gives $f(\ell_0) = f(u_0)$, so $[\ell_0, u_0]$ is an interior $(1-\alpha)$-HDI for f. This contradicts uniqueness in Assumption 5. The case $\ell_\beta^\dagger \le a - \delta$ is also excluded: there $f_\beta(\ell_\beta^\dagger) \to 0$ but $f_\beta(u_\beta^\dagger) \to f(u^*) > 0$, contradicting (B.2). Hence $\ell_\beta^\dagger \to a$, and (B.1) together with strict monotonicity of F on (a, b) yields $u_\beta^\dagger \to u^*$.

For $t \in \mathbb{R}$, substituting $z = a + \beta w$ gives:

$$\begin{aligned} f_\beta(a + \beta t \mid x) &= \int_0^{(b-a)/\beta} f(a + \beta w \mid x)\sigma'(t - w)\, dw \\ &\xrightarrow{\beta \to 0^+} f(a^+ \mid x) \int_0^\infty \sigma'(t - w)\, dw = f(a^+ \mid x)\sigma(t), \end{aligned} \tag{B.3}$$

by dominated convergence.

Define $t_\beta \coloneqq (\ell_\beta^\dagger - a)/\beta$. Then we have $\ell_\beta^\dagger - a \to 0$. We argue $\{t_\beta\}$ is bounded.

If $t_\beta \to -\infty$, we directly bound the integral:

$$f_\beta(\ell_\beta^\dagger) = \int_0^{(b-a)/\beta} f(a+\beta w)\sigma'(t_\beta - w)\,dw \le f_{\max}\int_0^\infty \sigma'(t_\beta - w)\,dw = f_{\max}\sigma(t_\beta).$$

As $t_\beta \to -\infty$, $\sigma(t_\beta) \to 0$, so $f_\beta(\ell_\beta^\dagger) \to 0$. But $f_\beta(u_\beta^\dagger) \to f(u^*) > 0$, contradicting (B.2).

If $t_\beta \to +\infty$ (so $\ell_\beta^\dagger - a \gg \beta$), we apply a boundary-layer argument. We have

$$f_\beta(\ell_\beta^\dagger) = \int f(\ell_\beta^\dagger - \beta w)\sigma'(w)dw.$$

Since $\ell_\beta^\dagger - a \gg \beta$, for any fixed w, $\ell_\beta^\dagger - \beta w > a$ for β small enough, meaning $f(\ell_\beta^\dagger - \beta w) \to f(a^+)$. By dominated convergence, $f_\beta(\ell_\beta^\dagger) \to f(a^+)\int \sigma'(w)dw = f(a^+) > f(u^*)$, which again contradicts (B.2). Therefore $\{t_\beta\}$ is bounded.

Along any convergent subsequence $t_\beta \to t$, from (B.2) and (B.3) we have

$$f_\beta(\ell_\beta^\dagger) \to f(a^+)\sigma(t) = f(u^*).$$

Solving $\sigma(t) = f(u^*)/f(a^+) \in (0,1)$ yields the unique $t = \kappa$, where

$$\kappa = \sigma^{-1}\left(\frac{f(u^*)}{f(a^+)}\right) = \log\left(\frac{f(u^*)}{f(a^+) - f(u^*)}\right).$$

Uniqueness of subsequential limits gives $t_\beta \to \kappa$, i.e., $\ell_\beta^\dagger = a + \kappa\beta + o(\beta)$.

From (B.1), $F(u_\beta^\dagger) - F(\ell_\beta^\dagger) = 1 - \alpha + O(\beta)$. Since F is $f_{\max}$-Lipschitz and $F(y) = 0$ for $y \le a$,

$$|F(\ell_\beta^\dagger) - F(a)| \le f_{\max}(\ell_\beta^\dagger - a)_+ = O(\beta).$$

With $F(a) = 0$ (so $F(u^*) = 1 - \alpha$) we get $F(u_\beta^\dagger) = F(u^*) + O(\beta)$. Strict monotonicity with $F'(u^*) = f(u^*) > 0$ gives local invertibility, so $u_\beta^\dagger = u^* + O(\beta)$. Therefore,

$$\begin{aligned}\mathrm{Len}(C_\beta^\dagger) - \mathrm{Len}(C^*) &= (u_\beta^\dagger - \ell_\beta^\dagger) - (u^* - a)\\ &= (u_\beta^\dagger - u^*) - (\ell_\beta^\dagger - a)\\ &= O(\beta) - (\kappa\beta + o(\beta)) = O(\beta).\end{aligned}$$

Then $m_\beta^\dagger = (\ell_\beta^\dagger + u_\beta^\dagger)/2 = m^* + O(\beta)$ and $h_\beta^\dagger = (u_\beta^\dagger - \ell_\beta^\dagger)/2 = h^* + O(\beta)$. □

Remark 3 (Compatibility with Theorem 2). For P_X-a.e. x, assume either the interior assumptions (Assumption 1) or the boundary assumptions (Assumption 5) hold. Proposition 5 gives $|m_{\beta_n}^\dagger(x) - m^*(x)| = O(\beta_n)$ and $|h_{\beta_n}^\dagger(x) - h^*(x)| = O(\beta_n)$ at active support boundaries. Since $\beta_n \to 0$, the smoothing-bias term (c.2) in (A.2) still vanishes P_X-a.s.; the remaining three terms (a), (b), (c.1) are unchanged. Conditional coverage at level $1-\alpha$ holds verbatim, since its proof in §A.9(ii) only uses the calibration error $|\widehat{q} - 1|$ and the radius-fitting error $|\widehat{h} - \phi_0(\widehat{m}; X)|$, both agnostic to whether the HDI is interior or boundary-pinned. Hence Theorem 2(i)–(iii) hold in the boundary regime as well, with the rate of (iii) governed by β_n rather than β_n^2.

C Experimental Details

Computational Resources. All experiments were conducted on a server equipped with 2x NVIDIA A100 80GB PCIe GPUs, running CUDA 12.6 with driver version 560.35.03. The peak GPU memory usage is approximately 600 MB for the base (sequential) implementation, and increases to approximately 3 GB when using the Fast variant (Appendix D.2) with fold parallelism ($K = 4$).

C.1 Real-world Datasets Details

Table 3 provides a summary of the seven real-world regression datasets used in Section 6.2. These datasets span a variety of domains, sample sizes (n), and feature dimensions (d), offering a comprehensive testbed for evaluating the efficiency and conditional reliability of the methods. All datasets are publicly available under standard open-source licenses, as detailed in the table below.

Table 3: Overview of the real-world datasets used in our main study. n and d represent the number of samples and features, respectively.

Dataset	Description	Samples (n)	Dim (d)	Ref.	License
Bike	Hourly bike sharing demand	10,886	18	[14]	CC BY 4.0
Bio	Protein tertiary structure properties	45,730	9	[37]	CC BY 4.0
Blog	Blog feedback volume prediction	52,397	280	[6]	CC BY 4.0
Facebook-1	Facebook comment volume (Variant 1)	40,948	53	[44]	CC BY 4.0
Facebook-2	Facebook comment volume (Variant 2)	81,311	53	[44]	CC BY 4.0
Homes	King County house sales prices	21,613	19	[22]	CC0
Supercon.	Superconductivity critical temperature	21,263	81	[21]	CC BY 4.0

C.2 Preprocessing

Synthetic data. Covariates are generated as $X \sim \mathrm{Unif}(-2, 2)$ and are used without further normalization. Targets are generated directly from the specified conditional distributions (Section 6.1).

Real data. On each run, we apply the following preprocessing steps fit on the training split only:

1. **Feature standardization:** we standardize each feature to zero mean and unit variance using the training split statistics, and apply the same transform to validation/calibration/test.
2. **Target stabilization:** for heavy-tailed targets, specifically in *Blog*, *Facebook-1*, and *Facebook-2*, we apply dataset-specific standard transformations commonly used in prior work, i.e., $\log(1 + Y)$.
3. **Response rescaling:** following the procedure in [41], we further rescale Y by a constant $c = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(X,Y)\in\mathcal{D}_{\text{train}}} |Y|$ computed on the training split, i.e., $Y \leftarrow Y/c$. This improves numerical conditioning and does not affect coverage; it only rescales interval lengths by a constant shared across methods within each repetition.

C.3 Implementation Details

General training setup and backbone MLP. Unless otherwise specified, all learning-based methods are trained with a batch size of 512 for a maximum of 1000 epochs, employing an early stopping patience of 100 based on validation performance. All neural networks are optimized with Adam. They share the same backbone architecture: a two-hidden-layer MLP with 64 ReLU units per layer and a linear output layer. Method-specific heads enforce required constraints, e.g.:

- Center of CoCP outputs $m(x) \in \mathbb{R}$.
- Radius of CoCP outputs $h(x) > 0$ by applying a Softplus nonlinearity.
- CQR outputs $(\widehat{q}_{\text{lo}}(x), \widehat{q}_{\text{hi}}(x))$ and ensures monotonicity ($\widehat{q}_{\text{lo}} \le \widehat{q}_{\text{hi}}$).

CoCP. We use $K = 4$ fold cross-fitting. For each fold, the radius h is trained on the current fold while the center m is trained on the remaining $K - 1$ folds. The procedure involves an initial MSE warm-up for m, followed by $T = 5$ alternating iterations between h and m, concluding with a final adjustment of h. We set $\beta = 0.005$ and employ exponential learning rate decay for h. For synthetic data, m uses a constant learning rate of 10^{-3}; h decays from $5 \cdot 10^{-4}$ to 0. For real data, m decays from $2 \cdot 10^{-3}$ to 10^{-3}; h decays from $5 \cdot 10^{-4}$ to 0. Note that the MSE warm-up is used only as a practical initialization: in the idealized population dynamics, the alternating updates are well-defined for any finite center, whereas in finite-sample training it places the boundary kernels in data-dense regions and reduces noisy or nearly zero gradients when β is small.

Split. Standard split conformal prediction using absolute residual scores (learning rate 10^{-3}).

CQR and CHR. CQR uses the shared backbone (learning rate 10^{-3}). For CHR, we employ an alternative faster implementation via CIR [20], using 98 interval components and the shared backbone (learning rate 10^{-3}).

C-HDR. We fit a conditional normalizing flow (2 layers, width 64, learning rate $2 \cdot 10^{-3}$). At test time, we approximate HDRs via Monte Carlo sampling using 5000 samples per test covariate. To ensure comparability with interval-based methods, we compute the convex hull of these samples to enforce a single connected prediction interval.

CPCP. We use the split fractions 0.4 and 0.4, bandwidth $\delta = 0.02$, and learning rate $5 \cdot 10^{-4}$, enabling clipping (`use_clip=true`, multiplier 5.0) and mixture augmentation (`use_mix=true`, $\lambda_{\text{mix}} = 0.5$).

RCP. For RCP, we train the conditional quantile function of the nonconformity score on a fraction 0.5 of the calibration set (learning rate 10^{-3}), and we employ the same calibration step as CoCP via the normalized score.

CPL. CPL relies on a user-specified finite-dimensional function class through a feature map $\Phi(x)$, used to define conditional moment constraints. We consider $\Phi(x) = [1,\ x^\top]^\top$ in synthetic data and $\Phi(x) = [1,\ \mathrm{emb}(x)]^\top$ in real data, where $\mathrm{emb}(x)$ is the embedding of x from the shared MLP model. We run 100 outer iterations. In each outer iteration, we train the threshold network for 200 steps on real data (resp. 500 steps on 1D synthetic data), with learning rate 10^{-3}, and we update the nonnegative multipliers with learning rate 0.2. We use smoothing parameter 0.1, length penalty weight 1.0, and split ratio 0.7 for the internal CPL split.

Naive CC. The RKHS-based conditional conformal method of [18] can be extremely computationally expensive in our setting. We therefore include it only on synthetic data, and use a simplified variant that omits test imputation. Under i.i.d. sampling and large sample sizes, this "naive" variant is empirically close to the full method while being significantly more tractable. We use an RBF kernel

$$k_\gamma(x, x') = \exp(-\gamma \|x - x'\|_2^2),$$

with $(\gamma = 0.5, \lambda = 0.00025)$ and a finite-dimensional feature map $\Phi(x)$ as in CPL. In our implementation, the nonnegative interval radius is parameterized as

$$t(x) = \max\{0,\ K(x, X_{\text{cal}})w + \Phi(x)^\top \nu\},$$

where $K(x, X_{\text{cal}})$ is the vector of kernel evaluations against the calibration covariates.

C.4 Metrics and Estimation

Let $\mathcal{D}_{\text{test}} = \{(x_i, y_i)\}_{i=1}^{n_{\text{test}}}$ be the test set, let $\widehat{C}(x) = [\widehat{\ell}(x), \widehat{u}(x)]$ be the predicted interval, and define $z_i = \mathbf{1}\{y_i \in \widehat{C}(x_i)\}$. We evaluate the performance using the following metrics:

Coverage. For synthetic datasets, the conditional distribution is known, so we compute the exact conditional coverage $c_i = \mathbb{P}(Y \in \widehat{C}(x_i) \mid X = x_i)$. For continuous outcomes, $c_i = F(\widehat{u}(x_i) \mid x_i) - F(\widehat{\ell}(x_i) \mid x_i)$, where $F(\cdot \mid x_i)$ is the conditional CDF of $Y \mid X = x_i$. We report

$$\text{Coverage} = \begin{cases} \dfrac{1}{n_{\text{test}}} \displaystyle\sum_{i=1}^{n_{\text{test}}} c_i, & \text{for synthetic datasets,} \\ \dfrac{1}{n_{\text{test}}} \displaystyle\sum_{i=1}^{n_{\text{test}}} z_i, & \text{for real datasets.} \end{cases}$$

Length. We report the average interval width:

$$\text{Length} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \big(\widehat{u}(x_i) - \widehat{\ell}(x_i)\big).$$

ConMAE. For synthetic datasets, we measure the mean absolute deviation of conditional coverage from the target level $1 - \alpha$:

$$\text{ConMAE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \big|c_i - (1 - \alpha)\big|.$$

MSCE. Following Braun et al. [5], we estimate the MSCE using a partition-based estimator. We run K-means on the standardized test covariates to obtain G clusters, with $G = 10$ throughout. Let $\mathcal{I}_g$ be the index set of cluster g, and define $\widehat{c}_g = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} z_i$. We then compute

$$\mathrm{MSCE} = \sum_{g=1}^{G} \frac{|\mathcal{I}_g|}{n_{\mathrm{test}}} \big(\widehat{c}_g - (1-\alpha)\big)^2.$$

WSC. We evaluate WSC following Cauchois et al. [8], using the implementation in `covmetrics` [5]. For $v \in \mathbb{R}^d$ and $a < b$, define the slab

$$S_{v,a,b} = \{x \in \mathbb{R}^d : a \le v^\top x \le b\}.$$

Let $\hat{\mathbb{P}}_n(S) = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathbf{1}\{x_i \in S\}$ denote the empirical mass of a slab S. Given a mass threshold $\delta \in (0, 1]$ (we use $\delta = 0.1$), WSC is the minimum empirical coverage over all scanned slabs with empirical mass at least δ:

$$\mathrm{WSC} = \min_{v \in \mathcal{V}} \min_{\substack{a<b: \\ \hat{\mathbb{P}}_n(S_{v,a,b}) \ge \delta}} \frac{\sum_{i=1}^{n_{\mathrm{test}}} z_i \, \mathbf{1}\{x_i \in S_{v,a,b}\}}{\sum_{i=1}^{n_{\mathrm{test}}} \mathbf{1}\{x_i \in S_{v,a,b}\}}.$$

ERT. We compute ERT using `covmetrics` [5]. ERT treats $\{(x_i, z_i)\}_{i=1}^{n_{\mathrm{test}}}$ as a probabilistic classification problem and compares the risk of the constant predictor $1 - \alpha$ with that of an auditor $a : \mathcal{X} \to [0, 1]$ under a proper scoring rule ℓ. To avoid evaluating the auditor on the same data used for training, ERT is estimated using k-fold ($k = 5$) cross-validation:

$$\ell\text{-ERT} = \frac{1}{n_{\mathrm{test}}} \sum_{j=1}^{k} \sum_{i \in \mathcal{I}_j} \Big[\ell(1-\alpha, z_i) - \ell\big(a^{(-j)}(x_i), z_i\big) \Big],$$

where $\mathcal{I}_1, \ldots, \mathcal{I}_k$ is a partition of the test set into k folds, and $a^{(-j)}$ is trained on all samples outside fold $\mathcal{I}_j$. Larger ERT indicates larger deviations from conditional coverage. We report both ℓ_1-ERT and ℓ_2-ERT, where ℓ_1 is the score in Braun et al. [5] and ℓ_2 is the Brier score,

$$\ell_1(p, z) = \mathrm{sgn}\big(p - (1-\alpha)\big)\big((1-\alpha) - z\big), \qquad \ell_2(p, z) = (p - z)^2,$$

using Logistic Regression on standardized covariates as the auditor; hence ERT is a diagnostic over this linear auditor class rather than a certificate against all nonlinear forms of conditional miscoverage. Due to finite-sample variability, ERT can occasionally be slightly negative; such values should be interpreted as effectively zero.

D Additional Results

D.1 Runtime Comparison

In this section, we compare the computational efficiency of CoCP against the baselines. Table 4 and Table 5 report the average training and inference times (in seconds) across all datasets. Note that the Naive CC method is evaluated only on synthetic benchmarks due to the difficulty of identifying robust hyperparameters for larger real-world datasets; these entries are therefore marked with "-".

As shown in Table 4, CoCP generally requires more training time than standard baselines like Split CP and CQR. For example, on the *Facebook-2* dataset, CoCP takes approximately 1844 seconds to train, compared to 66 seconds for CQR. This increased offline computational cost is an expected trade-off: CoCP relies on an alternating optimization process and K-fold cross-fitting to achieve its geometric interval alignment and superior conditional reliability.

Crucially, this computational overhead is strictly confined to the **offline training phase**. In practical deployments, real-time efficiency is paramount. As demonstrated in Table 5, CoCP's **inference time** is practically negligible (typically < 0.05 seconds). It matches the real-time efficiency of the fastest baselines (e.g., CQR and Split CP) and is orders of magnitude faster than generative methods like C-HDR, which require computationally heavy Monte Carlo sampling at test time.

Overall, the implementation is most appropriate when models are retrained infrequently and prediction is repeated. In settings with frequent retraining or limited compute, the additional offline cost may outweigh the gains in interval efficiency.

Table 4: Comparison of **Training Time** (in seconds) across synthetic and real-world datasets. Results are averaged over 10 random splits (mean ± std).

Dataset	Split	CPL	Naive CC	CQR	CPCP	RCP	C-HDR	CHR	CoCP
Synthetic Datasets									
Normal	18.8 (11.3)	100.7 (12.3)	146.8 (55.7)	18.3 (5.3)	12.1 (3.0)	22.5 (11.1)	23.9 (9.5)	26.2 (8.0)	397.1 (39.3)
LogNormal	18.4 (10.2)	95.4 (10.2)	133.2 (20.5)	17.3 (5.8)	11.2 (3.8)	22.0 (9.2)	19.5 (5.4)	24.7 (7.6)	374.2 (46.3)
Exponential	15.6 (7.3)	93.2 (6.7)	116.8 (23.8)	17.0 (7.5)	9.1 (2.9)	19.0 (7.4)	18.9 (6.2)	23.0 (6.9)	390.6 (35.5)
Real-World Datasets									
Bike	27.7 (3.1)	54.8 (4.2)	-	27.8 (6.9)	4.7 (0.9)	29.3 (3.2)	9.1 (1.1)	30.3 (5.7)	259.4 (22.6)
Bio	134.0 (25.4)	167.8 (26.0)	-	105.0 (19.9)	25.1 (8.2)	145.1 (27.0)	44.4 (9.5)	186.9 (4.8)	1358.4 (93.6)
Blog	30.3 (6.6)	65.8 (9.2)	-	39.6 (6.8)	34.6 (15.1)	37.3 (8.0)	40.0 (6.2)	37.2 (6.8)	1310.8 (209.6)
Facebook-1	26.3 (4.7)	59.7 (5.6)	-	32.2 (4.6)	16.5 (9.9)	31.1 (4.9)	28.6 (1.6)	33.4 (4.4)	819.3 (35.9)
Facebook-2	64.4 (7.9)	97.7 (7.9)	-	65.9 (8.2)	24.9 (10.6)	74.2 (8.2)	57.6 (5.2)	83.5 (20.5)	1844.4 (132.1)
Homes	21.3 (3.2)	55.8 (3.8)	-	21.5 (2.5)	7.9 (2.1)	24.6 (2.9)	17.4 (1.8)	26.2 (5.1)	418.1 (28.1)
Supercon.	48.2 (6.6)	81.9 (6.6)	-	37.6 (9.1)	9.3 (3.1)	52.0 (6.0)	16.1 (1.4)	70.9 (17.4)	681.3 (40.0)

Table 5: Comparison of **Inference Time** (in seconds) across synthetic and real-world datasets. Results are averaged over 10 random splits (mean ± std). CoCP maintains real-time efficiency comparable to the fastest baselines.

Dataset	Split	CPL	Naive CC	CQR	CPCP	RCP	C-HDR	CHR	CoCP
Synthetic Datasets									
Normal	0.02 (0.01)	0.01 (0.01)	0.18 (0.02)	0.00 (0.00)	0.05 (0.08)	0.03 (0.01)	4.04 (3.28)	0.06 (0.02)	0.01 (0.01)
LogNormal	0.01 (0.01)	0.01 (0.01)	0.40 (0.12)	0.00 (0.00)	0.03 (0.02)	0.03 (0.01)	3.74 (3.29)	0.05 (0.01)	0.01 (0.01)
Exponential	0.03 (0.06)	0.00 (0.00)	0.34 (0.07)	0.00 (0.00)	0.03 (0.01)	0.04 (0.05)	2.63 (2.05)	0.07 (0.03)	0.01 (0.01)
Real-World Datasets									
Bike	0.01 (0.00)	0.00 (0.00)	-	0.00 (0.00)	0.01 (0.00)	0.03 (0.05)	0.72 (0.01)	0.03 (0.00)	0.00 (0.00)
Bio	0.03 (0.00)	0.01 (0.01)	-	0.01 (0.01)	0.07 (0.04)	0.05 (0.01)	5.20 (6.98)	0.12 (0.02)	0.00 (0.00)
Blog	0.05 (0.02)	0.03 (0.02)	-	0.01 (0.01)	0.12 (0.03)	0.11 (0.05)	31.53 (24.33)	0.25 (0.14)	0.04 (0.06)
Facebook-1	0.03 (0.00)	0.01 (0.01)	-	0.00 (0.00)	0.08 (0.05)	0.07 (0.07)	9.81 (7.06)	0.11 (0.02)	0.01 (0.01)
Facebook-2	0.06 (0.06)	0.01 (0.01)	-	0.01 (0.00)	0.12 (0.04)	0.13 (0.08)	16.76 (15.42)	0.21 (0.04)	0.01 (0.01)
Homes	0.01 (0.00)	0.01 (0.01)	-	0.00 (0.00)	0.04 (0.01)	0.04 (0.01)	4.19 (4.47)	0.07 (0.01)	0.02 (0.02)
Supercon.	0.02 (0.01)	0.01 (0.01)	-	0.00 (0.00)	0.08 (0.09)	0.03 (0.01)	7.71 (3.40)	0.06 (0.01)	0.02 (0.02)

D.2 Accelerating Training: The Fast Variant

As introduced in Section 4, the base algorithm's K-fold training introduce an $O(K)$ computational overhead. To mitigate this, we propose a **Fast variant** that significantly accelerates training without compromising performance via two key improvements:

1. **Fold Parallelism:** Since training within each fold is mutually independent, we leverage multiprocessing to execute all K-fold simultaneously. This reduces wall-clock time at the cost of a K fold increase in peak GPU memory; sequential execution remains an option for memory-constrained environments.
2. **Strict Early Stopping:** We introduce a minimum improvement threshold ($\Delta_{\min} = 10^{-6}$). Validation loss must improve by at least this amount to reset the patience counter, preventing redundant computation on marginal updates in flat loss regions.

Performance and Speedup. Table 6 presents a head-to-head comparison between the Base implementation and the Fast variant across all benchmarks. The results demonstrate a substantial reduction in training time. On synthetic datasets, the Fast variant achieves a $3.1\times$ to $3.6\times$ speedup. On real-world datasets, the acceleration is even more pronounced, typically ranging from $3.2\times$ to $5.1\times$. For instance, on the *Blog* dataset, training time drops from 1310.8 seconds to just 258.6 seconds, and on the large *Facebook-2* dataset, it decreases from 1844.4 seconds to 462.7 seconds.

The Fast variant yields similar interval lengths and conditional-reliability diagnostics (ConMAE, MSCE, WSC, and ERT) across the evaluated datasets, while inference remains below 0.05 seconds for both implementations. These results confirm that CoCP scales efficiently to large-scale tasks via parallelization and robust early stopping, ensuring high practicality for real-world deployments, although it does not eliminate CoCP's higher offline cost relative to CQR.

Table 6: Comparison between the Base implementation and the Fast variant. Metrics are averaged over 10 random splits (mean ± std). The Fast variant achieves significant training speedups (up to $\approx 5\times$) while strictly preserving interval efficiency and all conditional reliability metrics.

Dataset	Variant	Train Time (s) (↓)	Coverage	Length (↓)	ConMAE / MSCE (↓)	WSC (↑)	ℓ_1-ERT (↓)	ℓ_2-ERT (↓)
Synthetic Datasets (Conditional metric: ConMAE)								
Normal	Base	397.1 (39.3)	0.9007 (0.0037)	1.5977 (0.0153)	0.0066 (0.0013)	-	-	-
	Fast	**109.1** (10.4)	0.9004 (0.0036)	1.5964 (0.0144)	0.0067 (0.0014)	-	-	-
LogNormal	Base	374.2 (46.3)	0.9002 (0.0029)	0.9599 (0.0132)	0.0066 (0.0019)	-	-	-
	Fast	**112.4** (23.0)	0.9002 (0.0026)	0.9602 (0.0099)	0.0058 (0.0015)	-	-	-
Exponential	Base	390.6 (35.5)	0.8975 (0.0026)	1.1277 (0.0121)	0.0077 (0.0022)	-	-	-
	Fast	**124.7** (8.9)	0.8967 (0.0022)	1.1290 (0.0095)	0.0077 (0.0014)	-	-	-
Real-World Datasets (Conditional metrics: MSCE, WSC, ℓ_1-ERT, ℓ_2-ERT)								
Bike	Base	259.4 (22.6)	0.8988 (0.0059)	0.6053 (0.0190)	0.0006 (0.0004)	0.7982 (0.0096)	0.0131 (0.0078)	-0.0001 (0.0004)
	Fast	**80.0** (5.7)	0.8991 (0.0078)	0.6233 (0.0179)	0.0008 (0.0004)	0.7889 (0.0139)	0.0165 (0.0125)	0.0001 (0.0008)
Bio	Base	1358.4 (93.6)	0.9013 (0.0056)	1.4231 (0.0142)	0.0002 (0.0001)	0.8478 (0.0091)	0.0059 (0.0038)	-0.0000 (0.0001)
	Fast	**401.7** (37.0)	0.9011 (0.0042)	1.4152 (0.0162)	0.0002 (0.0001)	0.8485 (0.0084)	0.0030 (0.0054)	-0.0000 (0.0001)
Blog	Base	1310.8 (209.6)	0.8994 (0.0046)	2.7025 (0.0897)	0.0004 (0.0001)	0.8416 (0.0069)	0.0177 (0.0036)	-0.0016 (0.0007)
	Fast	**258.6** (16.9)	0.8990 (0.0035)	2.7180 (0.1107)	0.0004 (0.0001)	0.8393 (0.0077)	0.0200 (0.0033)	-0.0015 (0.0006)
Facebook-1	Base	819.3 (35.9)	0.8998 (0.0047)	1.6380 (0.0232)	0.0002 (0.0001)	0.8470 (0.0074)	0.0056 (0.0051)	-0.0004 (0.0003)
	Fast	**218.4** (22.5)	0.8989 (0.0047)	1.6338 (0.0411)	0.0002 (0.0000)	0.8469 (0.0068)	0.0072 (0.0049)	-0.0004 (0.0004)
Facebook-2	Base	1844.4 (132.1)	0.8988 (0.0024)	1.5376 (0.0141)	0.0001 (0.0001)	0.8547 (0.0072)	0.0103 (0.0023)	-0.0001 (0.0001)
	Fast	**462.7** (45.0)	0.8981 (0.0027)	1.5308 (0.0181)	0.0001 (0.0001)	0.8547 (0.0068)	0.0109 (0.0031)	-0.0001 (0.0002)
Homes	Base	418.1 (28.1)	0.9017 (0.0065)	0.5242 (0.0087)	0.0003 (0.0002)	0.8283 (0.0145)	0.0052 (0.0070)	-0.0003 (0.0003)
	Fast	**111.8** (9.8)	0.9003 (0.0064)	0.5222 (0.0090)	0.0003 (0.0001)	0.8272 (0.0097)	0.0039 (0.0077)	-0.0004 (0.0003)
Supercon.	Base	681.3 (40.0)	0.9041 (0.0044)	0.8167 (0.0156)	0.0003 (0.0002)	0.8245 (0.0069)	0.0127 (0.0091)	-0.0006 (0.0006)
	Fast	**167.6** (14.9)	0.9034 (0.0054)	0.8134 (0.0135)	0.0004 (0.0001)	0.8247 (0.0098)	0.0114 (0.0070)	-0.0009 (0.0005)

D.3 Sensitivity Analysis and Ablation

We evaluate CoCP's robustness and core components using the Fast variant. We analyze three hyperparameters: smoothing temperature β, number of folds K, and alternating updates T, varying one while fixing the others at the default $(\beta, K, T) = (0.005, 4, 5)$.

We report Length and ConMAE for synthetic benchmarks. For clarity, real-world evaluation focuses on the four largest datasets (*Bio*, *Blog*, *Facebook-1*, *Facebook-2*), reporting Length, MSCE, WSC, and ℓ_1-ERT (ℓ_2-ERT is omitted due to identical trends). In all figures, color-coded dashed lines denote the CQR baseline, providing a fixed reference for standard center-invariant intervals.

Temperature parameter β. We evaluate $\beta \in \{0.002, 0.005, 0.01, 0.02, 0.05\}$. Figure 4 demonstrates CoCP's resilience to the smoothing temperature. Performance is highly stable across $\beta \in [0.002, 0.02]$, consistently outperforming CQR. However, excessive smoothing ($\beta = 0.05$) distorts the geometric boundary signal, causing intervals to widen (visible as a U-shape in Length on real datasets) and conditional reliability to degrade (e.g., inflated ConMAE under Exponential noise). The default $\beta = 0.005$ strikes an optimal balance, ensuring sufficient gradient flow without compromising geometric precision.

Number of folds K. Figure 5 evaluates $K \in \{2, 3, 4, 5\}$, which dictates the $(K-1) : 1$ data allocation for center versus radius updates. While synthetic benchmarks show mild sensitivity, real-world datasets exhibit clear gains as K increases to 4: interval lengths drop noticeably (e.g., on *Blog* and *Facebook-1*), and conditional metrics strictly dominate CQR. Performance plateaus at $K = 5$, confirming $K = 4$ as a robust default that balances data utilization with ensemble variance reduction.

Alternating iterations T (Ablation). Figure 6 evaluates $T \in \{0, 1, 2, 3, 4, 5\}$, serving as a crucial ablation of the center-refinement step. At $T = 0$ (radius learning only), CoCP yields interval lengths comparable to or even worse than CQR. Under symmetric Normal noise, increasing T has negligible impact since the initial center that estimates the conditional mean is already near-optimal. However, under skewed noise and across all four real-world datasets, enabling alternating updates ($T > 0$) triggers a sharp reduction in interval length. Although conditional reliability metrics may exhibit initial fluctuations, they stabilize within 3 to 5 iterations—maintaining or improving upon the $T = 0$ performance while consistently outperforming the CQR reference. This definitively validates CoCP's core premise: co-optimizing translation and scaling via the push-pull mechanism is fundamentally superior to the fixed-center paradigm.

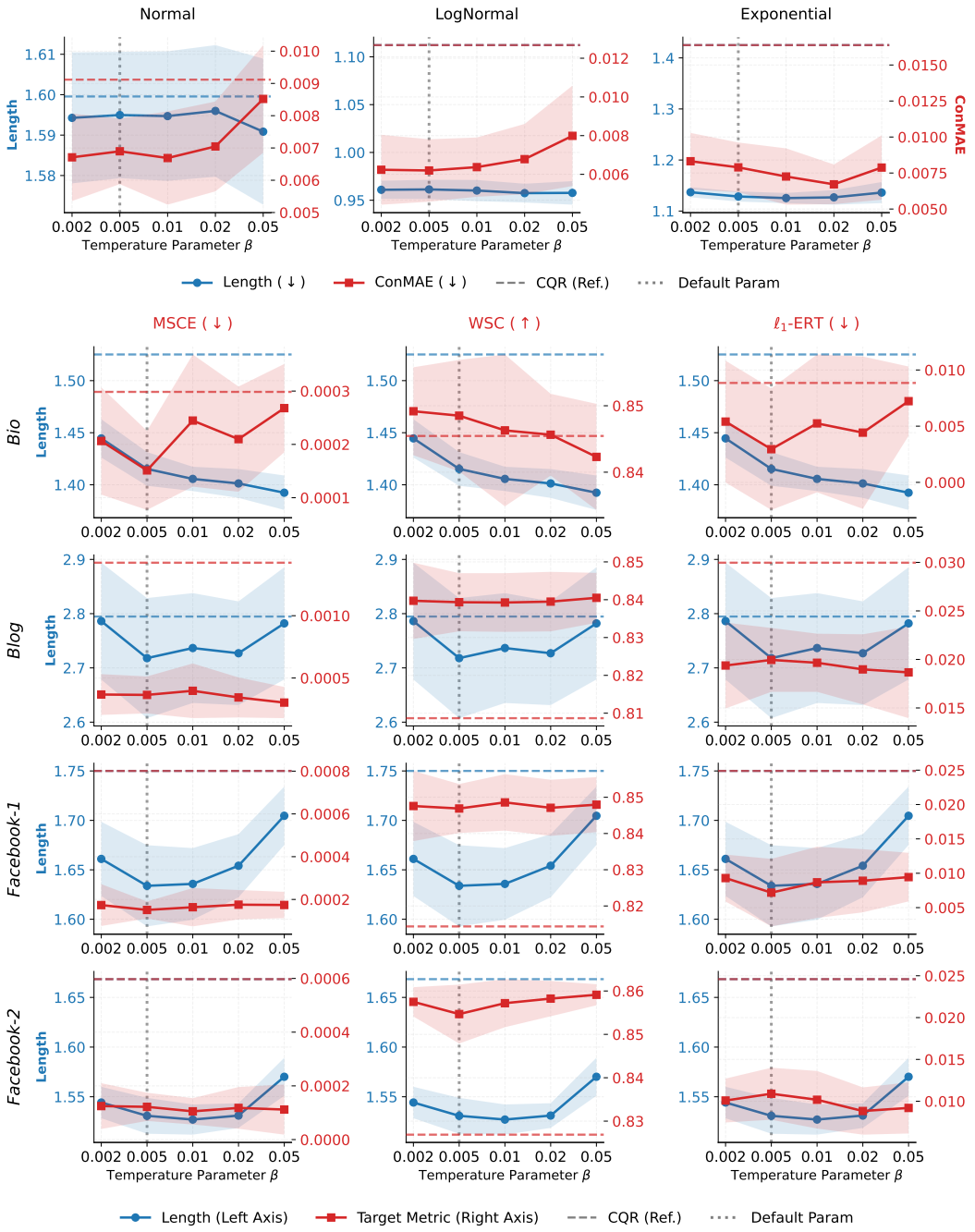

Figure 4: Sensitivity to the smoothing temperature $\beta \in \{0.002, 0.005, 0.01, 0.02, 0.05\}$ using the Fast variant. **Top:** Synthetic benchmarks (Length and ConMAE). **Bottom:** The four largest real datasets (Length, MSCE, WSC, and ℓ_1-ERT). The solid lines show CoCP performance, dashed lines denote the CQR reference (color-coded by axis), and the gray dotted line marks the default $\beta = 0.005$. CoCP remains stable and outperforms CQR across $\beta \leq 0.02$, with geometric distortion and performance degradation occurring only under excessive smoothing ($\beta = 0.05$).

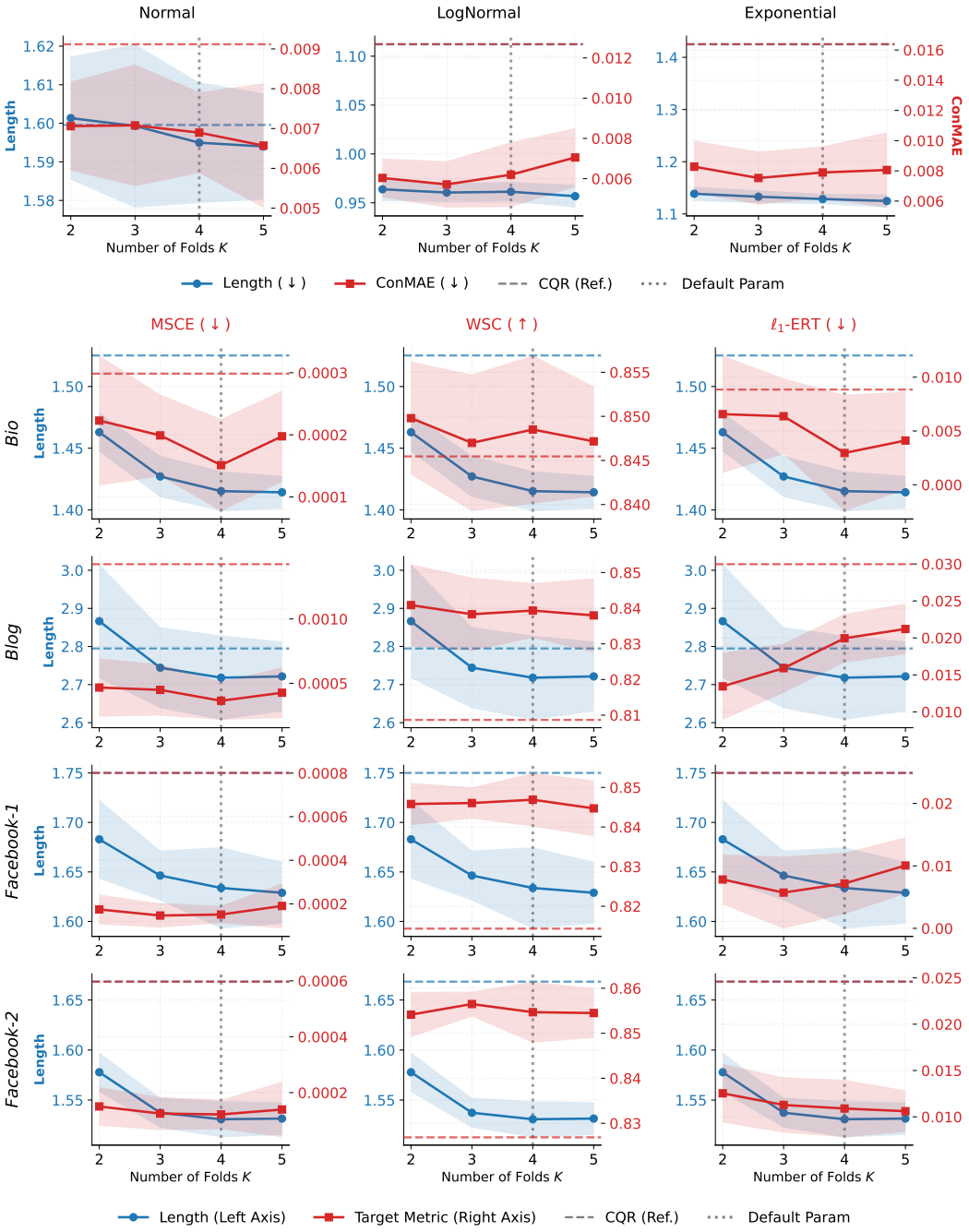

Figure 5: Sensitivity to the number of folds $K \in \{2, 3, 4, 5\}$ using the Fast variant. **Top:** Synthetic benchmarks. **Bottom:** The four largest real datasets. The gray dotted line marks the default $K = 4$. Increasing K from 2 to 4 yields clear efficiency gains and stabilizes conditional reliability, particularly on real datasets, with performance saturating at $K = 4 \sim 5$.

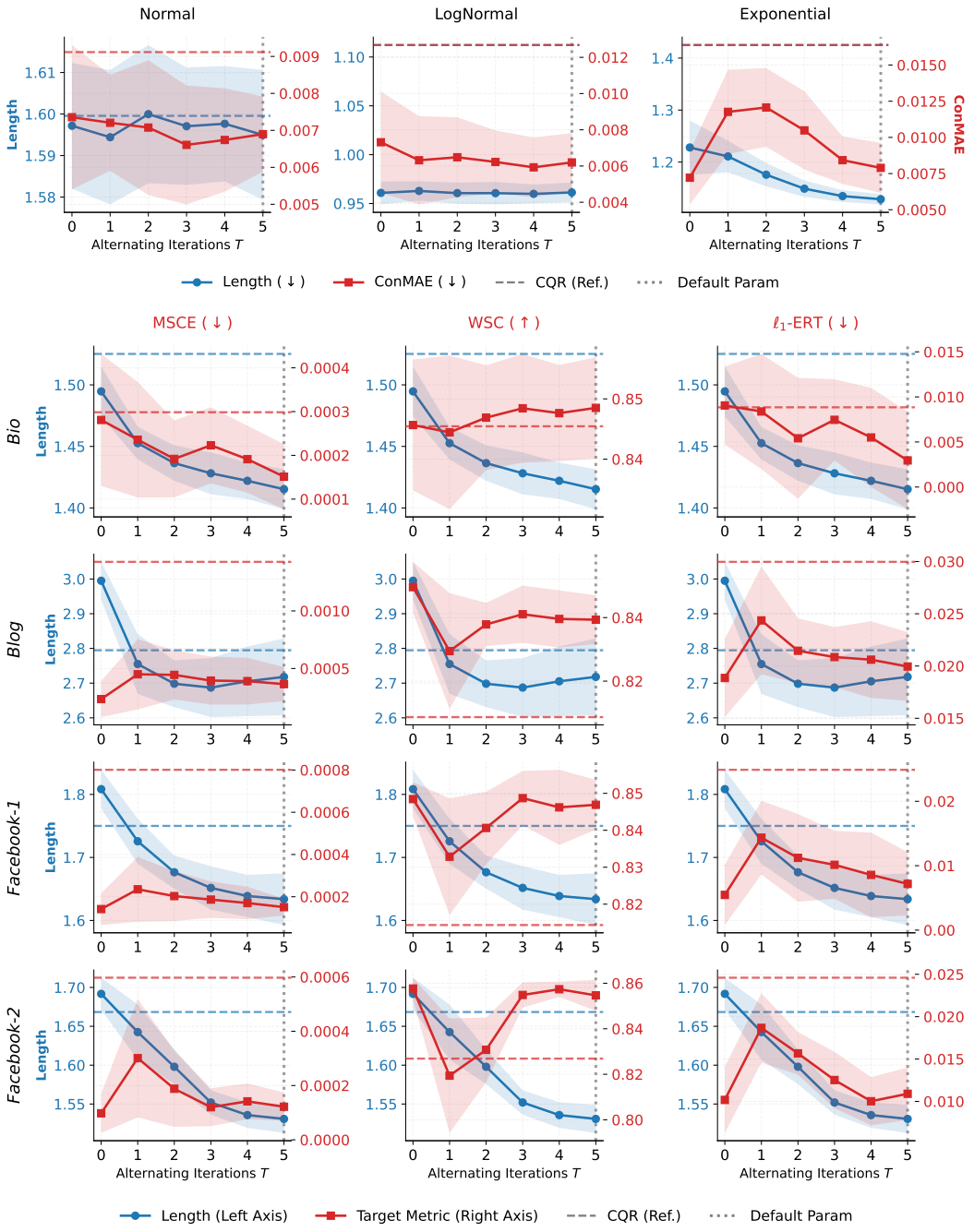

Figure 6: Sensitivity to alternating updates $T \in \{0, 1, 2, 3, 4, 5\}$ (Fast variant), serving as an ablation of the center-refinement step. **Top:** Synthetic benchmarks. **Bottom:** The four largest real datasets. The $T = 0$ case represents the no-refinement baseline (learning radius only). While $T = 0$ shows no significant advantage over the CQR reference (dashed lines) in interval length, enabling alternating updates ($T > 0$) yields substantial gains. Performance typically plateaus within 3 to 5 iterations, at which point CoCP consistently outperforms CQR across all metrics.

D.4 Sample-Adaptive Smoothing Strategy

While a global smoothing parameter β performs robustly on average (Appendix D.3), it requires tuning based on the scale of the target variable and may be suboptimal under severe heteroscedasticity. A fixed β can under-smooth wide intervals (vanishing gradients) and over-smooth narrow ones (distorting the indicator geometry). To achieve scale invariance, we propose a sample-adaptive strategy that scales $\beta(x)$ proportionally to the local interval radius, preserving a consistent geometric shape across the covariate space.

Geometric preservation and closed-form approximation. Let $\tilde{\mathbf{1}}_\beta(y; m, h) = \sigma\big(\frac{h-(y-m)}{\beta}\big) - \sigma\big(\frac{-h-(y-m)}{\beta}\big)$ denote the smoothed interval indicator. To maintain a "flat" interior, we require $\tilde{\mathbf{1}}_\beta$ to retain a high value of $1-\epsilon$ (for a small tolerance $\epsilon > 0$) at a fractional distance $p \in (0,1)$ of the radius h from the center m. Evaluating $\tilde{\mathbf{1}}_\beta(m + ph; m, h) = 1 - \epsilon$ yields:

$$\sigma\left(\frac{(1-p)h}{\beta}\right) - \sigma\left(-\frac{(1+p)h}{\beta}\right) = 1 - \epsilon. \tag{D.1}$$

In practice, $\beta \ll h$, making the second sigmoid term exponentially vanishing (≈ 0). Omitting it provides a highly accurate closed-form approximation for the adaptive smoothing parameter:

$$\beta(x) \approx \frac{(1-p)h(x)}{\log\big(\frac{1-\epsilon}{\epsilon}\big)}. \tag{D.2}$$

The truncation error is bounded by $O(\epsilon^{(1+p)/(1-p)})$ (e.g., $\sim 10^{-18}$ for $p = 0.8, \epsilon = 0.01$), rendering this closed-form expression virtually exact.

Implementation via QR warm-up. Since the optimal radius $h(x)$ is unknown a priori, we approximate it using a lightweight Quantile Regression (QR) warm-up. We first train a base QR model to estimate initial bounds $\widehat{q}_{\alpha/2}(x)$ and $\widehat{q}_{1-\alpha/2}(x)$, yielding a proxy radius $\widehat{h}(x) = [\widehat{q}_{1-\alpha/2}(x) - \widehat{q}_{\alpha/2}(x)]/2$. For each sample x_i, we then compute the instance-dependent $\beta(x_i)$ using $\widehat{h}(x_i)$ and proceed with the standard CoCP optimization. The complete procedure is summarized in Algorithm 2. Note that this fixed-(p, ϵ) adaptive rule is a practical scale-normalized heuristic. To strictly satisfy the vanishing schedule $\beta_n \to 0$ required by Theorem 2, one can set $\beta_{i,n} = c_n \widehat{h}(x_i)$ with $c_n \to 0$.

Algorithm 2 Sample-Adaptive Smoothing Strategy for CoCP

Require: Training data $\mathcal{D}_{\text{train}}$, calibration data $\mathcal{D}_{\text{cal}}$, miscoverage $\alpha \in (0,1)$, folds K, iterations T, interior proportion $p \in (0,1)$, tolerance $\epsilon > 0$

Ensure: Calibrated interval function $\widehat{C}(\cdot)$

1: **Phase 1: QR Warm-up**

2: Train a base quantile regression model on $\mathcal{D}_{\text{train}}$ to estimate $\widehat{q}_{\alpha/2}(x)$ and $\widehat{q}_{1-\alpha/2}(x)$.

3: For each $(X_i, Y_i) \in \mathcal{D}_{\text{train}}$, compute the proxy radius:

$$\widehat{h}(X_i) = \frac{1}{2}\left(\widehat{q}_{1-\alpha/2}(X_i) - \widehat{q}_{\alpha/2}(X_i)\right)$$

4: Compute the instance-dependent smoothing parameter:

$$\beta_i = \frac{(1-p)\widehat{h}(X_i)}{\log((1-\epsilon)/\epsilon)}$$

5: **Phase 2: Adaptive Alternating Optimization**

6: Execute the K-fold alternating optimization (Lines 1-11 in Algorithm 1), replacing the global β with the sample-specific β_i in the smooth pinball loss (3) and smooth interval loss (4).

7: **Phase 3: Calibration**

8: Compute nonconformity scores and calibrate on $\mathcal{D}_{\text{cal}}$ (Lines 12-15 in Algorithm 1).

9: **return** $\widehat{C}(x)$

Empirical evaluation. We fix the tolerance $\epsilon = 0.01$ and evaluate the interior proportion $p \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ using the Fast variant of CoCP. Table 7 summarizes the results across all synthetic and real-world benchmarks.

The adaptive strategy yields performance that is highly competitive with—and occasionally superior to—the well tuned fixed-β baseline. On complex and skewed distributions (e.g., Exponential, *Blog*, *Facebook*), the adaptive formulation effectively enhances conditional reliability (lower MSCE/ℓ_1-ERT, higher WSC) while maintaining tight intervals. Crucially, this strategy renders CoCP invariant to the absolute scale of the target variable, effectively alleviating the burden of hyperparameter tuning. The performance remains remarkably stable across different choices of p, establishing $p \in [0.6, 0.8]$ (e.g., $p = 0.7$) as a robust, plug-and-play default for practical deployments.

Table 7: Performance comparison of the Fixed β strategy versus the Sample-Adaptive Smoothing strategy across different interior proportions p. Best results per row are highlighted in **bold**. Values in blue indicate an improvement over the Fixed β baseline. The adaptive strategy matches or frequently exceeds the baseline, offering a robust, scale-invariant default.

Dataset	Metric	Fixed β (Fast CoCP)	Adaptive β (Fast CoCP)				
			$p = 0.5$	$p = 0.6$	$p = 0.7$	$p = 0.8$	$p = 0.9$
Synthetic Datasets							
Normal	Length (↓)	1.5964 (0.0144)	1.5978 (0.0159)	1.5978 (0.0162)	1.5959 (0.0167)	**1.5938** (0.0161)	1.5953 (0.0170)
	ConMAE (↓)	0.0067 (0.0014)	0.0064 (0.0013)	**0.0063** (0.0013)	0.0065 (0.0011)	0.0067 (0.0014)	0.0072 (0.0014)
LogNormal	Length (↓)	0.9602 (0.0099)	0.9603 (0.0118)	0.9599 (0.0117)	0.9595 (0.0111)	0.9599 (0.0105)	**0.9581** (0.0096)
	ConMAE (↓)	**0.0058** (0.0015)	0.0065 (0.0020)	0.0063 (0.0015)	0.0064 (0.0016)	0.0059 (0.0013)	0.0059 (0.0015)
Exponential	Length (↓)	1.1290 (0.0095)	1.1632 (0.0312)	1.1513 (0.0318)	1.1403 (0.0265)	1.1294 (0.0267)	**1.1209** (0.0147)
	ConMAE (↓)	0.0077 (0.0014)	0.0068 (0.0014)	0.0067 (0.0013)	**0.0065** (0.0016)	0.0071 (0.0020)	0.0077 (0.0018)
Real-World Datasets							
Bike	Length (↓)	0.6233 (0.0179)	0.6136 (0.0157)	0.6077 (0.0119)	**0.6014** (0.0166)	0.6082 (0.0181)	0.6157 (0.0216)
	MSCE (↓)	0.0008 (0.0004)	0.0009 (0.0005)	0.0009 (0.0004)	0.0008 (0.0004)	**0.0007** (0.0002)	0.0007 (0.0003)
	WSC (↑)	0.7889 (0.0139)	0.7893 (0.0201)	0.7897 (0.0182)	0.7796 (0.0191)	**0.7961** (0.0149)	0.7940 (0.0134)
	ℓ_1-ERT (↓)	0.0165 (0.0125)	0.0185 (0.0075)	0.0175 (0.0095)	0.0155 (0.0110)	**0.0132** (0.0118)	0.0150 (0.0080)
	ℓ_2-ERT (↓)	0.0001 (0.0008)	0.0004 (0.0007)	0.0003 (0.0009)	0.0001 (0.0010)	**-0.0000** (0.0007)	-0.0000 (0.0006)
Bio	Length (↓)	1.4152 (0.0162)	**1.3923** (0.0139)	1.3972 (0.0146)	1.3975 (0.0147)	1.3960 (0.0165)	1.4047 (0.0179)
	MSCE (↓)	**0.0002** (0.0001)	0.0002 (0.0001)	0.0003 (0.0001)	0.0002 (0.0001)	0.0002 (0.0001)	0.0002 (0.0001)
	WSC (↑)	**0.8485** (0.0084)	0.8443 (0.0103)	0.8464 (0.0092)	0.8435 (0.0105)	0.8458 (0.0086)	0.8455 (0.0088)
	ℓ_1-ERT (↓)	**0.0030** (0.0054)	0.0057 (0.0067)	0.0071 (0.0054)	0.0064 (0.0046)	0.0050 (0.0060)	0.0047 (0.0048)
	ℓ_2-ERT (↓)	**0.0000** (0.0001)	0.0000 (0.0001)	0.0001 (0.0002)	0.0000 (0.0001)	-0.0000 (0.0001)	-0.0000 (0.0001)
Blog	Length (↓)	**2.7180** (0.1107)	2.8474 (0.0913)	2.8348 (0.1151)	2.8057 (0.1110)	2.7814 (0.0990)	2.7317 (0.1062)
	MSCE (↓)	0.0004 (0.0001)	0.0003 (0.0001)	**0.0003** (0.0001)	0.0003 (0.0001)	0.0004 (0.0001)	0.0004 (0.0001)
	WSC (↑)	0.8393 (0.0077)	0.8432 (0.0072)	**0.8434** (0.0083)	0.8432 (0.0101)	0.8404 (0.0087)	0.8402 (0.0089)
	ℓ_1-ERT (↓)	0.0200 (0.0033)	**0.0163** (0.0035)	0.0181 (0.0044)	0.0171 (0.0038)	0.0189 (0.0042)	0.0195 (0.0045)
	ℓ_2-ERT (↓)	-0.0015 (0.0006)	-0.0015 (0.0007)	-0.0015 (0.0007)	**-0.0017** (0.0005)	-0.0014 (0.0010)	-0.0014 (0.0009)
Facebook-1	Length (↓)	**1.6338** (0.0411)	1.7040 (0.0305)	1.6969 (0.0286)	1.6802 (0.0313)	1.6652 (0.0410)	1.6450 (0.0305)
	MSCE (↓)	0.0002 (0.0000)	**0.0001** (0.0001)	0.0001 (0.0001)	0.0001 (0.0001)	0.0002 (0.0001)	0.0002 (0.0001)
	WSC (↑)	0.8469 (0.0068)	0.8478 (0.0077)	**0.8498** (0.0058)	0.8491 (0.0068)	0.8476 (0.0065)	0.8486 (0.0066)
	ℓ_1-ERT (↓)	**0.0072** (0.0049)	0.0075 (0.0042)	0.0079 (0.0038)	0.0081 (0.0050)	0.0078 (0.0042)	0.0087 (0.0038)
	ℓ_2-ERT (↓)	-0.0004 (0.0004)	-0.0004 (0.0003)	-0.0004 (0.0004)	-0.0004 (0.0004)	**-0.0005** (0.0003)	-0.0004 (0.0003)
Facebook-2	Length (↓)	**1.5308** (0.0181)	1.5656 (0.0164)	1.5574 (0.0161)	1.5519 (0.0108)	1.5405 (0.0167)	1.5322 (0.0155)
	MSCE (↓)	0.0001 (0.0001)	0.0001 (0.0001)	**0.0001** (0.0000)	0.0001 (0.0001)	0.0001 (0.0001)	0.0001 (0.0001)
	WSC (↑)	0.8547 (0.0068)	**0.8598** (0.0040)	0.8584 (0.0031)	0.8590 (0.0032)	0.8592 (0.0034)	0.8589 (0.0033)
	ℓ_1-ERT (↓)	0.0109 (0.0031)	0.0105 (0.0032)	**0.0096** (0.0030)	0.0096 (0.0024)	0.0097 (0.0029)	0.0106 (0.0032)
	ℓ_2-ERT (↓)	-0.0001 (0.0002)	-0.0001 (0.0001)	**-0.0002** (0.0001)	-0.0001 (0.0001)	-0.0001 (0.0001)	-0.0001 (0.0002)
Homes	Length (↓)	**0.5222** (0.0090)	0.5279 (0.0098)	0.5278 (0.0108)	0.5292 (0.0070)	0.5291 (0.0119)	0.5269 (0.0092)
	MSCE (↓)	**0.0003** (0.0001)	0.0004 (0.0002)	0.0004 (0.0001)	0.0004 (0.0002)	0.0003 (0.0001)	0.0004 (0.0002)
	WSC (↑)	**0.8272** (0.0097)	0.8271 (0.0136)	0.8235 (0.0150)	0.8250 (0.0150)	0.8250 (0.0097)	0.8230 (0.0137)
	ℓ_1-ERT (↓)	0.0039 (0.0077)	0.0064 (0.0064)	0.0059 (0.0055)	0.0038 (0.0075)	0.0048 (0.0073)	**0.0029** (0.0073)
	ℓ_2-ERT (↓)	-0.0004 (0.0003)	-0.0004 (0.0003)	-0.0003 (0.0003)	-0.0004 (0.0004)	-0.0004 (0.0004)	-0.0004 (0.0004)
Supercon.	Length (↓)	0.8134 (0.0135)	**0.7618** (0.0179)	0.7696 (0.0149)	0.7698 (0.0140)	0.7779 (0.0126)	0.7932 (0.0102)
	MSCE (↓)	0.0004 (0.0001)	0.0004 (0.0001)	0.0003 (0.0002)	**0.0003** (0.0002)	0.0004 (0.0002)	0.0004 (0.0001)
	WSC (↑)	0.8247 (0.0098)	0.8252 (0.0098)	**0.8275** (0.0081)	0.8251 (0.0082)	0.8262 (0.0137)	0.8258 (0.0081)
	ℓ_1-ERT (↓)	0.0114 (0.0070)	0.0129 (0.0043)	**0.0099** (0.0074)	0.0129 (0.0053)	0.0128 (0.0059)	0.0126 (0.0048)
	ℓ_2-ERT (↓)	**-0.0009** (0.0005)	-0.0006 (0.0005)	-0.0006 (0.0004)	-0.0005 (0.0004)	-0.0007 (0.0004)	-0.0007 (0.0006)

D.5 Center Refinement and Point Prediction

Because the learned center $m(x)$ targets the HDI midpoint rather than the conditional mean, it need not serve as a point predictor. Figure 7 separates the two objectives across all ten benchmarks. The

$T=0$ control shares the MSE initialization, fold allocation, radius fitting, cross-fold averaging, and calibration of full CoCP, but omits center refinement. It already matches or slightly improves upon the separately trained Split predictor in RMSE (left panel). Raising T to 5 shifts the center toward the interval-optimal target, shortening the interval on nine of ten datasets (middle panel). The largest reductions appear under skewness (*Exponential*: -6.9%) and on larger real benchmarks (*Facebook-1*: -9.9%, *Blog*: -8.1%). Conditional error (right panel) improves or holds steady on most datasets; the few apparent increases occur where MSCE is already near its floor (e.g., *Facebook-2*: absolute MSCE stays below 10^{-4}). When point-prediction fidelity matters, the $T=0$ center or a separate predictor can be retained for point decisions, while CoCP is used solely for uncertainty quantification.

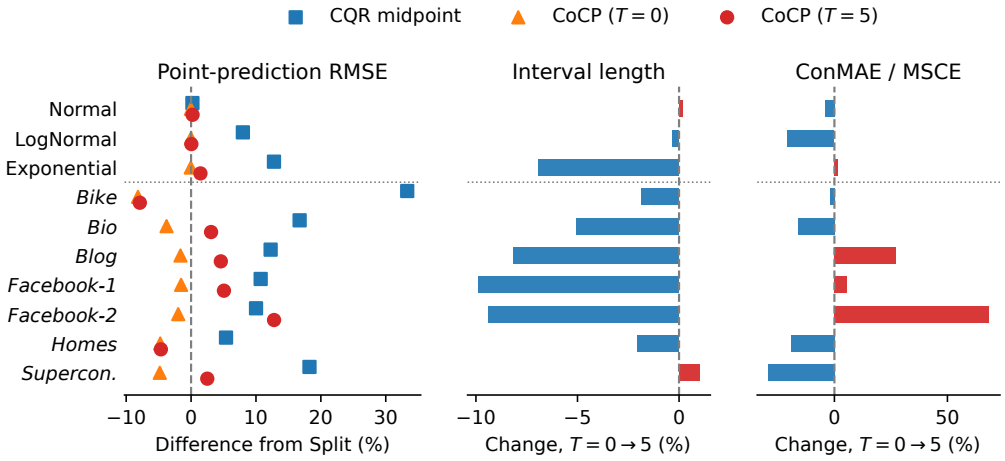

Figure 7: Point-prediction accuracy versus interval efficiency under center refinement ($\alpha = 0.1$). **Left:** RMSE of each center relative to the Split predictor (zero line = parity; negative values indicate lower RMSE). **Middle:** percentage change in interval length from $T=0$ (no refinement) to $T=5$. **Right:** percentage change in conditional error (ConMAE for synthetic, MSCE for real data). The dotted line separates synthetic (above) from real (below) datasets.

D.6 Fold Variability and Local Residual Modality

CoCP averages K fold-specific predictors. Figure 8 reports the normalized spread among them at $(\alpha, K, T) = (0.1, 4, 5)$, averaged over ten splits. On synthetic data, center disagreement is 1.8–2.4% of interval length and radius disagreement is 3.0–4.8% of half-length, small enough that ensembling provides only marginal stability gains. On real data the corresponding ranges rise to 8–18% and 16–25%, consistent with the stronger benefit of $K \geq 4$ observed in Figure 5.

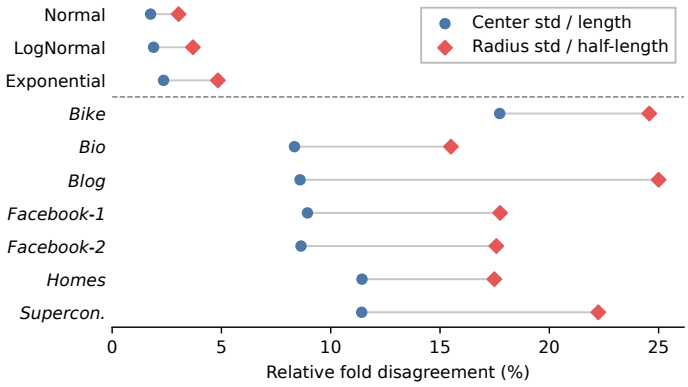

Figure 8: Relative fold disagreement at $(\alpha, K, T) = (0.1, 4, 5)$, averaged over ten splits. Center standard deviation is normalized by mean interval length; radius standard deviation by mean half-length. The dashed line separates synthetic (above) from real (below) datasets.

To assess whether widespread residual multimodality might explain the larger fold variability on real data, we apply Hartigan's dip test locally. On each split, signed residuals $Y - \widehat{g}(X)$ from the MSE predictor are formed, covariates are standardized and projected onto the first $q = \min(d, 32)$ principal components, and the dip test is applied to the $k = 200$ nearest calibration residuals of 100 test anchors. Table 8 shows that raw rejection rates never exceed 0.015; all Benjamini–Hochberg and Benjamini–Yekutieli adjusted rates are zero. The conclusion is stable across $k \in \{100, 200, 400\}$ and $q \in \{16, 32, 64\}$. Failure to reject does not establish unimodality, since each neighborhood pools nearby but non-identical covariates. These findings are consistent with finite-sample estimator variability contributing to the larger fold disagreement on real data, without ruling out departures from conditional unimodality.

Table 8: Hartigan's dip test for local residual unimodality ($q = \min(d, 32)$ PCA components, $k = 200$ nearest neighbors). Mean $\pm$ std over ten splits.

Dataset	Median dip	Raw rejection	BH rejection	BY rejection
Bike	0.0171	0.000±0.000	0.000±0.000	0.000±0.000
Bio	0.0181	0.003±0.007	0.000±0.000	0.000±0.000
Blog	0.0190	0.008±0.012	0.000±0.000	0.000±0.000
Facebook-1	0.0196	0.003±0.007	0.000±0.000	0.000±0.000
Facebook-2	0.0202	0.015±0.015	0.000±0.000	0.000±0.000
Homes	0.0175	0.000±0.000	0.000±0.000	0.000±0.000
Supercon.	0.0176	0.000±0.000	0.000±0.000	0.000±0.000

D.7 Comparison with Asymmetric Tail Allocation

An alternative to translating the center is to optimize a single global split of the two tail probabilities. We implement this baseline (A-CQR) with intervals

$$[\widehat{q}_{\lambda\alpha}(x),\, \widehat{q}_{1-(1-\lambda)\alpha}(x)], \quad \lambda \in \{0.1, \ldots, 0.9\}.$$

Candidate models are fitted on the training split, the validation split selects the λ minimizing conformalized length, and the selected model is recalibrated on the independent calibration split. Table 9 compares A-CQR with CoCP at $\alpha = 0.1$. A-CQR already improves markedly over equal-tailed CQR on the skewed benchmarks. CoCP is shorter on nine of ten datasets (*Blog* excepted) and achieves lower conditional error on all ten, consistent with its ability to adapt the center locally as a function of x rather than selecting a single global λ.

Table 9: Comparison with validation-tuned A-CQR at $\alpha = 0.1$ (mean over ten splits). The conditional metric is ConMAE on synthetic data and MSCE on real data.

	Coverage		Length (↓)		ConMAE / MSCE (↓)	
Dataset	A-CQR	**CoCP**	A-CQR	**CoCP**	A-CQR	**CoCP**
Normal	0.9010	0.9002	1.5967	**1.5951**	0.0098	**0.0072**
LogNormal	0.9000	0.9006	0.9622	**0.9606**	0.0083	**0.0065**
Exponential	0.8976	0.8968	1.1575	**1.1278**	0.0121	**0.0079**
Bike	0.9045	0.8999	0.7485	**0.6230**	0.0018	**0.0008**
Bio	0.8997	0.9011	1.4851	**1.4162**	**0.0002**	**0.0002**
Blog	0.8989	0.8986	**2.4652**	2.7090	0.0006	**0.0004**
Facebook-1	0.8984	0.9012	1.6618	**1.6389**	0.0010	**0.0002**
Facebook-2	0.8969	0.8985	1.5704	**1.5324**	0.0005	**0.0001**
Homes	0.9014	0.9015	0.5590	**0.5230**	0.0011	**0.0003**
Supercon.	0.9038	0.9040	0.8676	**0.8071**	0.0006	**0.0003**

D.8 Robustness Across Coverage Levels

The main experiments (Tables 1 and 2) target $\alpha = 0.1$. To assess sensitivity to the coverage level, we retrain and independently calibrate every method at $\alpha \in \{0.05, 0.2\}$ using the same default

hyperparameters, without any level-specific tuning. Tables 10 and 11 report within-dataset ranks on the ten original benchmarks. For real data, the conditional-reliability rank averages the MSCE, WSC, ℓ_1-ERT, and ℓ_2-ERT ranks (with WSC ranked in descending order). CoCP achieves the best overall mean rank in both length and conditional reliability at every evaluated level: $2.40 \,/\, 1.55$ at $\alpha = 0.05$ and $2.20 \,/\, 1.82$ at $\alpha = 0.2$. The advantage is strongest at the default $\alpha = 0.1$ ($1.40 \,/\, 1.00$) but remains favorable at both alternative levels despite no per-level tuning.

Table 10: Per-dataset ranks at $\alpha = 0.05$. Each entry is length rank / conditional-reliability rank; smaller is better for every reported rank. The conditional rank is based on ConMAE for synthetic data and is the mean of the MSCE, WSC, ℓ_1-ERT, and ℓ_2-ERT ranks for real data. WSC is ranked in descending order before averaging. Bold marks the best rank for each dataset and criterion.

	Synthetic			Real							Overall
Method	Normal	LogNormal	Exponential	*Bike*	*Bio*	*Blog*	*Facebook-1*	*Facebook-2*	*Homes*	*Supercon.*	Mean
Split	9.00 / 9.00	9.00 / 9.00	9.00 / 9.00	6.00 / 9.00	9.00 / 9.00	8.00 / 9.00	9.00 / 9.00	9.00 / 9.00	9.00 / 9.00	8.00 / 9.00	8.50 / 9.00
CPL	**1.00** / 8.00	3.00 / 8.00	4.00 / 8.00	5.00 / 5.00	6.00 / 5.25	9.00 / 7.75	6.00 / 5.00	6.00 / 5.00	8.00 / 7.25	5.00 / 6.75	5.30 / 6.60
CQR	5.00 / 3.00	5.00 / 6.00	5.00 / 7.00	8.00 / 2.50	3.00 / 4.00	2.00 / 4.75	3.00 / 7.75	4.00 / 8.00	4.00 / 5.25	3.00 / 2.50	4.20 / 5.08
A-CQR	3.00 / 4.00	2.00 / 3.00	**1.00** / 4.00	7.00 / 7.75	2.00 / 2.75	**1.00 / 2.25**	**1.00** / 4.75	**1.00** / 6.75	5.00 / 6.00	2.00 / 5.75	2.50 / 4.70
CPCP	4.00 / 7.00	8.00 / 7.00	6.00 / 6.00	3.00 / 3.75	8.00 / 7.00	6.00 / 3.50	8.00 / 5.75	8.00 / 4.75	6.00 / 5.50	4.00 / 4.75	6.10 / 5.50
RCP	2.00 / 6.00	6.00 / 4.00	7.00 / 3.00	4.00 / 5.50	7.00 / 6.00	7.00 / 6.50	7.00 / 3.00	7.00 / 2.50	7.00 / 6.00	6.00 / 6.00	6.00 / 4.85
C-HDR	6.00 / 2.00	7.00 / **1.00**	8.00 / **1.00**	2.00 / 6.00	4.00 / 7.00	5.00 / 3.75	5.00 / 3.50	5.00 / 3.25	2.00 / 2.25	7.00 / 2.75	5.10 / 3.25
CHR	8.00 / 5.00	4.00 / 5.00	2.00 / 5.00	9.00 / 4.00	5.00 / 2.75	3.00 / 4.50	4.00 / 5.25	2.00 / 4.75	3.00 / 2.50	9.00 / 6.00	4.90 / 4.47
CoCP	7.00 / **1.00**	**1.00** / 2.00	3.00 / 2.00	**1.00 / 1.50**	**1.00 / 1.25**	4.00 / 3.00	2.00 / **1.00**	3.00 / **1.00**	**1.00 / 1.25**	**1.00 / 1.50**	**2.40 / 1.55**

Table 11: Per-dataset ranks at $\alpha = 0.20$. Each entry is length rank / conditional-reliability rank; smaller is better for every reported rank. The conditional rank is based on ConMAE for synthetic data and is the mean of the MSCE, WSC, ℓ_1-ERT, and ℓ_2-ERT ranks for real data. WSC is ranked in descending order before averaging. Bold marks the best rank for each dataset and criterion.

	Synthetic			Real							Overall
Method	Normal	LogNormal	Exponential	*Bike*	*Bio*	*Blog*	*Facebook-1*	*Facebook-2*	*Homes*	*Supercon.*	Mean
Split	2.00 / 9.00	9.00 / 9.00	9.00 / 9.00	5.00 / 9.00	9.00 / 8.75	8.00 / 9.00	9.00 / 9.00	9.00 / 9.00	**1.00** / 9.00	4.00 / 9.00	6.50 / 8.97
CPL	**1.00** / 8.00	4.00 / 8.00	4.00 / 8.00	9.00 / 7.75	5.00 / 7.00	9.00 / 4.25	8.00 / 3.50	8.00 / 3.00	9.00 / 8.00	7.00 / 6.75	6.40 / 6.42
CQR	4.00 / 3.00	8.00 / 4.00	8.00 / 4.00	4.00 / 6.75	6.00 / 2.50	5.00 / 4.25	5.00 / 6.00	5.00 / 5.25	5.00 / 4.75	3.00 / 3.50	5.30 / 4.40
A-CQR	3.00 / 5.00	**1.00** / 2.00	3.00 / 2.00	3.00 / 6.50	3.00 / 2.50	3.00 / **2.50**	3.00 / 3.50	3.00 / 4.00	4.00 / 6.00	2.00 / 3.25	2.80 / 3.73
CPCP	8.00 / 7.00	6.00 / 6.00	6.00 / 6.00	7.00 / 3.50	7.00 / 4.25	6.00 / 4.25	6.00 / 3.50	7.00 / 2.50	7.00 / 6.00	5.00 / 6.50	6.50 / 4.95
RCP	7.00 / 6.00	5.00 / 5.00	7.00 / 5.00	6.00 / 2.00	8.00 / 5.50	7.00 / 4.75	7.00 / **1.00**	6.00 / **1.25**	8.00 / 4.75	8.00 / 5.00	6.90 / 4.03
C-HDR	5.00 / 2.00	7.00 / 7.00	5.00 / 7.00	8.00 / 3.50	4.00 / 8.25	4.00 / 8.00	4.00 / 7.50	4.00 / 8.00	6.00 / 2.50	9.00 / 3.25	5.60 / 5.70
CHR	9.00 / 4.00	3.00 / 3.00	2.00 / 3.00	2.00 / 5.00	**1.00** / 5.25	**1.00** / 5.25	**1.00** / 7.50	**1.00** / 7.00	2.00 / 3.00	6.00 / 6.75	2.80 / 4.97
CoCP	6.00 / **1.00**	2.00 / **1.00**	**1.00 / 1.00**	**1.00 / 1.00**	2.00 / **1.00**	2.00 / 2.75	2.00 / 3.50	2.00 / 5.00	3.00 / **1.00**	**1.00 / 1.00**	**2.20 / 1.82**

D.9 Sensitivity to Data Availability

We vary the amount of training and calibration data on the LogNormal and Exponential benchmarks at $\alpha = 0.1$. In the training-size experiment, every method is refitted for each $n \in \{2000, 5000, 10000, 20000\}$. In the calibration-size experiment, the fitted models are held fixed while calibration subsets of size $n_{\text{cal}} \in \{100, 250, 500, 1000, 2000, 4000\}$ are resampled 50 times per training run. Figure 9 compares Split, CQR, A-CQR, CHR, and CoCP via within-dataset metric ratios relative to CQR, averaged over the two benchmarks.

Across training sizes, CoCP reduces Length by approximately 13–18% and ConMAE by 25–46% relative to CQR. These gains are already present at $n = 2{,}000$, although the ConMAE ratio is not monotonic in n. As calibration size increases from 100 to 4,000, CoCP's mean Length ratio decreases from 0.885 to 0.830 and its mean ConMAE ratio from 0.909 to 0.545. At $n_{\text{cal}} = 100$, CHR and A-CQR can be slightly shorter on average, but CoCP maintains the lowest mean ConMAE ratio at every calibration size. The relative gains are thus not confined to the largest data regime.

Table 12 extends the comparison to three additional small real-world datasets: *Concrete* ($n = 1{,}030$) [48], *Community* ($n = 1{,}994$) [38], and *Star* ($n = 2{,}161$) [1]. CoCP produces the shortest intervals on *Concrete* and *Community* and the second-shortest on *Star*. Its MSCE is lowest on *Community* and tied for lowest on *Star*; on *Concrete*, A-CQR has the edge. These results confirm that CoCP's efficiency gains persist at small sample sizes, while conditional-reliability advantages become more dataset-dependent in this regime.

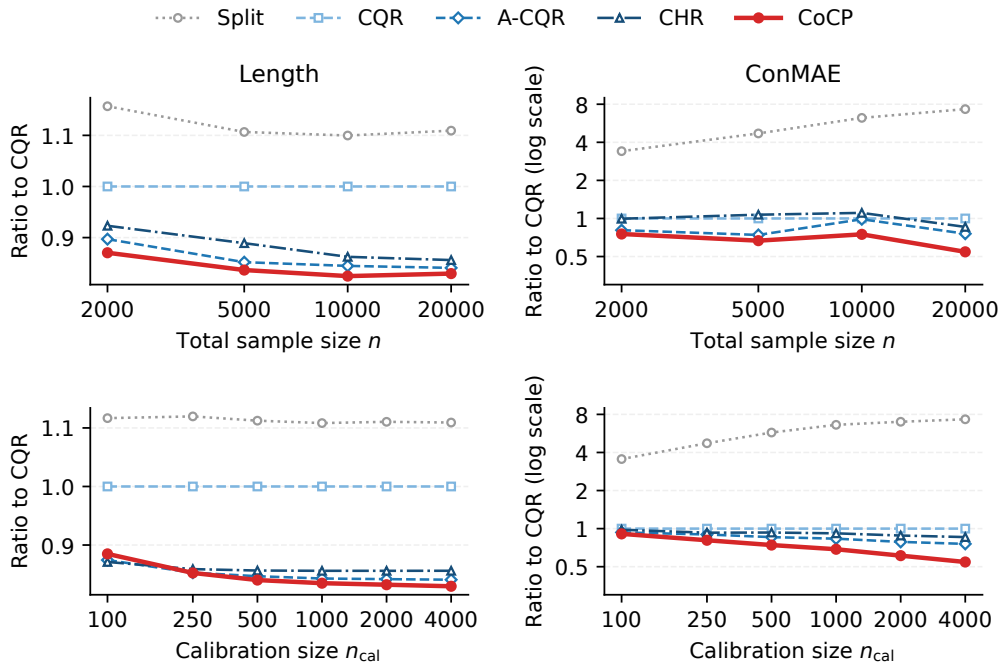

Figure 9: Sensitivity to training size (top row) and calibration size (bottom row) on the LogNormal and Exponential benchmarks at $\alpha = 0.1$. Results are ratios to CQR (lower is better; CQR $= 1$), averaged over the two benchmarks. In the top row, every method is refitted at each total sample size n; in the bottom row, the trained models are held fixed and only the calibration set is resampled. Horizontal axes use a logarithmic scale; the ConMAE axis (right column) is also logarithmic.

Table 12: Complete results on three additional small real datasets at $\alpha = 0.1$ (mean $\pm$ std over ten splits). Arrows indicate the preferred direction; bold marks the best value for each dataset and metric.

Dataset	Method	Coverage	Length (↓)	MSCE (↓)	WSC (↑)	ℓ_1-ERT (↓)	ℓ_2-ERT (↓)
Community	Split	0.9035 (0.0287)	2.0124 (0.1940)	0.0133 (0.0068)	0.5671 (0.0886)	0.0521 (0.0153)	-0.0096 (0.0092)
	CPL	0.9000 (0.0121)	1.7506 (0.0905)	0.0058 (0.0040)	0.6311 (0.0528)	0.0241 (0.0148)	-0.0185 (0.0077)
	CQR	0.9023 (0.0202)	1.7128 (0.0752)	0.0032 (0.0008)	0.6522 (0.0396)	0.0118 (0.0175)	**-0.0242** (0.0065)
	A-CQR	0.9018 (0.0213)	1.6729 (0.1043)	0.0035 (0.0010)	0.6609 (0.0493)	0.0091 (0.0198)	-0.0231 (0.0063)
	CPCP	0.9045 (0.0390)	1.8380 (0.3240)	0.0054 (0.0017)	0.6577 (0.0895)	0.0439 (0.0182)	-0.0123 (0.0096)
	RCP	0.9283 (0.0193)	1.9499 (0.1778)	0.0033 (0.0012)	**0.7125** (0.0460)	0.0376 (0.0185)	-0.0129 (0.0061)
	C-HDR	0.8935 (0.0242)	1.9120 (0.5636)	0.0042 (0.0018)	0.6364 (0.0491)	0.0291 (0.0168)	-0.0180 (0.0069)
	CHR	0.9170 (0.0235)	1.8082 (0.1942)	0.0025 (0.0011)	0.6950 (0.0473)	0.0139 (0.0188)	-0.0201 (0.0056)
	CoCP	0.9123 (0.0185)	**1.6202** (0.0913)	**0.0022** (0.0006)	0.6811 (0.0261)	**0.0077** (0.0231)	-0.0226 (0.0092)
Concrete	Split	0.8990 (0.0248)	0.6331 (0.0408)	0.0122 (0.0098)	0.5018 (0.1013)	0.0506 (0.0291)	0.0015 (0.0113)
	CPL	0.9252 (0.0320)	0.7168 (0.1066)	0.0077 (0.0035)	**0.5952** (0.0817)	0.0519 (0.0209)	0.0019 (0.0046)
	CQR	0.8976 (0.0205)	0.7031 (0.1083)	0.0063 (0.0040)	0.5312 (0.0602)	0.0283 (0.0192)	-0.0004 (0.0055)
	A-CQR	0.8898 (0.0285)	0.6843 (0.0363)	**0.0062** (0.0027)	0.5160 (0.0566)	**0.0157** (0.0275)	-0.0019 (0.0065)
	CPCP	0.9049 (0.0320)	0.6370 (0.0592)	0.0105 (0.0092)	0.5375 (0.1299)	0.0355 (0.0334)	0.0025 (0.0128)
	RCP	0.9136 (0.0354)	0.6937 (0.0929)	0.0102 (0.0057)	0.5552 (0.1082)	0.0625 (0.0173)	0.0046 (0.0039)
	C-HDR	0.8898 (0.0270)	0.5908 (0.0709)	0.0087 (0.0074)	0.5187 (0.0734)	0.0188 (0.0186)	**-0.0040** (0.0034)
	CHR	0.9019 (0.0174)	0.5957 (0.0661)	0.0068 (0.0028)	0.5460 (0.0544)	0.0294 (0.0278)	-0.0013 (0.0029)
	CoCP	0.8966 (0.0260)	**0.5177** (0.0512)	0.0079 (0.0046)	0.5304 (0.0728)	0.0231 (0.0434)	-0.0013 (0.0059)
Star	Split	0.9009 (0.0228)	0.2528 (0.0127)	0.0026 (0.0014)	0.6729 (0.0391)	0.0231 (0.0134)	-0.0070 (0.0036)
	CPL	0.8977 (0.0232)	0.2733 (0.0238)	0.0038 (0.0014)	0.6525 (0.0383)	0.0361 (0.0206)	-0.0046 (0.0041)
	CQR	0.9076 (0.0235)	0.3034 (0.0202)	0.0030 (0.0009)	0.6770 (0.0400)	**0.0117** (0.0129)	**-0.0098** (0.0037)
	A-CQR	0.8984 (0.0205)	0.2731 (0.0131)	0.0038 (0.0009)	0.6591 (0.0351)	0.0319 (0.0220)	-0.0073 (0.0046)
	CPCP	0.9012 (0.0491)	0.2884 (0.0322)	0.0058 (0.0049)	0.6720 (0.1068)	0.0469 (0.0230)	-0.0033 (0.0059)
	RCP	0.9049 (0.0256)	0.2999 (0.0258)	0.0046 (0.0016)	0.6340 (0.0647)	0.0522 (0.0177)	-0.0017 (0.0062)
	C-HDR	0.9116 (0.0217)	**0.2137** (0.0138)	0.0034 (0.0014)	0.6950 (0.0426)	0.0243 (0.0114)	-0.0075 (0.0043)
	CHR	0.9188 (0.0151)	0.2809 (0.0151)	**0.0023** (0.0008)	0.6987 (0.0368)	0.0171 (0.0191)	-0.0071 (0.0040)
	CoCP	0.9178 (0.0155)	0.2515 (0.0119)	**0.0023** (0.0007)	**0.7151** (0.0409)	0.0266 (0.0163)	-0.0058 (0.0026)

E Limitations and Future Work

Although CoCP generally produces shorter intervals and improves empirical conditional-reliability measures on the evaluated benchmarks, several theoretical and practical limitations remain.

Assumption of unimodality. CoCP is designed to output a single, contiguous prediction interval. As proven in our theoretical analysis, this naturally recovers the optimal Highest Density Interval (HDI) when the underlying conditional distribution is unimodal. However, for highly multimodal distributions, the true Highest Density Region (HDR) may consist of disjoint regions. While CoCP will still output a marginally valid interval that captures the required mass, a single contiguous interval is fundamentally suboptimal for multimodal data. Extending the push-pull geometry to multi-interval outputs (e.g., via mixture-density centers) is a promising direction.

Dependency on base estimator consistency. As discussed in Section 5, while finite-sample marginal coverage is always guaranteed regardless of model quality, CoCP's asymptotic optimality strongly depends on the statistical consistency of the base estimators (Assumption 2). In practice, due to the highly non-convex nature of neural network optimization, strict theoretical consistency cannot be guaranteed. Consequently, under severe data sparsity, suboptimal local minima, or model misspecification, the learned geometry may deviate from the true HDI.

Offline computational overhead. To reduce co-adaptation and stabilize gradient dynamics, CoCP relies on a K-fold alternating optimization scheme. While this does not affect real-time inference speed (which remains on par with standard CQR), it increases the offline training cost. Although our Fast variant (Appendix D.2) significantly mitigates this overhead via fold-parallelization and strict early stopping, CoCP remains more computationally demanding to train than single-pass methods. Exploring joint optimization strategies that bypass strict alternation without sacrificing stability could further streamline the training pipeline.

Reliance on smoothing approximations. The geometric boundary-balancing mechanism relies on a temperature parameter β to smooth the hard interval indicator. Although our sample-adaptive strategy (Appendix D.4) largely eliminates the need for manual tuning and provides scale invariance, the optimization landscape is still an approximation of the true non-differentiable coverage objective. In extreme cases of heavy-tailed distributions or severe heteroscedasticity, the gradient signal from the smoothed boundaries might still suffer from vanishing or exploding behaviors.

Extension to multivariate and structured outputs. The current formulation of CoCP is tailored to univariate regression, where the boundary consists of exactly two endpoints. Extending this geometric push-pull mechanism to multivariate or structured outputs (e.g., bounding boxes in object detection, semantic segmentation masks) remains nontrivial. A natural first step would be to generalize CoCP to fixed geometric templates that share a similar center-scale structure, such as hyperrectangles, ellipsoids (Mahalanobis balls), or general L_p-norm balls. However, achieving optimal arbitrary shapes in higher dimensions requires matching the density along a continuous high-dimensional contour rather than discrete endpoints. Furthermore, while generative or representation-based pipelines [13] map data to latent spaces, translating a latent center may not preserve volume-optimality in the original space due to geometric distortions. Formulating a tractable, differentiable surrogate for multi-dimensional boundary balancing presents an exciting frontier for conformal regression.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the proposed CoCP method, its theoretical guarantees (marginal and asymptotic conditional coverage), and empirical performance, which are fully supported by the methodology in Section 4, theoretical analysis in Section 5, and experiments in Section 6.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: A dedicated discussion of limitations and future work is provided in Appendix E, addressing the unimodality assumption, dependency on base estimator consistency, computational overhead, and challenges with multivariate extensions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All assumptions (e.g., density regularity, estimation consistency, calibration regularity) are explicitly stated in Section 5. Complete proofs for all theorems, propositions, and lemmas are provided in Appendices A and B.

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The paper provides detailed pseudocode (Algorithms 1 and 2) and comprehensive implementation details, including network architectures, hyperparameters, data preprocessing steps, and baseline configurations in Appendix C.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The supplementary material provides the experimental code and instructions for reproducing the main comparisons.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Experimental settings, including data generation, data splits, network architectures, hyperparameters, and baseline configurations, are detailed in Section 6 and Appendix C. Additional experiments on center refinement, asymmetric CQR, coverage levels, and data availability are documented in Appendix D.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The main experimental results (Tables 1 and 2) and additional small-dataset results (Table 12) report means and standard deviations over 10 random data splits. Runtime, implementation variants, and local modality diagnostics also report variability across splits (Tables 4, 5, 6, 7, and 8). Further robustness analyses, including asymmetric CQR and alternative coverage levels, are reported in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix C details the hardware used (2x NVIDIA A100 80GB PCIe GPUs, CUDA 12.6, peak memory usage), and Appendix D.1 provides comprehensive runtime comparisons (training and inference times) for all methods.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The paper presents foundational methodological research in machine learning and strictly adheres to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: This is foundational theoretical and methodological work on conformal prediction. It aims to improve uncertainty quantification reliability, which generally has positive impacts, but it does not have direct foreseeable negative societal impacts or specific applications requiring mitigation strategies.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not introduce models or datasets with a high risk for misuse (e.g., no LLMs, image generators, or scraped datasets are involved).

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All real-world datasets used in the experiments are standard, publicly available benchmarks, and their original sources are properly cited in Appendix C.1 (Table 3).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes]

Justification: The CoCP implementation is publicly released with documentation, usage instructions, and an open-source license.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The research does not involve crowdsourcing or human subjects.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The research does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs were not used as a component of the core methods in this research.

Guidelines:

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