Beyond Marginal Coverage: Efficient Localized Conformal Prediction via Residual Rank Calibration

Xiangshi Li **Wenqing He**
Department of Statistical and Actuarial Sciences
Western University, London, Ontario, Canada

Abstract

Conformal prediction provides finite-sample marginal coverage guarantees under exchangeability, but marginal validity alone does not ensure that prediction intervals adapt to local structure in the conditional distribution. Existing adaptive methods such as conformalized quantile regression (CQR) and conformal histogram regression (CHR) improve conditional coverage by incorporating estimated quantiles or conditional densities into the conformity score, but doing so requires direct estimation of the conditional distribution of response Y given covariates X, which can be statistically demanding. We propose a residual-based local rank conformal histogram regression (RLR-CHR) method that replaces direct estimation of the conditional distribution of Y with a more stable target by decoupling location and scale effects via a learned noise proxy and constructing histogram-based conformity scores on the standardized residuals. An unweighted conformal test of the local ranks of these scores provides an alternative approach to localized calibration. We establish finite-sample marginal coverage for RLR-CHR under exchangeability and asymptotic conditional coverage under local stability of the residual score distribution. We further prove that RLR-CHR is asymptotically equivalent to its weighted localization counterpart RBC-CHR under suitable regularity conditions. Experiments on synthetic and real datasets demonstrate that RLR-CHR can produce more compact prediction intervals than existing conformal approaches while maintaining worst-slab coverage near nominal.

1 Introduction

Conformal prediction [Vovk et al., 2005] has received increasing attention as a practical framework for uncertainty quantification in regression. Given exchangeable observations, split conformal prediction constructs prediction sets with finite-sample marginal coverage without parametric assumptions [Lei et al., 2018], making it compatible with modern regression algorithms whose sampling distributions are difficult to characterize.

Although split conformal prediction sets have marginal guarantees over the joint distribution of the covariates X and the response Y, this marginal guarantee does not ensure that the coverage holds at a particular covariate value $X = x$. Vovk [2012], Barber et al. [2021b] showed that exact conditional coverage is unattainable without distributional assumptions, which has motivated methods that preserve marginal validity while improving local adaptivity. Two common strategies have emerged in the literature: localized calibration [Lei and Wasserman, 2014, Guan, 2019] which assigns larger weights to calibration points that are more similar to the test point, and distribution-aware scoring.

Distribution-aware methods such as conformalized quantile regression (CQR) [Romano et al., 2019] and conformal histogram regression (CHR) [Sesia and Romano, 2021] have been used to improve interval adaptivity. CQR uses estimated conditional quantiles to construct asymmetric intervals, while CHR uses an estimated conditional histogram to build conformity scores that capture skewness and

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

multimodality of the conditional distribution. These methods, however, require direct modeling of the conditional distribution of Y given X. This requirement places a single estimator under pressure to simultaneously capture the conditional mean, spread, and shape, which becomes particularly challenging in nonparametric settings.

In this paper, we propose a residual-based local-rank conformal histogram regression (RLR-CHR) approach. Instead of modeling the full conditional distribution directly, we decompose the problem through a progressive simplification. We first construct standardized residuals by centering the response with a point predictor and rescaling by an estimated spread. The conformity scores are then computed from these standardized residuals. After removing the main location and scale effects, the standardized residual is expected to have a more stable conditional distribution across covariate values. A histogram-based conformal scoring procedure is then applied to these standardized residuals to produce discrete residual CHR scores, which are subsequently converted into local rank statistics via local conformal calibration. Finally, an unweighted conformal rank test is applied to the augmented calibration set to generate the prediction set.

This construction follows the localized conformal prediction perspective in the sense that calibration is performed using points close to the test point in the localization coordinate. Existing localized conformal methods commonly construct weighted conformal quantiles around the test point and then apply an additional correction step to retain finite-sample validity [Guan, 2019], whereas RLR-CHR uses local rank comparisons of the standardized residual CHR scores. To explore the relationship between these two approaches, we study a weighted localization counterpart, RBC-CHR, in the Appendix and show that it is asymptotically equivalent to RLR-CHR under the regularity conditions.

The main contributions of this paper are as follows:

- **Method.** We propose RLR-CHR, a novel residual-based conformal prediction method that decouples location and scale effects via a learned scale estimate and applies histogram-based conformity scoring to the standardized residuals. Unlike existing localized conformal methods that rely on weighted quantile estimation, RLR-CHR performs calibration through local rank comparisons.
- **Theory.** We establish finite-sample marginal coverage guarantees for RLR-CHR under exchangeability and develop asymptotic conditional coverage theory under local stability of the residual CHR score distribution. We further prove that RLR-CHR is asymptotically equivalent to its weighted localization counterpart RBC-CHR under suitable regularity conditions.
- **Experiments.** We evaluate RLR-CHR on both synthetic and real datasets. The results demonstrate improved worst-slab coverage on most datasets and gains in interval efficiency from residual CHR scoring with local calibration.

1.1 Related work

Existing adaptive conformal methods differ in what they estimate and how they calibrate. Besides CQR and CHR, distributional conformal prediction constructs conformity scores from an estimated conditional distribution function [Chernozhukov et al., 2021]. Normalized residual methods rescale absolute residual scores using an estimated local scale [Papadopoulos et al., 2011, Lei et al., 2018]. Split localized conformal prediction and rectified conformity scores use estimates of conditional score quantiles to transform the scores [Han et al., 2022, Plassier et al., 2025]. RLR-CHR combines scale normalization with a conditional histogram of residual shape and calibrates the resulting scores through local ranks.

Local calibration can use partitions [Lei and Wasserman, 2014, Boström and Johansson, 2020], learned features [Kiyani et al., 2024b], or weights based on similarity to the test point [Guan, 2019, Amoukou and Brunel, 2023]. Weighted conformal prediction also provides a framework for covariate shift [Tibshirani et al., 2019], while randomized localization gives model-free local guarantees [Hore and Barber, 2025]. Our analysis concerns an unweighted rank test in a single fitted coordinate. The explicit condition on score stability and the equivalence to weighted localization describe its connection to these methods; a comparison of guarantees appears in Appendix B.12.

The class of prediction sets also matters for efficiency. Density-based regions, thresholded intervals, and structured prediction sets can be disconnected [Izbicki et al., 2020, 2022, Luo and Zhou, 2025,

Gao et al., 2025]. Length optimization under conditional coverage requirements has also been studied directly [Kiyani et al., 2024a]. RLR-CHR returns a single interval, so its oracle benchmark is the shortest interval with the required conditional coverage. Appendix B.12 discusses further connections, including concurrent work on conditional coverage.

The remainder of the paper is organized as follows. Section 2 introduces the residual-based construction and the RLR-CHR calibration approach. Section 3 presents the finite-sample and asymptotic theoretical results. Simulation studies and real data analyses are reported in Section 4. We conclude the paper with a discussion in Section 5. The weighted localization construction, proofs, and regularity conditions are deferred to the Appendix.

2 Methodology

2.1 Background

Consider exchangeable data $\{(X_i, Y_i)\}_{i=1}^n$ drawn from an unknown distribution over $\mathcal{X} \times \mathbb{R}$. The split conformal prediction framework [Lei et al., 2018] partitions the observed data into a training set and a calibration set, with $\mathcal{I}_{\text{train}}$ and $\mathcal{I}_{\text{cal}}$ denoting the corresponding index sets. A regression estimator $\hat{f}$ is fitted on the training set $\mathcal{I}_{\text{train}}$ and evaluated on the held-out calibration set.

Given a measurable conformity score function $s(x, y)$, the calibration scores $E_i = s(X_i, Y_i)$ are computed for each $i \in \mathcal{I}_{\text{cal}}$. Let $m := |\mathcal{I}_{\text{cal}}|$ denote the calibration set size.

The adjusted empirical $(1-\alpha)$-quantile of the calibration scores is defined as

$$t_{1-\alpha}^{\text{Split}} = \text{the } \lceil (1-\alpha)(m+1) \rceil \text{-th smallest of } \{E_i : i \in \mathcal{I}_{\text{cal}}\} \cup \{\infty\}. \tag{1}$$

The split conformal prediction set for a test covariate X_{n+1} is then given by $\widehat{C}(X_{n+1}) = \{y \in \mathbb{R} : s(X_{n+1}, y) \leq t_{1-\alpha}^{\text{Split}}\}$.

By exchangeability of the data, this construction guarantees a finite-sample marginal coverage:

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

When the absolute residual score $E_i = |Y_i - \hat{f}(X_i)|$ is used, the resulting prediction interval is

$$\widehat{C}(X_{n+1}) = \left[\hat{f}(X_{n+1}) - t_{1-\alpha}^{\text{Split}},\ \hat{f}(X_{n+1}) + t_{1-\alpha}^{\text{Split}}\right].$$

Its width $2t_{1-\alpha}^{\text{Split}}$ is therefore the same for all test points. This uniformity fails to adapt to heteroscedasticity or varying tail behavior in the conditional distribution of Y given X, motivating adaptive methods that employ richer, covariate-dependent conformity scores.

CQR and CHR are two widely used distribution-aware conformal methods. CQR [Romano et al., 2019] uses the score $s(x, y) = \max\{\hat{q}_{\alpha/2}(x) - y,\ y - \hat{q}_{1-\alpha/2}(x)\}$ to calibrate a prediction interval, where $\hat{q}_\tau(x)$ estimates the conditional τ-quantile of Y given $X = x$.

In contrast, CHR [Sesia and Romano, 2021] models the full conditional distribution. In CHR, the conditional density $\widehat{f}_{Y|X}(\cdot \mid x)$ is estimated nonparametrically through a histogram. For each covariate value, the estimated histogram is used to construct a sequence of intervals ordered by their estimated probability mass. The intervals are constructed to be nested: each successive interval contains the previous one. A response is scored by the first interval in this family that contains it, and split calibration selects the final level. By using the estimated distributional shape rather than only two quantiles, CHR can produce shorter prediction intervals under complex conditional distributions. The cost of this flexibility is that CHR requires a full conditional density estimate, which makes the estimation task considerably more difficult. This motivates the residual-based construction developed in the next subsection.

Another way to adapt prediction intervals to local structure is localized conformal prediction, first introduced by Guan [2019]. It uses a weighted score distribution around each test covariate and adjusts the quantile level to preserve marginal validity. For comparison, we use its G2 construction as a weighted calibration baseline.

2.2 The proposed RLR-CHR approach

The proposed RLR-CHR approach first constructs residual CHR scores from standardized residuals and then calibrates these scores through local rank comparisons.

2.2.1 Residual histograms and the RB-CHR baseline

Instead of constructing conformity scores directly from the conditional distribution of Y given X, the residual-based framework starts from a point predictor and removes the location component before modeling the remaining residual variation. This produces the residual-based conformal histogram regression (RB-CHR).

As in the split conformal procedure, we partition the data into a training set $\mathcal{I}_{\mathrm{train}}$ and a calibration set $\mathcal{I}_{\mathrm{cal}}$. Let $\hat{f}$ denote any regression estimator trained on $\mathcal{I}_{\mathrm{train}}$. Since $\hat{f}$ may overfit the training data [Hastie et al., 2009], in-sample residuals can be systematically smaller than out-of-sample residuals. To mitigate this effect, we employ K-fold cross-fitting within $\mathcal{I}_{\mathrm{train}}$ [Barber et al., 2021a], so that each residual is computed from a model that was not trained on the corresponding observation. For each fold $\mathcal{I}_{\mathrm{train}}^{(q)}$, we train $\hat{f}_{-q}$ on the remaining folds and compute

$$R_i^{\mathrm{oof}} = Y_i - \hat{f}_{-q}(X_i), \qquad i \in \mathcal{I}_{\mathrm{train}}^{(q)}, \quad q = 1, \ldots, K.$$

These out-of-fold residuals are then used to construct the residual histogram.

RB-CHR models the full conditional distribution of the residuals via histograms [Sesia and Romano, 2021]. Specifically, we partition the residual domain into B equal-width bins $[b_0, b_1), [b_1, b_2), \ldots, [b_{B-1}, b_B)$ with bin width Δ. For the j-th bin, define the conditional bin probability $\pi_j(x) := \mathbb{P}(R \in [b_{j-1}, b_j) \mid X = x)$, $j \in \{1, \ldots, B\}$. These probabilities are estimated from the training pairs $(X_i, R_i^{\mathrm{oof}})$, yielding $\hat{\pi}(x) = (\hat{\pi}_1(x), \ldots, \hat{\pi}_B(x))$.

Given histogram probabilities π at a covariate value x, we seek consecutive bins with total probability at least τ. The index sets S^- and S^+ specify the bins that must be included and those that may be selected, respectively. The function $\mathcal{S}$ returns the endpoint indices of the shortest interval satisfying these constraints:

$$\mathcal{S}(x, \pi, S^-, S^+, \tau) := \operatorname*{arg\,min}_{1 \le \ell \le u \le B} (u - \ell) \quad \text{s.t.} \sum_{j=\ell}^{u} \pi_j(x) \ge \tau, \quad S^- \subseteq [\ell, u] \subseteq S^+,$$

Here $[\ell, u]$ denotes the bin index set $\{\ell, \ldots, u\}$. Since bin j is $[b_{j-1}, b_j)$, selecting bins $\ell, \ldots, u$ gives the residual interval $[b_{\ell-1}, b_u]$. Let $p_t = t/T$, $t = 1, \ldots, T$, be a grid of target probability levels. Starting from the level closest to $1 - \alpha$, we apply $\mathcal{S}$ to obtain nested residual intervals $S_1(x) \subseteq \cdots \subseteq S_T(x) = \mathbb{R}$. For each $t < T$, $S_t(x)$ is the shortest residual bin interval whose estimated histogram mass is at least p_t. Thus larger t corresponds to a larger target mass p_t and a weakly larger residual set.

The CHR-type conformity score for a calibration point in $\mathcal{I}_{\mathrm{cal}}$ is defined as the smallest index t such that its residual falls in $S_t(X_i)$:

$$E_i^{\mathrm{RB}} := \min\{t \in \{1, \ldots, T\} : R_i \in S_t(X_i)\}, \quad i \in \mathcal{I}_{\mathrm{cal}},$$

where $R_i := Y_i - \hat{f}(X_i)$. Let t_{split} be the split conformal quantile of the calibration scores $\{E_i^{\mathrm{RB}} : i \in \mathcal{I}_{\mathrm{cal}}\}$, as in (1), with T replacing ∞. The RB-CHR prediction interval is

$$\widehat{C}^{\text{RB-CHR}}(x_{n+1}) = \hat{f}(x_{n+1}) + S_{t_{\mathrm{split}}}(x_{n+1}).$$

RB-CHR shifts the histogram modeling task from the conditional distribution of $Y \mid X$ to that of the residual $Y - \hat{f}(X) \mid X$, making the CHR construction compatible with an existing point predictor.

2.2.2 Standardization and residual CHR scores

RB-CHR removes the location component through a point predictor, but calibrates the residual CHR scores globally. When the conditional spread of Y varies with X, the residual distribution may still differ across covariate values. We therefore further standardize the residuals to remove this effect.

When the response has the approximate form $Y = \mu(X) + \sigma(X)\varepsilon$, and we have estimates $\hat{\mu}(X) \approx \mu(X)$ and $\hat{\sigma}(X) \approx \sigma(X)$, the standardized residual $\widetilde{R} = (Y - \hat{\mu}(X))/\hat{\sigma}(X)$ is expected to have a more stable conditional distribution across covariate values.

To account for this scale variation, we estimate a scale function from the out-of-fold residuals. We use cross-fitting within $\mathcal{I}_{\text{train}}$ to obtain predictions $\hat{\mu}_{\text{oof}}(X_i)$, with each training observation predicted by a model fitted without its fold. We denote by $\hat{\mu}$ the predictor constructed from the training data for calibration and test points. Using the out-of-fold residuals $R_i^{\text{oof}} = Y_i - \hat{\mu}_{\text{oof}}(X_i)$, we estimate the local scale by fitting a regression model $\hat{\sigma}_{\text{raw}}(x)$ to predict the conditional median of the absolute residuals, $\hat{\sigma}_{\text{raw}}(x) \approx \text{median}\big(|R^{\text{oof}}| \mid X = x\big)$. We apply a robust median-based rescaling to obtain the final scale estimate:

$$c^* := \text{median}\Big\{\frac{|Y_i - \hat{\mu}_{\text{oof}}(X_i)|}{\hat{\sigma}_{\text{raw}}(X_i)} : i \in \mathcal{I}_{\text{train}}\Big\}, \qquad \hat{\sigma}(x) := c^* \cdot \hat{\sigma}_{\text{raw}}(x).$$

The standardized out-of-fold residuals used to train the residual histogram are

$$\widetilde{R}_i^{\text{oof}} = \frac{Y_i - \hat{\mu}_{\text{oof}}(X_i)}{\hat{\sigma}(X_i)}, \qquad i \in \mathcal{I}_{\text{train}},$$

and the standardized residuals for the calibration points are $\widetilde{R}_i = (Y_i - \hat{\mu}(X_i))/\hat{\sigma}(X_i)$. The conditional histogram $\hat{\pi}(x)$ is estimated from the pairs $(X_i, \widetilde{R}_i^{\text{oof}})$, $i \in \mathcal{I}_{\text{train}}$. Applying the CHR construction to the standardized residuals for each covariate x, we obtain the nested residual intervals $\tilde{S}_1(x) \subseteq \cdots \subseteq \tilde{S}_T(x) = \mathbb{R}$. The corresponding conformity score for each calibration point is

$$E_i := \min\{t \in \{1, \ldots, T\} : \widetilde{R}_i \in \tilde{S}_t(X_i)\}, \qquad i \in \mathcal{I}_{\text{cal}}. \tag{2}$$

This bounded discrete score is the object calibrated by RLR-CHR.

2.2.3 Local-rank calibration

We now describe the proposed residual-based local-rank conformal histogram regression (RLR-CHR). After the residual CHR scores (2) are obtained, RLR-CHR calibrates them locally by comparing their ranks among nearby calibration points. The localization coordinate is defined as

$$Z_i = \log \hat{\sigma}(X_i),$$

where using the logarithm makes neighborhoods depend on relative scale differences, since $|Z_i - Z_j| = |\log \hat{\sigma}(X_i) - \log \hat{\sigma}(X_j)|$ is the absolute log-ratio of the estimated scales. To avoid ties in the localization coordinates, we henceforth construct neighborhoods using $\widetilde{Z}_i = Z_i + h_n \xi_i$, where $h_n > 0$ is deterministic and the ξ_i are i.i.d. random variables with distribution $\text{Unif}(0, 1)$, independent of all data and fitted objects. We use $h_n = 10^{-9}$ in our experiments.

Let $J = \mathcal{I}_{\text{cal}} \cup \{n + 1\}$ be the augmented index set containing the calibration points and the test point. For each $j \in J$, define $\mathcal{N}_j(J)$ as the neighborhood containing j and the k indices in $J \setminus \{j\}$ that are closest to j according to the absolute differences $|\widetilde{Z}_\ell - \widetilde{Z}_j|$ in their localization coordinates. For a candidate threshold e, set $E_\ell(e) = E_\ell$ for $\ell \in \mathcal{I}_{\text{cal}}$ and $E_{n+1}(e) = e$. The local rank score for the candidate threshold e at point j is defined as

$$A_j(e) := \sum_{\ell \in \mathcal{N}_j(J)} \mathbf{1}\{E_\ell(e) \le E_j(e)\},$$

which counts the number of indices in $\mathcal{N}_j(J)$ whose conformity scores are no larger than that of j; a larger $A_j(e)$ therefore indicates poorer local conformity. The conformal p-value is then

$$p(e) := \frac{1 + \sum_{i \in \mathcal{I}_{\text{cal}}} \mathbf{1}\{A_i(e) \ge A_{n+1}(e)\}}{m + 1}.$$

Since the neighborhood rule is permutation-equivariant, it implies $\mathbb{P}\{p(E_{n+1}) \le \alpha\} \le \alpha$.

Note that $p(e)$ is nonincreasing in e. When e increases, the test point's local rank $A_{n+1}(e)$ increases while each calibration point's rank $A_i(e)$ can only decrease. The acceptance set $\{e : p(e) > \alpha\}$ is therefore either empty or an initial segment $\{1, \ldots, e^\star\}$. We define $e^\star := \max\{e : p(e) > \alpha\}$ with the convention $\max \varnothing := 1$. Under this convention, replacing the empty prediction set with the first interval does not reduce coverage. The prediction set

$$\widehat{C}^{\text{RLR-CHR}}(X_{n+1}) = \hat{\mu}(X_{n+1}) + \hat{\sigma}(X_{n+1}) \cdot \tilde{S}_{e^\star}(X_{n+1})$$

is a single interval since the nested residual sets $\tilde{S}_t(x)$ are intervals by construction. Algorithm 1 summarizes the proposed RLR-CHR procedure.

Algorithm 1 RLR-CHR: Residual Local-Rank Conformal Histogram Regression

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$ split into $\mathcal{I}_{\text{train}}$ and $\mathcal{I}_{\text{cal}}$; test point X_{n+1}; $\alpha \in (0,1)$; number of nearest neighbors k excluding the center point; number of CHR levels T.

Output: Prediction interval $\widehat{C}^{\text{RLR-CHR}}(X_{n+1})$.

Construct $\hat{\mu}, \hat{\sigma}$ from $\mathcal{I}_{\text{train}}$ and use cross-fitting to obtain standardized out-of-fold residuals $\widetilde{R}_i^{\text{oof}}$.

Fit the residual CHR model to $(X_i, \widetilde{R}_i^{\text{oof}})$, $i \in \mathcal{I}_{\text{train}}$, and compute E_i for $i \in \mathcal{I}_{\text{cal}}$.

Set $J := \mathcal{I}_{\text{cal}} \cup \{n+1\}$ and compute $\widetilde{Z}_j = \log \hat{\sigma}(X_j) + h_n \xi_j$ for each $j \in J$.

Construct permutation-equivariant k-NN neighborhoods: for each $j \in J$, let $\mathcal{N}_j(J)$ be the k nearest neighbors of j in $\widetilde{Z}$-space plus j itself, so $|\mathcal{N}_j(J)| = k+1$.

Let $m = |\mathcal{I}_{\text{cal}}|$ and set $E_\ell(e) = E_\ell$ for $\ell \in \mathcal{I}_{\text{cal}}$ and $E_{n+1}(e) = e$.

Local rank calibration: set $e^\star = 1$.

For $e = T, T-1, \ldots, 1$:

Compute $A_{n+1}(e) = 1 + \big|\{\ell \in \mathcal{N}_{n+1}(J) \setminus \{n+1\} : E_\ell \le e\}\big|$.

For each $i \in \mathcal{I}_{\text{cal}}$, compute $A_i(e) = \sum_{\ell \in \mathcal{N}_i(J)} \mathbf{1}\{E_\ell(e) \le E_i\}$, updating only indices with $n+1 \in \mathcal{N}_i(J)$.

Compute $p(e) = \big(1 + \sum_{i \in \mathcal{I}_{\text{cal}}} \mathbf{1}\{A_i(e) \ge A_{n+1}(e)\}\big)/(m+1)$.

If $p(e) > \alpha$, set $e^\star = e$ and stop.

Return $\widehat{C}^{\text{RLR-CHR}}(X_{n+1}) = \hat{\mu}(X_{n+1}) + \hat{\sigma}(X_{n+1}) \cdot \tilde{S}_{e^\star}(X_{n+1})$.

The proposed RLR-CHR differs from the G2 localization procedure of Guan [2019]. G2 constructs localized weighted conformal quantiles and then applies an additional correction step, whereas RLR-CHR applies an unweighted conformal rank test to local ranks of the residual CHR scores. The same residual CHR scores can also be combined with Guan's G2 approach to obtain the counterpart RBC-CHR, which is described in Appendix A.2. Appendix B.10 discusses Guan's later construction and algorithm. Appendix B.11 illustrates the rank calculation.

3 Theoretical results

In this section, we establish finite-sample marginal coverage for RLR-CHR under exchangeability. We then show that conditional coverage approaches the nominal level under local stability of the residual CHR score distribution and that, under additional regularity conditions in Appendix A.4, the interval length converges to the oracle shortest length.

Let $\mathcal{I}_{\text{train}} \subset [n]$ denote the training index set used to fit all model components required for constructing the conformity score, and let $\mathcal{I}_{\text{cal}} \subset [n]$ denote a disjoint calibration index set for computing the conformal cutoff. Let $\mathcal{F}_n := \sigma(\mathcal{I}_{\text{train}}, \{(X_i, Y_i) : i \in \mathcal{I}_{\text{train}}\})$ denote the σ-algebra generated by the random split and the training sample used to construct the fitted score components.

The following theorem gives a marginal coverage guarantee by the standard exchangeability argument.

Theorem 1. *Assume that $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable and that all fitted objects used to construct $\widehat{C}^{\text{RLR-CHR}}$ are $\mathcal{F}_n$-measurable. Then, for any $\alpha \in (0,1)$,*

$$\mathbb{P}\Big(Y_{n+1} \in \widehat{C}^{\text{RLR-CHR}}(X_{n+1})\Big) \ge 1 - \alpha.$$

We next study RLR-CHR under an asymptotic regime. The key requirement is that the localization coordinate groups together points with similar residual CHR score distributions, so that the local calibration scores provide a good approximation to the score distribution at the test point.

For a test covariate $x \in \mathcal{X}$, the RLR-CHR prediction set takes the form

$$\widehat{C}^{\text{RLR-CHR}}(x) = \hat{\mu}(x) + \hat{\sigma}(x) \cdot \tilde{S}_{\hat{e}(x)}(x),$$

where $\hat{e}(x) \in \{1, \ldots, T\}$ is the score index selected by local-rank calibration. $\hat{e}(x)$ corresponds to the value $e^\star$ in Algorithm 1 when the test covariate is x. Writing $\tilde{S}_t(x) = [\tilde{\ell}_t(x), \tilde{u}_t(x)]$, the corresponding interval length is $\widehat{L}^{\text{RLR-CHR}}(x) = \hat{\sigma}(x)\big(\tilde{u}_{\hat{e}(x)}(x) - \tilde{\ell}_{\hat{e}(x)}(x)\big)$. Let $T = T_n$ and

$k = k_n$ denote the number of CHR levels and the neighborhood size. Let $\mathcal{D}_n := \sigma\big(\mathcal{F}_n, \{(X_j, Y_j) : j \in \mathcal{I}_{\mathrm{cal}}\}, \{\xi_j : j \in J\}\big)$. For $x \in \mathcal{X}$ and $t \in \{0, 1, \ldots, T\}$, let $F_x(t) := \mathbb{P}(E \le t \mid X = x, \mathcal{F}_n)$, where E is the standardized residual CHR score computed from the fitted objects trained on $\mathcal{I}_{\mathrm{train}}$.

Assumption 1. Conditional on $\mathcal{F}_n$, the augmented covariates, and $\{\xi_j : j \in J\}$, define the neighborhood discrepancy

$$\eta_{Z,n} := \sup_{r \in J} \sup_{j \in \mathcal{N}_r(J)} \max_{0 \le t \le T} \big|F_{X_j}(t) - F_{X_r}(t)\big|$$

and the largest score atom

$$\gamma_n := \sup_{x \in \mathcal{X}} \max_{1 \le t \le T} \mathbb{P}(E = t \mid X = x, \mathcal{F}_n).$$

The score distributions are locally stable in the chosen coordinate and their atoms vanish: $\eta_{Z,n} = o_{\mathbb{P}}(1)$ and $\gamma_n = o_{\mathbb{P}}(1)$.

Appendices A.4 and A.5 discuss the scope of Assumption 1 and give sufficient conditions for the assumption.

Theorem 2. *Assume the data are i.i.d., the fitted objects are $\mathcal{F}_n$-measurable, Assumption 1 holds, and the localization parameters satisfy $m \to \infty$, $k \to \infty$, $k/m \to 0$, $\log(T \vee 2)/k \to 0$, and $k \log(m \vee 2)/m \to 0$. Then*

$$\begin{aligned}&\Big|\mathbb{P}\big(Y_{n+1} \in \widehat{C}^{\text{RLR-CHR}}(X_{n+1}) \mid X_{n+1}, \mathcal{D}_n\big) - (1-\alpha)\Big| \\ &\qquad = O_{\mathbb{P}}\left(\eta_{Z,n} + \gamma_n + \sqrt{\frac{\log(T \vee 2)}{k}} + \sqrt{\frac{k \log(m \vee 2)}{m}} + \frac{k}{m}\right).\end{aligned}$$

Theorem 2 can be interpreted as follows. If the localization coordinate forms neighborhoods in which the residual CHR score distribution varies little, and if the neighborhood size grows but remains small relative to the calibration sample, then the conditional coverage of RLR-CHR at the test covariate approaches $1 - \alpha$. The rate separates the sources of error: $\eta_{Z,n}$ captures the locality error of the score distribution, γ_n is the discreteness and tie cost from the residual CHR score atoms, $\sqrt{\log(T \vee 2)/k}$ reflects the local empirical fluctuation over the T score levels, and $\sqrt{k \log(m \vee 2)/m} + k/m$ is the cost of using overlapping k-NN neighborhoods in the calibration sample.

Remark 1. *Theorem 2 evaluates conditional coverage at the realized test covariate, with convergence in probability under the random design. The locality term $\eta_{Z,n}$ uses the maximum discrepancy over the constructed neighborhoods, so the bound does not quantify how the rate varies with local covariate density. Appendix B.5 reports empirical coverage across strata defined by covariate density.*

We next consider interval length. Under the conditions below, the CHR sets approximate the shortest conditional residual intervals, so $F_x(t)$ is close to $p_t = t/T$. Local rank calibration then selects an interval with conditional probability close to $1 - \alpha$.

Write G_x for the conditional distribution of the standardized residual $\widetilde{R} = (Y - \hat{\mu}(X))/\hat{\sigma}(X)$ given $X = x$ and $\mathcal{F}_n$, let $S_x^*(p) = [\ell_x^*(p), u_x^*(p)]$ be a shortest interval with probability p under G_x, and let $w_x^*(p) = u_x^*(p) - \ell_x^*(p)$. Conditions (E1)–(E4) of Appendix A.4 require that $\hat{\sigma}(x)$ be uniformly bounded above and away from zero, that G_x have a density bounded uniformly in x, that the shortest interval be unique with a width that is Lipschitz in p uniformly in x, and that the CHR sets approximate $S_x^*(p_t)$ uniformly in x and t.

Theorem 3. *Assume the conditions of Theorem 2 other than Assumption 1. Assume in addition that the covariate space is compact, $T = T_n \to \infty$, $\log(m \vee 2)/k \to 0$, and conditions (E1)–(E4) hold. Let $\rho_{\mathrm{chr},n}$ be the approximation rate in (E4) and set $\delta_n' := \rho_{\mathrm{chr},n} + 1/T$. Then Assumption 1 holds with $\eta_{Z,n} = O_{\mathbb{P}}(\delta_n')$ and $\gamma_n = O_{\mathbb{P}}(\delta_n')$, and*

$$\sup_{x \in \mathcal{X}} \big|\widehat{L}^{\text{RLR-CHR}}(x) - L^*(x)\big| = O_{\mathbb{P}}\left(\delta_n' + \sqrt{\frac{\log((T \vee 2)(m \vee 2))}{k}} + \sqrt{\frac{k \log(m \vee 2)}{m}} + \frac{k}{m}\right),$$

where $L^*(x)$ is the length of the shortest interval having conditional probability $1 - \alpha$ for Y given $X = x$.

Remark 2. *The location-scale model, together with the regularity and estimation conditions in Appendix A.4, satisfies (E1)–(E4) (Proposition 2). An example is given in Appendix A.4, and a setting that violates the location-scale assumption is shown in Appendix B.8.*

RBC-CHR, which uses weighted localization, has analogous guarantees for conditional coverage and convergence to the oracle length in Appendix A.6. Under the hypotheses of Theorem 3 and the local endpoint condition (C1) in Appendix A.4, its rate of convergence to the oracle length has the same order as that of RLR-CHR, and Theorem 7 in Appendix A.6 establishes asymptotic equivalence between the two methods. Proofs of the theoretical results are given in Appendices A.3–A.8.

4 Experiments

We conducted both synthetic and real data experiments to evaluate the proposed methods and compare them with the main competitors in terms of conditional coverage and interval length under heteroscedasticity. Following Cauchois et al. [2021], we employed worst-slab coverage (WSC) to approximate conditional coverage. Formally, WSC is defined as

$$\mathrm{WSC}_\delta \;=\; \inf_{\{S:\, P(X\in S)\geq\delta\}} P\big(Y \in \widehat{C}(X) \mid X \in S\big), \qquad S = \{x : a \leq v^\top x \leq b\},$$

which searches over all covariate slabs with probability mass at least δ and reports the smallest conditional coverage. A method with WSC_δ close to $1-\alpha$ achieves approximate coverage at the nominal level uniformly over these slabs. We use $\alpha = \delta = 0.1$ unless stated otherwise, and estimate WSC by searching and evaluating the slab on separate test subsamples; Appendix B.13 provides details of the WSC estimator.

We compare the proposed methods with conformal competitors that target distributional adaptivity or local coverage, including Split-CP, CQR [Romano et al., 2019], CHR [Sesia and Romano, 2021], and the following methods: Norm-SC, which applies split conformal prediction to normalized absolute residual scores [Papadopoulos et al., 2011, Lei et al., 2018]; MCR, a Mondrian conformal regressor that partitions calibration points into scale groups before calibration [Boström and Johansson, 2020]; G2-Residual, which applies Guan's G2 localized calibration [Guan, 2019] to normalized absolute residual scores; and randomized localized conformal prediction (RLCP) [Hore and Barber, 2025] for the real datasets. These comparisons help assess the contributions of different components of RLR-CHR (Table 2).

We used the same deep neural network architecture for all conditional quantile and point estimation within each study. For CHR, RBC-CHR, and RLR-CHR, we used the same CHR setup. The neighborhood size k for experiments on real data is chosen based on the theoretical guidance in Remark 5, which shows that the balancing of the localization terms in the conditional coverage rate yields $k \asymp \sqrt{m\log(T\vee 2)/\log(m\vee 2)}$. We report means and standard errors (SEs) over 100 repetitions. Additional implementation details are given in Appendix B.1.[1]

4.1 Synthetic data

We begin with a one-dimensional setup under non-Gaussian noise. Let $X \sim \mathrm{Uniform}(0,1)$ and generate $Y = f(X) + \sigma(X)\varepsilon$, where $f(x) = 2\sin(2\pi x) + 0.5(x-0.5)$, $\sigma(x) = 0.15$ for $x < 0.70$, and $\sigma(x) = 1.50$ for $x \geq 0.70$.

We generate the error as $\varepsilon = (V-1) + 0.1G$, where $V \sim \mathrm{Exp}(1)$ and $G \sim \mathcal{N}(0,1)$ are independent of each other and of X. It has mean zero and a right-skewed distribution, with most of its probability mass in the interval $[-1, 6]$. The high-noise region has ten times the noise scale of the low-noise region and accounts for 30% of the covariate distribution. This proportion exceeds the nominal miscoverage level $\alpha = 0.10$, creating a setting in which methods need to truly adapt to local heteroscedasticity.

Figure 1 reports the worst-slab coverage and average interval length for all methods as the sample size increases. RLR-CHR and RBC-CHR have similar WSC and interval length and generally produce more compact prediction intervals than the other methods except CHR. G2-Residual achieves comparable WSC but produces longer intervals, whereas Norm-SC remains below the target coverage. These comparisons indicate that standardization and localization alone do not account for RLR-CHR's

[1] Code and data documentation are available at `https://github.com/Xiangshi-L/rlr-chr`.

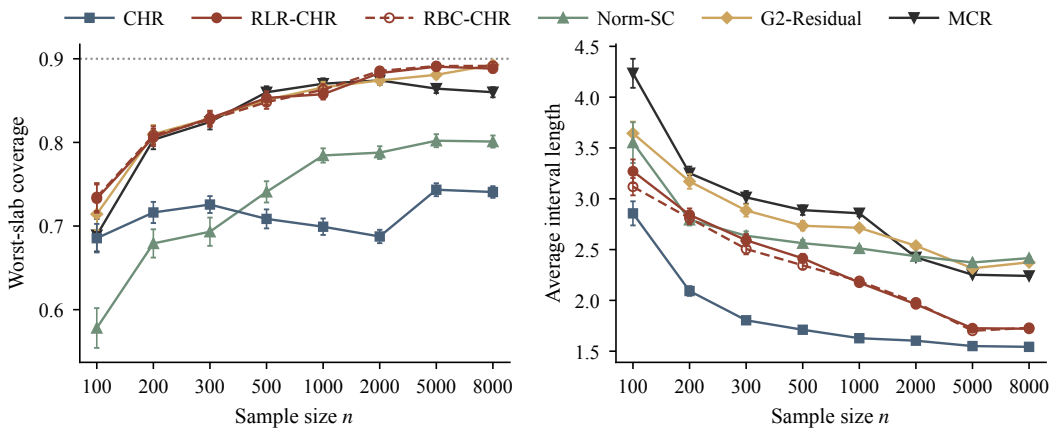


Figure 1: WSC and average interval length on synthetic data. Points are means over 100 repetitions; error bars show ± 1 SE. The sample size n is divided equally between training and calibration data.

performance. CHR yields shorter intervals at the cost of persistently low WSC, indicating that its direct modeling of the conditional response distribution with global calibration does not adequately adapt to local heteroscedasticity in this design.

4.2 Real data

We evaluate the proposed methods on eight public regression datasets spanning robotics, computer systems, energy demand, parking occupancy, and video transcoding: SARCOS Robot Arm Inverse Dynamics [Vijayakumar, 2006], SGEMM GPU kernel performance [Paredes and Ballester-Ripoll, 2017], Tetouan City Power Consumption Zone 1 [Salam and El Hibaoui, 2018], Parking Birmingham Occupancy [Stolfi, 2017], Naval Propulsion Compressor Decay [Coraddu et al., 2014], MV [Torgo, 2014], Online Video Characteristics and Transcoding Time [Deneke, 2014], and French Load [Keisler, 2024]. Dataset sources, preprocessing details, and complete result tables are reported in Appendix B.

The main real data results are summarized in Table 1. We report (i) the empirical WSC and (ii) the relative average interval length compared with that produced by CHR. Consistent with the asymptotic equivalence result established in Theorem 7, RBC-CHR and RLR-CHR have very close WSC values and interval lengths, so we focus on RLR-CHR in the discussion below.

In terms of interval efficiency, RLR-CHR maintains WSC close to the nominal level on most datasets while producing shorter intervals than the competing methods. Relative to CHR, shorter intervals are generally accompanied by WSC closer to nominal. The competing adaptive methods further illuminate this pattern. Norm-SC retains an absolute residual score with global calibration; it produces longer prediction intervals and lower WSC than RLR-CHR. G2-Residual applies Guan's localized calibration to the normalized residual scores, but produces longer prediction intervals with broadly comparable WSC. RLCP's median interval widths are also larger than RLR-CHR's mean lengths, with WSC generally farther from the nominal level. These comparisons highlight the importance of the progressive simplification strategy: scale normalization alone is insufficient, and localized calibration is more effective after the conformity score has been simplified into a residual CHR score.

Additional experiments clarify the scope of RLR-CHR. In the experiment reported in Appendix B.8, RLR-CHR retains WSC near the nominal level with shorter intervals even though the location-scale assumption (A2) is violated. We also conducted experiments on datasets from the CQR and CHR studies [Romano et al., 2019, Sesia and Romano, 2021]. RLR-CHR achieves WSC near the nominal level but produces more conservative intervals than CHR, especially on datasets with substantial zero inflation, which can be identified before fitting. This can happen because standardization maps the atom at zero to a location that depends on X, making the residual distribution harder to estimate (Appendix B.4). On Video Time, RLR-CHR produces extremely wide intervals in two of 100 repetitions. This behavior arises from large score atoms induced by discretization, which violate the small-atom condition in Assumption 1 (Appendix B.3). Further diagnostics and sensitivity analyses are given in Appendices B.6 and B.7.

Table 1: WSC and relative interval length on real data. Entries are means with standard errors in parentheses. Relative length is the mean ratio to CHR (CHR = 1). WSC values closest to the nominal level and the smallest relative lengths are bolded.

	SARCOS		SGEMM GPU		Tetouan		Parking	
Method	WSC	Rel. length	WSC	Rel. length	WSC	Rel. length	WSC	Rel. length
Split-CP	0.861 (.004)	1.061 (.004)	0.796 (.006)	1.334 (.015)	0.872 (.004)	1.070 (.005)	0.766 (.007)	1.200 (.010)
Norm-SC	0.884 (.003)	0.962 (.004)	0.893 (.003)	1.054 (.012)	0.887 (.003)	1.004 (.005)	0.834 (.006)	1.125 (.009)
CQR	0.887 (.003)	0.988 (.005)	0.881 (.003)	1.251 (.010)	0.888 (.003)	1.014 (.005)	0.876 (.003)	1.075 (.007)
CHR	0.884 (.003)	1.000 (.000)	0.892 (.003)	1.000 (.000)	0.889 (.003)	1.000 (.000)	**0.890** (.003)	1.000 (.000)
MCR	**0.902** (.003)	0.910 (.003)	0.896 (.002)	1.270 (.014)	0.892 (.003)	0.993 (.005)	0.885 (.003)	1.152 (.008)
G2-Residual	0.895 (.003)	0.926 (.004)	0.903 (.003)	1.053 (.009)	0.893 (.003)	1.000 (.004)	0.868 (.004)	1.122 (.007)
RBC-CHR	0.897 (.003)	0.791 (.004)	**0.898** (.002)	0.733 (.008)	0.894 (.003)	**0.864** (.004)	0.886 (.003)	0.889 (.006)
RLR-CHR	0.896 (.003)	**0.788** (.004)	0.897 (.003)	**0.732** (.008)	**0.899** (.003)	0.869 (.004)	0.881 (.003)	**0.887** (.006)
RLCP	0.893 (.003)	0.968 (.004)†	0.867 (.004)	1.209 (.016)†	0.898 (.003)	1.046 (.005)†	0.877 (.003)	1.126 (.008)†

	Naval		MV		Video Time		French Load	
Method	WSC	Rel. length	WSC	Rel. length	WSC	Rel. length	WSC	Rel. length
Split-CP	0.726 (.007)	1.087 (.019)	0.726 (.010)	1.269 (.029)	0.611 (.006)	1.039 (.012)	0.842 (.004)	1.100 (.011)
Norm-SC	0.734 (.008)	1.014 (.018)	0.790 (.006)	0.968 (.021)	0.825 (.004)	0.900 (.010)	0.890 (.003)	1.002 (.010)
CQR	0.758 (.007)	1.066 (.019)	0.700 (.009)	1.136 (.020)	0.730 (.009)	1.045 (.011)	0.855 (.004)	1.045 (.008)
CHR	0.734 (.007)	1.000 (.000)	0.707 (.008)	1.000 (.000)	0.768 (.010)	1.000 (.000)	0.880 (.003)	1.000 (.000)
MCR	0.799 (.005)	0.779 (.013)	**0.882** (.004)	0.996 (.021)	0.873 (.003)	1.037 (.011)	0.896 (.002)	0.985 (.010)
G2-Residual	0.811 (.005)	1.070 (.017)	0.878 (.004)	0.906 (.012)	0.869 (.003)	0.927 (.009)	**0.897** (.003)	0.945 (.007)
RBC-CHR	**0.834** (.005)	0.655 (.010)	0.865 (.005)	0.511 (.009)	0.884 (.003)†	**0.520** (.014)†	0.892 (.003)	0.633 (.005)
RLR-CHR	0.832 (.004)	**0.654** (.010)	0.869 (.005)	**0.508** (.009)	**0.885** (.004)†	0.524 (.016)†	0.892 (.003)	**0.632** (.005)
RLCP	0.825 (.005)	1.018 (.018)†	0.873 (.005)	1.161 (.024)†	0.876 (.004)	0.719 (.008)†	0.890 (.003)	1.055 (.010)†

† The results for Video Time exclude two of the 100 repetitions for RLR-CHR and one for RBC-CHR. RLCP uses the median interval width within each repetition. Further details are discussed in Appendices B.3 and B.9.

5 Conclusion

In this paper, we propose a residual-based conformal prediction framework built on the principle of signal-noise separation. By applying a point predictor to remove the location component and a scale estimator to standardize the residuals, the framework reduces the conditional distribution modeling task to estimating the shape of standardized residuals.

Building on this framework, the proposed RLR-CHR standardizes the residuals, constructs CHR scores on the standardized residual scale, and performs local rank calibration. This local-rank step replaces the existing calibration based on weighted localized quantiles with an unweighted rank test.

RLR-CHR satisfies finite-sample marginal coverage under exchangeability. We establish that its conditional coverage converges asymptotically to the nominal level under local score stability. Under additional regularity conditions, its interval lengths converge to those of the oracle shortest prediction interval, and it is asymptotically equivalent to the counterpart based on weighted localization studied in the Appendix.

Experiments on synthetic and real data show that RLR-CHR improves worst-slab coverage and produces shorter prediction intervals than the existing methods compared on most datasets. The weighted and local-rank calibration rules also give similar coverage and interval lengths.

Several directions remain open for future research. Extending the residual-based framework to non-exchangeable settings, such as covariate shift or temporal dependence, would broaden its applicability. The localizer based on the estimated scale may be less effective when score variation due to skewness, multimodality, or tails is not captured by scale. A learned one-dimensional projection combining scale and higher-moment residual summaries may be a possible extension. Adapting these ideas to structured prediction tasks and multi-output regression is a further topic for future work.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and the area chair for their comments.

The research is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the CANSSI Ontario AI Applications in Statistical Sciences Research (CO AIStats) program. The authors declare no competing interests.

References

Agency for Healthcare Research and Quality. MEPS HC-181: 2015 Full Year Consolidated Data File [Dataset]. Medical Expenditure Panel Survey, 2017. URL https://meps.ahrq.gov/data_stats/download_data/pufs/h181/h181doc.shtml.

Agency for Healthcare Research and Quality. MEPS HC-192: 2016 Full Year Consolidated Data File [Dataset]. Medical Expenditure Panel Survey, 2018. URL https://meps.ahrq.gov/data_stats/download_data/pufs/h192/h192doc.shtml.

Salim I. Amoukou and Nicolas J.-B. Brunel. Adaptive conformal prediction by reweighting nonconformity score. arXiv preprint arXiv:2303.12695, 2023.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021a. doi: 10.1214/20-AOS1965.

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, 2021b. doi: 10.1093/imaiai/iaaa017.

Henrik Boström. Conformal prediction in Python with crepes. In *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*, pages 236–249. PMLR, 2024. URL https://proceedings.mlr.press/v230/bostrom24a.html.

Henrik Boström and Ulf Johansson. Mondrian conformal regressors. In *Proceedings of the Ninth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 128 of *Proceedings of Machine Learning Research*, pages 114–133. PMLR, 2020. URL https://proceedings.mlr.press/v128/bostrom20a.html.

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

Luben M. C. Cabezas, Mateus P. Otto, Rafael Izbicki, and Rafael B. Stern. Regression trees for fast and adaptive prediction intervals. *Information Sciences*, 686:121369, 2025. doi: 10.1016/j.ins.2024.121369.

Maxime Cauchois, Suyash Gupta, and John C. Duchi. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 22 (81):1–42, 2021. URL https://www.jmlr.org/papers/v22/20-753.html.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021. doi: 10.1073/pnas.2107794118.

Anton Conrad, Rustam Isaev, Denis Belomestny, Eric Moulines, and Sergey Samsonov. Beyond marginal validity: Finite-sample guarantees for localized conformal prediction. arXiv preprint arXiv:2608.06206, 2026.

Andrea Coraddu, Luca Oneto, Alessandro Ghio, Stefano Savio, Davide Anguita, and Massimo Figari. Condition Based Maintenance of Naval Propulsion Plants [Dataset]. UCI Machine Learning Repository, 2014. URL https://doi.org/10.24432/C5K31K.

Will Cukierski. Bike Sharing Demand [Dataset]. Kaggle competition, 2014. URL https://www.kaggle.com/competitions/bike-sharing-demand.

Tewodros Deneke. Online Video Characteristics and Transcoding Time Dataset. UCI Machine Learning Repository, 2014. URL https://doi.org/10.24432/C58C9K.

Hadi Fanaee-T and João Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2–3):113–127, 2014. doi: 10.1007/s13748-013-0040-3.

Shai Feldman, Stephen Bates, and Yaniv Romano. Improving conditional coverage via orthogonal quantile regression. In *Advances in Neural Information Processing Systems*, volume 34, pages 2060–2071. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/1006ff12c465532f8c574aeaa4461b16-Abstract.html.

Chao Gao, Liren Shan, Vaidehi Srinivas, and Aravindan Vijayaraghavan. Volume optimality in conformal prediction with structured prediction sets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18495–18527. PMLR, 2025. URL https://proceedings.mlr.press/v267/gao25k.html.

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, 2025. doi: 10.1093/jrsssb/qkaf008.

Leying Guan. Conformal prediction with localization. arXiv preprint arXiv:1908.08558v3, 2019.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023. doi: 10.1093/biomet/asac040.

Xing Han, Ziyang Tang, Joydeep Ghosh, and Qiang Liu. Split localized conformal prediction. arXiv preprint arXiv:2206.13092, 2022.

Bruce E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748, 2008. doi: 10.1017/S0266466608080304.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, second edition, 2009. doi: 10.1007/978-0-387-84858-7.

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

Rafael Izbicki, Gilson Shimizu, and Rafael B. Stern. Flexible distribution-free conditional predictive bands using density estimators. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3068–3077. PMLR, 2020. URL https://proceedings.mlr.press/v108/izbicki20a.html.

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. URL https://www.jmlr.org/papers/v23/20-797.html.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Dk7QQp8jHEo.

Julie Keisler. french_load_consumption [Dataset]. OpenML, data ID 46340, 2024. URL https://www.openml.org/d/46340.

Shayan Kiyani, George Pappas, and Hamed Hassani. Length optimization in conformal prediction. In *Advances in Neural Information Processing Systems*, volume 37, pages 99519–99563. Curran Associates, Inc., 2024a. doi: 10.52202/079017-3158.

Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction with learned features. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 24749–24769. PMLR, 2024b. URL https://proceedings.mlr.press/v235/kiyani24a.html.

Batiste Le Bars and Pierre Humbert. On volume minimization in conformal regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 3030–3053. PMLR, 2025. URL https://proceedings.mlr.press/v267/bars25a.html.

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. doi: 10.1111/rssb.12021.

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

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

Rui Luo and Zhixin Zhou. Conformal thresholded intervals for efficient regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19216–19223, 2025. doi: 10.1609/aaai.v39i18.34115.

Yinjie Min, Liuhua Peng, and Changliang Zou. A unified theory of conditional coverage in conformal prediction with applications. arXiv preprint arXiv:2605.11602, 2026.

Harris Papadopoulos, Vladimir Vovk, and Alex Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011. doi: 10.1613/jair.3198.

Enrique Paredes and Rafael Ballester-Ripoll. SGEMM GPU Kernel Performance [Dataset]. UCI Machine Learning Repository, 2017. URL https://doi.org/10.24432/C5MK70.

Vincent Plassier, Alexander Fishkov, Victor Dheur, Mohsen Guizani, Souhaib Ben Taieb, Maxim Panov, and Eric Moulines. Rectifying conformity scores for better conditional coverage. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 49459–49492. PMLR, 2025. URL https://proceedings.mlr.press/v267/plassier25a.html.

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

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

Raphael Rossellini, Rina Foygel Barber, and Rebecca Willett. Integrating uncertainty awareness into conformalized quantile regression. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1540–1548. PMLR, 2024. URL https://proceedings.mlr.press/v238/rossellini24a.html.

Abdulwahed Salam and Abdelaaziz El Hibaoui. Power Consumption of Tetouan City [Dataset]. UCI Machine Learning Repository, 2018. URL https://doi.org/10.24432/C5B034.

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. In *Advances in Neural Information Processing Systems*, volume 34, pages 6304–6315. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/31b3b31a1c2f8a370206f111127c0dbd-Abstract.html.

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

Daniel Stolfi. Parking Birmingham [Dataset]. UCI Machine Learning Repository, 2017. URL https://doi.org/10.24432/C51K5Z.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html.

Luis Torgo. mv [Dataset]. OpenML, data ID 344, 2014. URL https://www.openml.org/d/344.

Lars van der Laan and Ahmed M. Alaa. Self-calibrating conformal prediction. In *Advances in Neural Information Processing Systems*, volume 37, pages 107138–107170. Curran Associates, Inc., 2024. doi: 10.52202/079017-3402.

Sethu Vijayakumar. SARCOS robot arm inverse dynamics data [Dataset]. Gaussian Processes for Machine Learning website, 2006. URL https://gaussianprocess.org/gpml/data/.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490. PMLR, 2012. URL https://proceedings.mlr.press/v25/vovk12.html.

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

Ran Xie, Rina Foygel Barber, and Emmanuel J. Candès. Boosted conformal prediction intervals. In *Advances in Neural Information Processing Systems*, volume 37, pages 71868–71899. Curran Associates, Inc., 2024. doi: 10.52202/079017-2296.

A Technical details and proofs

A.1 Setup and notation

We use the notation of Sections 2 and 3. The random split into training and calibration sets is independent of the data, and all fitted objects are $\mathcal{F}_n$-measurable.

Neighborhoods use the perturbed coordinate $\widetilde{Z}_i$ of Section 2.2.3, with any remaining ties in distance broken using ξ_j. The coordinates are almost surely distinct, and the neighborhood rule is permutation-equivariant. For a test covariate x, let $\widetilde{Z}(x) = \log \hat{\sigma}(x) + h_n \xi_{n+1}$. When taking a supremum over x, we hold all auxiliary variables ξ_j, $j \in J$, fixed.

Write $E(x, y)$ for the score in (2) evaluated at (x, y), so $E_i = E(X_i, Y_i)$ for $i \in J$. We set $\tilde{S}_T(x) = \mathbb{R}$. By nesting, for $t \in \{1, \ldots, T\}$,

$$y \in \hat{\mu}(x) + \hat{\sigma}(x)\tilde{S}_t(x) \quad \Longleftrightarrow \quad E(x, y) \leq t.$$

Under i.i.d. sampling, the test observation is independent of the calibration sample and auxiliary variables given $\mathcal{F}_n$. Hence, with F_x as defined in Section 3, for any $\mathcal{D}_n$-measurable threshold $\tau(x) \in \{0, 1, \ldots, T\}$,

$$\mathbb{P}(E_{n+1} \leq \tau(x) \mid X_{n+1} = x, \mathcal{D}_n) = F_x(\tau(x)).$$

This identity expresses conditional coverage in terms of the selected score level.

A.2 RBC-CHR construction and algorithm

RBC-CHR uses the same fitted models and standardized residual CHR scores $E_i \in \{1, \ldots, T\}$ as RLR-CHR. It applies Guan's G2 calibration [Guan, 2019] on the same k-nearest-neighbor neighborhoods. G2 selects the probability level of a local score quantile by checking two inequalities on the calibration sample.

For a test covariate x, let $\mathcal{N}_r(x)$ denote the neighborhoods constructed as in Appendix A.1. The normalized binary weights are $p_{rj}(x) = \mathbf{1}\{j \in \mathcal{N}_r(x)\}/(k+1)$. For a finite multiset $\mathcal{V} = \{v_1, \ldots, v_s\}$ and $p \in (0, 1]$, let $Q_p(\mathcal{V})$ denote the left empirical p-quantile:

$$Q_p(\mathcal{V}) := v_{(\lceil ps \rceil)},$$

where $v_{(1)} \leq \cdots \leq v_{(s)}$ are the sorted values.

For each candidate probability level $p \in \mathcal{G} := \{g/(m+1) : g = 1, \ldots, m+1\}$, G2 first computes the local quantile at the test point by appending T to the scores of its calibration neighbors:

$$\bar{t}_x(p) \;:=\; Q_p(\,\{E_j :\; j \in \mathcal{N}_{n+1}(x) \setminus \{n+1\}\} \;\cup\; \{T\}\,).$$

The value T corresponds to $\tilde{S}_T(x) = \mathbb{R}$, while 0 lies below all realized scores. For each calibration point r, G2 forms two corrected quantiles, $v_{r,1}^{(x)}(p)$ and $v_{r,2}^{(x)}(p)$, by setting the test score to $\bar{t}_x(p)$ and 0, respectively, whenever the test point belongs to $\mathcal{N}_r(x)$.

The acceptance fractions $A_1^{(x)}(p)$ and $A_2^{(x)}(p)$ count calibration scores no greater than their respective corrected quantiles, with denominator $m+1$. The second fraction also counts the test point with its score set to zero. Choose $p^*(x)$ as the smallest $p \in \mathcal{G}$ for which both fractions are at least $1 - \alpha$, or $\bar{t}_x(p) = T$, and set $\hat{t}(x) = \bar{t}_x(p^*(x))$. At $x = X_{n+1}$, the selected score index is $t^\star = \hat{t}(x)$. Algorithm 2 gives the complete RBC-CHR procedure. Table 2 compares conformity scores and calibration rules across methods.

Algorithm 2 RBC-CHR with k-NN Local Calibration (G2) [Guan, 2019]

Input: Data $\{(X_i, Y_i)\}_{i=1}^n$ split into $\mathcal{I}_{\text{train}}$ and $\mathcal{I}_{\text{cal}}$; test point X_{n+1}; $\alpha \in (0,1)$; neighborhood size k; number of CHR levels T.

Output: Prediction interval $\widehat{C}^{\text{RBC-CHR}}(X_{n+1})$.

Train $\hat{\mu}$, $\hat{\sigma}$, and the CHR model on $\mathcal{I}_{\text{train}}$. Compute E_i for $i \in \mathcal{I}_{\text{cal}}$ and $\widetilde{Z}_j = \log \hat{\sigma}(X_j) + h_n \xi_j$ for $j \in \mathcal{I}_{\text{cal}} \cup \{n+1\}$.

Construct k-NN neighborhoods in $\widetilde{Z}$-space for $J = \mathcal{I}_{\text{cal}} \cup \{n+1\}$.

Let $m = |\mathcal{I}_{\text{cal}}|$ and write $x = X_{n+1}$.

G2 search:

For $p \in \mathcal{G} := \{g/(m+1) : g = 1, \ldots, m+1\}$:

Compute $\bar{t}_x(p) := Q_p(\{E_\ell : \ell \in \mathcal{N}_{n+1}(J) \setminus \{n+1\}\} \cup \{T\})$, the $\lceil p(k+1) \rceil$-th smallest score after adding T.

Set $v_{r,1}^{(x)}(p) := Q_p(\{E_\ell : \ell \in \mathcal{N}_r(J) \cap \mathcal{I}_{\text{cal}}\} \cup \{\bar{t}_x(p)\}^{\mathbf{1}\{n+1 \in \mathcal{N}_r(J)\}})$ for each $r \in \mathcal{I}_{\text{cal}}$.

Set $v_{r,2}^{(x)}(p) := Q_p(\{E_\ell : \ell \in \mathcal{N}_r(J) \cap \mathcal{I}_{\text{cal}}\} \cup \{0\}^{\mathbf{1}\{n+1 \in \mathcal{N}_r(J)\}})$ for each $r \in \mathcal{I}_{\text{cal}}$.

Compute $A_1^{(x)}(p) := \frac{1}{m+1} \sum_{r \in \mathcal{I}_{\text{cal}}} \mathbf{1}\{E_r \le v_{r,1}^{(x)}(p)\}$.

Compute $A_2^{(x)}(p) := \frac{1}{m+1} \Big(\sum_{r \in \mathcal{I}_{\text{cal}}} \mathbf{1}\{E_r \le v_{r,2}^{(x)}(p)\} + 1 \Big)$.

Choose $p^*(x)$ as the smallest $p \in \mathcal{G}$ such that either $\bar{t}_x(p) = T$, or $A_1^{(x)}(p) \ge 1-\alpha$ and $A_2^{(x)}(p) \ge 1-\alpha$.

Set $t^\star := \bar{t}_x(p^*(x))$.

Return $\widehat{C}^{\text{RBC-CHR}}(X_{n+1}) = \hat{\mu}(X_{n+1}) + \hat{\sigma}(X_{n+1}) \cdot \tilde{S}_{t^\star}(X_{n+1})$.

Table 2: Comparison of conformity scores and calibration rules across methods.

Method	**Conformity score**	**Calibration rule**
Norm-SC	Normalized absolute residual	Global
CHR	CHR score on response	Global
RB-CHR	CHR score on unstandardized residuals	Global
G2-Residual	Normalized absolute residual	G2 weighted
RBC-CHR	CHR score on standardized residuals	G2 weighted
RLR-CHR	Same as RBC-CHR	Local rank

A.3 Finite-sample validity

Theorem 4. *Let $\widehat{C}^{\text{RBC-CHR}}$ denote the prediction set produced by Algorithm 2. Assume that $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable and that all fitted objects used to construct $\widehat{C}^{\text{RBC-CHR}}$ are $\mathcal{F}_n$-measurable. Then, for any $\alpha \in (0,1)$,*

$$\mathbb{P}\Big(Y_{n+1} \in \widehat{C}^{\text{RBC-CHR}}(X_{n+1})\Big) \ge 1 - \alpha.$$

Proof. Condition on $\mathcal{F}_n$. The score function $E(\cdot,\cdot)$ is fixed, and the triples $\{(E_j, \widetilde{Z}_j, \xi_j)\}_{j \in J}$ are exchangeable because the observations are exchangeable and the auxiliary variables are i.i.d. and independent of the data. The neighborhood rule in Appendix A.1 is permutation-equivariant, and so is the induced weight matrix $(p_{rj})_{r,j \in J}$ [Guan, 2023].

Algorithm 2 applies G2 to the bounded scores. Selecting level T returns the whole response space because $\tilde{S}_T(x) = \mathbb{R}$. By the definition of the prediction interval and the G2 exchangeability argument [Guan, 2019, Theorem 3.2],

$$\begin{aligned}\mathbb{P}\big(Y_{n+1} \in \widehat{C}^{\text{RBC-CHR}}(X_{n+1}) \mid \mathcal{F}_n\big) &= \mathbb{P}\big(E_{n+1} \le \hat{t}(X_{n+1}) \mid \mathcal{F}_n\big) \\ &\ge 1 - \alpha.\end{aligned}$$

Taking expectations over $\mathcal{F}_n$ proves the result. □

Proof of Theorem 1 for RLR-CHR. As in the preceding proof, the triples $\{(E_j, \widetilde{Z}_j, \xi_j)\}_{j\in J}$ are exchangeable conditional on $\mathcal{F}_n$. At the true test score $E_{n+1} = E(X_{n+1}, Y_{n+1})$, define

$$W_j := A_j(E_{n+1}), \qquad j \in J.$$

The permutation-equivariant neighborhood construction makes $(W_j)_{j\in J}$ a measurable, permutation-equivariant function of these triples. Thus the local ranks are exchangeable conditional on $\mathcal{F}_n$.

The conformal p-value at the true test score is

$$p_{X_{n+1}}(E_{n+1}) = \frac{1 + \sum_{i\in\mathcal{I}_{\mathrm{cal}}} \mathbf{1}\{W_i \geq W_{n+1}\}}{m+1}.$$

Conditional exchangeability makes this p-value super-uniform. On the event $\{p_{X_{n+1}}(E_{n+1}) > \alpha\}$, the true score is accepted, so $E_{n+1} \leq \hat{e}(X_{n+1})$ by the definition of the selected level. By the definition of the prediction interval,

$$\begin{aligned}\mathbb{P}\big(Y_{n+1} \in \widehat{C}^{\text{RLR-CHR}}(X_{n+1}) \mid \mathcal{F}_n\big) &\geq \mathbb{P}\big(p_{X_{n+1}}(E_{n+1}) > \alpha \mid \mathcal{F}_n\big) \\ &\geq 1 - \alpha.\end{aligned}$$

Taking expectations over $\mathcal{F}_n$ gives the marginal coverage bound. □

A.4 Regularity conditions for asymptotic results

The conditions fall into three tiers according to the conclusion they support.

1. *Marginal coverage* (Theorem 1) requires exchangeability of the observations and measurability of the fitted objects with respect to the training split. No assumptions on the distribution of Y given X are required.
2. *Conditional coverage* (Theorem 2) requires the sampling and tuning conditions of the theorem together with Assumption 1: local stability of the residual CHR score distribution along $\widetilde{Z}$ and vanishing score atoms. The standardized residual distribution may depend on X, provided that points that are neighbours in $\widetilde{Z}$ have similar score distributions. Covariate-dependent skewness is therefore permitted to the extent that it varies slowly along $\widetilde{Z}$, with its effect entering the rate through $\eta_{Z,n}$.
3. *Length optimality and equivalence* (Theorems 3 and 7) require $T = T_n \to \infty$, the sampling and localization conditions stated in Theorem 3, and conditions (E1)–(E4) below on the conditional residual distribution and the approximation of its shortest intervals. Equivalence in Theorem 7, and hence the length conclusion of Theorem 6, also uses the local endpoint bound (C1). The oracle is the shortest conditional interval for Y.

We consider an asymptotic sequence indexed by $n \to \infty$. For each n, the data are i.i.d. and randomly split into disjoint training and calibration index sets $\mathcal{I}_{\mathrm{train}}$ and $\mathcal{I}_{\mathrm{cal}}$. Let $m := |\mathcal{I}_{\mathrm{cal}}|$ and $n_{\mathrm{tr}} := |\mathcal{I}_{\mathrm{train}}|$ denote the calibration and training sample sizes, respectively. Let $T = T_n$ and $k = k_n$. Miscoverage $\alpha \in (0, 1)$ is fixed.

For a real-valued function g defined on a set $\mathcal{S}$, we write

$$\|g\|_\infty := \sup_{s\in\mathcal{S}} |g(s)|.$$

A.4.1 Sampling and localization

(A1) The pairs $\{(X_i, Y_i)\}_{i\geq 1}$ are i.i.d. from a distribution P on $\mathcal{X} \times \mathbb{R}$, where $\mathcal{X} \subset \mathbb{R}^p$ is compact.

Let $\hat{\mu} = \hat{\mu}_n$ and $\hat{\sigma} = \hat{\sigma}_n$ be trained on $\mathcal{I}_{\mathrm{train}}$.

(B1) The fitted objects $(\hat{\mu}, \hat{\sigma}, \{\tilde{S}_t(\cdot)\}_{t=1}^T)$ are $\mathcal{F}_n$-measurable.

(D1) The coordinates used for localization satisfy $\mathbb{P}(\widetilde{Z}_i = \widetilde{Z}_j) = 0$ for $i \neq j$. This holds by construction for the continuous perturbation in Appendix A.1, even when the raw fitted scales have ties.

(D2) The calibration size and neighborhood size satisfy $m \to \infty$, $k \to \infty$, and

$$\frac{k}{m} \to 0, \qquad \frac{\log(T \vee 2)}{k} \to 0, \qquad \frac{k \log(m \vee 2)}{m} \to 0, \qquad \frac{\log(m \vee 2)}{k} \to 0.$$

The last condition in (D2) is used for the uniform bounds in Corollary 2 and Theorems 3, 7, and 6. It controls the additional error from a union bound over the polynomially many neighborhood configurations obtained as the test point varies.

For the conditional coverage bound under Assumption 1, write

$$\Delta_{n,Z}^{\mathrm{RLR}} := \eta_{Z,n} + \gamma_n + \sqrt{\frac{\log(T \vee 2)}{k}} + \sqrt{\frac{k \log(m \vee 2)}{m}} + \frac{k}{m}.$$

Appendix A.6 uses the same rate for RBC-CHR, denoted by $\Delta_{n,Z}^{\mathrm{RBC}}$.

A.4.2 Conditions for length optimality

We use the notation of Section 3 and impose the following conditions.

(E1) *Bounded fitted scale.* There exist constants $0 < \hat{\sigma}_{\min} \le \hat{\sigma}_{\max} < \infty$, independent of n, such that, with probability tending to one,

$$\hat{\sigma}_{\min} \le \hat{\sigma}(x) \le \hat{\sigma}_{\max} \qquad \text{for all } x \in \mathcal{X}.$$

(E2) *Uniformly bounded conditional density.* For every $x \in \mathcal{X}$ the distribution G_x has a density $g(\cdot \mid x)$ with $\kappa := \sup_{x \in \mathcal{X}} \|g(\cdot \mid x)\|_\infty < \infty$.

(E3) *Uniform uniqueness and Lipschitz width.* There is $\eta_0 \in (0, \alpha \wedge (1-\alpha))$ such that for every $x \in \mathcal{X}$ and every $p \in [1-\alpha-\eta_0, 1-\alpha+\eta_0]$ the shortest interval $S_x^*(p)$ is unique, and there is $L_w < \infty$, not depending on x, with $|w_x^*(p) - w_x^*(p')| \le L_w |p - p'|$ for all such p, p'.

(E4) *Conditional CHR consistency.* There is $\rho_{\mathrm{chr},n} \to 0$ with

$$\sup_{x \in \mathcal{X}} \max_{1 \le t \le T-1} \Big(|\tilde{\ell}_t(x) - \ell_x^*(p_t)| + |\tilde{u}_t(x) - u_x^*(p_t)| \Big) = O_{\mathbb{P}}(\rho_{\mathrm{chr},n}).$$

Condition (E4) requires uniform approximation to the shortest conditional residual intervals at every finite grid level. For distributions with unbounded support it presumes a histogram range growing with T; it is a separate estimation condition. For RBC-CHR length optimality and RBC–RLR equivalence, we additionally impose the following local endpoint condition.

(C1) For the fixed $\eta_0 \in (0, \alpha \wedge (1-\alpha))$ of (E3), or (A4b) in the location-scale model, there exists $L_S < \infty$ such that, with probability tending to one,

$$\sup_{x \in \mathcal{X}} \Big(|\tilde{\ell}_{t+1}(x) - \tilde{\ell}_t(x)| + |\tilde{u}_{t+1}(x) - \tilde{u}_t(x)| \Big) \le \frac{L_S}{T}, \quad 1 \le t \le T-2, \quad |p_t - (1-\alpha)| \le \eta_0.$$

A.4.3 Sufficient conditions under a location-scale model

To make (E1)–(E4) more concrete, we give sufficient conditions under a location-scale model and illustrate them with a Gaussian example. Proposition 2 shows that these conditions imply (E1)–(E4) and gives the corresponding approximation rate.

(A2) There exist measurable functions $\mu : \mathcal{X} \to \mathbb{R}$ and $\sigma : \mathcal{X} \to (0, \infty)$ and an error term $\varepsilon \perp X$ with continuous density f such that

$$Y = \mu(X) + \sigma(X)\varepsilon.$$

For $p \in (0,1)$, let $S^*(p) = [\ell^*(p), u^*(p)]$ denote a measurable shortest interval satisfying $\mathbb{P}(\varepsilon \in S^*(p)) = p$. Under this model, the oracle interval is $C^*(x) = \mu(x) + \sigma(x) S^*(1-\alpha)$, with length $L^*(x) = \sigma(x)\{u^*(1-\alpha) - \ell^*(1-\alpha)\}$.

(A3) There exist constants $0 < \sigma_{\min} < \sigma_{\max} < \infty$ such that

$$\sigma_{\min} \le \sigma(x) \le \sigma_{\max} \quad \forall x \in \mathcal{X}.$$

(A4) The following conditions control the behavior of the error density and the oracle endpoints.

(A4a) The density f is globally bounded:

$$\|f\|_\infty := \sup_{z\in\mathbb{R}} f(z) \;<\; \infty.$$

(A4b) There exist $\eta_0 \in (0, \alpha \wedge (1-\alpha))$ and constants $0 < f_{\min} \le f_{\max} < \infty$ such that for all $p \in [1-\alpha-\eta_0,\, 1-\alpha+\eta_0]$,

$$f(\ell^*(p)),\; f(u^*(p)) \in [f_{\min}, f_{\max}],$$

and the maps $p \mapsto \ell^*(p)$ and $p \mapsto u^*(p)$ are Lipschitz on that interval.

(A4c) Let $p_t := t/T$ for $t = 1, \dots, T$ and define the bound on the endpoints at the finite grid levels

$$B_T := 1 + \max_{1\le t\le T-1} \left(|\ell^*(p_t)| \vee |u^*(p_t)|\right).$$

Assume $B_T < \infty$ for each finite T. As B_T may grow with T when the error distribution has unbounded support, we impose the compatibility condition

$$B_T\, r_{\sigma,n} \;\to\; 0.$$

(A5a) For every $p \in [1-\alpha-\eta_0,\, 1-\alpha+\eta_0]$, the oracle shortest interval $S^*(p) = [\ell^*(p), u^*(p)]$ is unique among intervals with probability content p. Strict unimodality of f is sufficient for this uniqueness.

For notational convenience, set $S^*(1) := \mathbb{R}$ with endpoints $\ell^*(1) = -\infty$ and $u^*(1) = +\infty$.

(B2) There exist deterministic sequences $r_{\mu,n} \to 0$ and $r_{\sigma,n} \to 0$ such that

$$\|\hat\mu - \mu\|_\infty = O_{\mathbb{P}}(r_{\mu,n}), \qquad \|\hat\sigma - \sigma\|_\infty = O_{\mathbb{P}}(r_{\sigma,n}),$$

and with probability $\to 1$, $\inf_x \hat\sigma(x) \ge \sigma_{\min}/2$.

(C2) There exists a sequence $r_{\mathrm{chr},n} \to 0$ such that

$$\sup_{x\in\mathcal{X}} \max_{1\le t\le T-1} \Big(|\tilde\ell_t(x) - \ell^*(p_t)| + |\tilde u_t(x) - u^*(p_t)|\Big) = O_{\mathbb{P}}(r_{\mathrm{chr},n}).$$

The level $t = T$ is excluded from this bound because both intervals equal $\mathbb{R}$.

(C3) The grid size satisfies $T = T_n \to \infty$ and $r_{\mathrm{chr},n} \to 0$.

Rates

Under the location-scale conditions, Lemma 3 and Corollary 1 show that $\eta_{Z,n}$ and γ_n in Assumption 1 are both $O_{\mathbb{P}}(\delta_n)$. We therefore use the common rate

$$\Delta_n^{\mathrm{RBC}} = \Delta_n^{\mathrm{RLR}} := \delta_n + \sqrt{\frac{\log(T\vee 2)}{k}} + \sqrt{\frac{k\log(m\vee 2)}{m}} + \frac{k}{m},$$

where

$$\delta_n := r_{\mu,n} + B_T\, r_{\sigma,n} + r_{\mathrm{chr},n} + \frac{1}{T}.$$

The first-stage estimators enter these bounds through their sup-norm rates in (B2). The factor B_T accounts for the effect of scale estimation on the oracle endpoints, and (A4c) ensures that this contribution vanishes as T grows. These rates enter the proof of Corollary 2; their uniform versions below also bound the rates for convergence to oracle length and RBC–RLR equivalence under the location-scale conditions.

For the results that hold uniformly in x under these conditions, define

$$\Delta_{n,\mathrm{unif}}^{\mathrm{RBC}} = \Delta_{n,\mathrm{unif}}^{\mathrm{RLR}} := \delta_n + \sqrt{\frac{\log((T\vee 2)(m\vee 2))}{k}} + \sqrt{\frac{k\log(m\vee 2)}{m}} + \frac{k}{m}.$$

These rates vanish under Condition (D2) together with $\delta_n \to 0$.

Gaussian example

Consider the model

$$X \sim \mathrm{Unif}[0,1], \quad Y = \sin(2\pi X) + (1+X)\varepsilon, \quad \varepsilon \sim N(0,1), \quad \varepsilon \perp X,$$
$$S^*(p) = \left[-\Phi^{-1}\left(\frac{1+p}{2}\right), \Phi^{-1}\left(\frac{1+p}{2}\right)\right].$$

The location and scale are Lipschitz, the scale lies in $[1,2]$, and the Gaussian density satisfies (A4a), (A4b), and (A5a). Here $B_T = 1+\Phi^{-1}(1-1/(2T)) \asymp \sqrt{2\log T}$, so (A4c) requires $\sqrt{\log T}\, r_{\sigma,n} \to 0$.

Take $\hat{\mu} = \mu$ and $\hat{\sigma} = \sigma$, and choose $\tilde{S}_t = S^*(p_t)$ for $t < T$ and $\tilde{S}_T = \mathbb{R}$. The estimation errors in (B2) and (C2) are then zero. The endpoint derivatives are bounded on the fixed window in (C1), and the perturbed localizer $\widetilde{Z} = \log(1+X) + h_n\xi$ has no ties. Taking $T = \lfloor m^{1/4} \rfloor$ and $k = \lfloor m^{1/2} \rfloor$ satisfies (C3) and (D2).

Within the location-scale model, (A4b) and (A5a) impose the main restrictions on the error distribution. They require a unique shortest interval for probability levels near $1-\alpha$, Lipschitz oracle endpoints, and density bounded away from zero at those endpoints. For example, a continuous density that attains its maximum throughout an interval with probability greater than $1-\alpha$ does not satisfy (A5a), because the shortest interval at the target probability is not unique.

Conditions (B2) and (C2) require control of the largest estimation error over the covariate space, including regions with little probability mass. In (C2), the endpoint approximation must also hold at every finite grid level, including levels approaching 0 and 1 as T grows. Uniform rates for local-linear regression are available under additional design, moment, smoothness, and bandwidth conditions [Hansen, 2008].

Appendices B.6 and B.8 give the diagnostic for score stability and the design with covariate-dependent skewness, respectively.

A.5 Sufficiency of the localization coordinate

Assumption 1 is stated for a given localization coordinate, and it is natural to ask what makes a coordinate a good one. The following result treats a continuous coordinate without perturbation.

Proposition 1. *Assume (D2). Conditional on $\mathcal{F}_n$, suppose a measurable $g : \mathcal{X} \to \mathbb{R}$ and a function H satisfy*

$$F_x(t) = H\big(g(x), t\big) \qquad \text{for all } x \in \mathcal{X}, \quad t \in \{0, 1, \ldots, T\}, \tag{A.5.1}$$

where $H(\cdot, t)$ is L-Lipschitz uniformly in t. Suppose also that the conditional distribution of $Z = g(X)$ has compact interval support on which its density is bounded below by a constant $\underline{f} > 0$, uniformly in n. Then, for neighborhoods formed with $\widetilde{Z} = Z$,

$$\eta_{Z,n} = O_{\mathbb{P}}(Lk/m).$$

If, in addition, $\gamma_n = o_{\mathbb{P}}(1)$, then Assumption 1 holds.

Proof. Condition on $\mathcal{F}_n$ and set $N = m+1$. The N localization coordinates in the augmented calibration sample are i.i.d. with distribution function G_n. Thus $U_i = G_n(Z_i)$, $i = 1, \ldots, N$, are i.i.d. $\mathrm{Uniform}(0,1)$. The density lower bound gives

$$|z - z'| \le \underline{f}^{-1} |G_n(z) - G_n(z')|$$

for all z, z' in the support of Z.

Let $U_{(1)} \le \cdots \le U_{(N)}$ be the order statistics, with $U_{(0)} = 0$ and $U_{(N+1)} = 1$. The spacings admit the representation

$$\big(U_{(i)} - U_{(i-1)}\big)_{i=1}^{N+1} \overset{d}{=} \left(\frac{V_i}{\sum_{q=1}^{N+1} V_q}\right)_{i=1}^{N+1},$$

where $V_1, \ldots, V_{N+1}$ are independent $\mathrm{Exp}(1)$ variables. Since G_n preserves the ordering of the coordinates, the image of each k-NN neighborhood spans at most $2k+1$ consecutive spacings. On

the event $\sum_{q=1}^{N+1} V_q \geq (N+1)/2$, every such span is bounded by

$$\frac{2}{N+1} \max_{0 \leq i \leq N} \sum_{q=i+1}^{(i+2k+1)\wedge(N+1)} V_q.$$

The complement of this event has probability at most $\exp(-c_0 N)$ for a universal $c_0 > 0$. For any block of at most $2k+1$ exponential variables and any $a > 0$, Chernoff's inequality with parameter $1/2$ gives

$$\mathbb{P}\left\{\sum V_q > 2(2k+1) + 2a\right\} \leq e^{-a}.$$

Taking $a = 2\log(N \vee 2)$ and applying a union bound over at most $N+1$ blocks yields

$$\max_{0 \leq i \leq N} \{U_{(i+2k+1)\wedge(N+1)} - U_{(i)}\} = O_{\mathbb{P}}\left(\frac{k + \log(N \vee 2)}{N}\right).$$

The density lower bound then yields

$$\sup_{r \in J} \sup_{j \in \mathcal{N}_r(J)} |Z_j - Z_r| = O_{\mathbb{P}}\left(\frac{k + \log(m \vee 2)}{m}\right) = O_{\mathbb{P}}(k/m),$$

where the last equality uses $\log(m \vee 2)/k \to 0$ from (D2). By (A.5.1) and the Lipschitz property of H,

$$\max_{0 \leq t \leq T} |F_{X_j}(t) - F_{X_r}(t)| \leq L|Z_j - Z_r|$$

for every $r \in J$ and $j \in \mathcal{N}_r(J)$. Taking the supremum over centers and their neighbors gives $\eta_{Z,n} = O_{\mathbb{P}}(Lk/m)$. If $\gamma_n = o_{\mathbb{P}}(1)$, Assumption 1 follows. □

For the perturbed coordinate $\widetilde{Z}_i = Z_i + h_n \xi_i$ of Section 2.2.3, the distance from any selected neighbor to its center in the Z coordinate is at most the unperturbed k-NN radius plus $2h_n$. The same argument therefore gives

$$\eta_{Z,n} = O_{\mathbb{P}}(Lk/m + Lh_n).$$

Taking $h_n = O(k/m)$ retains the rate $O_{\mathbb{P}}(Lk/m)$.

Condition (A.5.1) means that $g(X)$ captures the covariate dependence of the conditional score distribution. Assumption 1 also allows this distribution to depend on X beyond $g(X)$, provided that the resulting score CDF variation within neighborhoods vanishes.

The choice $Z = \log \hat{\sigma}(X)$ is appropriate when the remaining score variation is mainly associated with fitted scale. A response atom at y_0 illustrates a possible limitation of this choice. Standardization places the atom at $(y_0 - \hat{\mu}(x))/\hat{\sigma}(x)$, whose position depends on both fitted location and scale. The factorization in (A.5.1) may then fail, and the atom can contribute nonvanishing score CDF variation within neighborhoods formed by fitted scale. Appendix B.6 examines score distributions across groups defined by fitted scale.

A.6 Additional results for RBC-CHR and equivalence

This subsection gives asymptotic results for RBC-CHR analogous to those for RLR-CHR and states the asymptotic equivalence of the two methods.

Recall that $\widehat{C}^{\text{RBC-CHR}}$ is the prediction interval produced by Algorithm 2. For a test covariate $x \in \mathcal{X}$, its length is

$$\widehat{L}^{\text{RBC-CHR}}(x) = \hat{\sigma}(x)\big(\tilde{u}_{\hat{t}(x)}(x) - \tilde{\ell}_{\hat{t}(x)}(x)\big).$$

We use the conditional coverage rate $\Delta_{n,Z}^{\text{RBC}}$ defined in Appendix A.4.

Theorem 5. *Assume the data are i.i.d., the fitted objects are $\mathcal{F}_n$-measurable, Assumption 1 holds, and the localization parameters satisfy $m \to \infty$, $k \to \infty$, $k/m \to 0$, $\log(T \vee 2)/k \to 0$, and $k\log(m \vee 2)/m \to 0$. Then*

$$\left|\mathbb{P}\big(Y_{n+1} \in \widehat{C}^{\text{RBC-CHR}}(X_{n+1}) \mid X_{n+1}, \mathcal{D}_n\big) - (1-\alpha)\right| = O_{\mathbb{P}}(\Delta_{n,Z}^{\text{RBC}}).$$

Theorem 6. *Assume the conditions of Theorem 3 and the local endpoint condition (C1). Let $\delta'_n = \rho_{\mathrm{chr},n} + 1/T$ as in that theorem. Then*

$$\sup_{x\in\mathcal{X}} \left|\widehat{L}^{\text{RBC-CHR}}(x) - L^*(x)\right| = O_{\mathbb{P}}\left(\delta'_n + \sqrt{\frac{\log((T\vee 2)(m\vee 2))}{k}} + \sqrt{\frac{k\log(m\vee 2)}{m}} + \frac{k}{m}\right),$$

where $L^*(x)$ is the length of the shortest interval having conditional probability $1-\alpha$ for Y given $X = x$.

Theorem 7. *Assume the conditions of Theorem 3 and the local endpoint condition (C1). Let $\delta'_n = \rho_{\mathrm{chr},n} + 1/T$ as in that theorem. Then*

$$\begin{aligned}&\sup_{x\in\mathcal{X}} d_H\big(\widehat{C}^{\text{RBC-CHR}}(x), \widehat{C}^{\text{RLR-CHR}}(x)\big)\\ &\qquad = O_{\mathbb{P}}\left(\delta'_n + \sqrt{\frac{\log((T\vee 2)(m\vee 2))}{k}} + \sqrt{\frac{k\log(m\vee 2)}{m}} + \frac{k}{m}\right),\end{aligned}$$

where d_H denotes the Hausdorff distance, and the same bound holds for $\sup_{x\in\mathcal{X}}\big|\widehat{L}^{\text{RBC-CHR}}(x) - \widehat{L}^{\text{RLR-CHR}}(x)\big|$.

A.7 Auxiliary lemmas for asymptotic theory

Define

$$\mathcal{G}_n := \sigma\Big(\mathcal{F}_n,\ \{X_j : j\in\mathcal{I}_{\mathrm{cal}}\},\ \{\xi_j : j\in J\}\Big).$$

Conditional on $\mathcal{G}_n$, the neighborhoods among calibration points are deterministic.

Lemma 1. *Under (A1) and (B1), conditional on $\mathcal{G}_n$, the responses $\{Y_j\}_{j\in\mathcal{I}_{\mathrm{cal}}}$ are independent, and therefore the scores $\{E_j\}_{j\in\mathcal{I}_{\mathrm{cal}}}$ are independent.*

Proof. Given $\mathcal{G}_n$, the covariates $\{X_j\}_{j\in\mathcal{I}_{\mathrm{cal}}}$ and the fitted objects are fixed. Under i.i.d. sampling and the split, the calibration responses $\{Y_j\}$ are independent conditional on $\{X_j\}$ and $\mathcal{F}_n$. Since each E_j is a measurable function of (X_j, Y_j) and the fitted objects, the scores $\{E_j\}$ are independent conditional on $\mathcal{G}_n$. □

Lemma 2. *Under (A2), (A3), and (B2), for any x,*

$$\frac{Y-\hat\mu(x)}{\hat\sigma(x)} = \varepsilon\cdot\frac{\sigma(x)}{\hat\sigma(x)} + \frac{\mu(x)-\hat\mu(x)}{\hat\sigma(x)}.$$

On $\{\inf_x \hat\sigma(x)\ge\sigma_{\min}/2\}$,

$$\sup_x\left|\frac{\mu(x)-\hat\mu(x)}{\hat\sigma(x)}\right| \le \frac{2}{\sigma_{\min}}\|\hat\mu-\mu\|_\infty, \qquad \sup_x\left|\frac{\sigma(x)}{\hat\sigma(x)}-1\right| \le \frac{2}{\sigma_{\min}}\|\hat\sigma-\sigma\|_\infty.$$

Proof. The first identity follows by algebraic rearrangement. The inequalities follow from dividing by $\hat\sigma(x)\ge\sigma_{\min}/2$ and using

$$\left|\frac{\sigma(x)}{\hat\sigma(x)}-1\right| = \frac{|\sigma(x)-\hat\sigma(x)|}{\hat\sigma(x)} \le \frac{2}{\sigma_{\min}}\|\hat\sigma-\sigma\|_\infty.$$

□

Lemma 3. *Assume (A2), (A3), (A4a), (A4c), (B2), (C2)–(C3). Then*

$$\sup_{x\in\mathcal{X}}\max_{1\le t\le T}\left|\mathbb{P}(E\le t \mid X=x,\mathcal{F}_n) - p_t\right| = O_{\mathbb{P}}\left(r_{\mu,n} + B_T\, r_{\sigma,n} + r_{\mathrm{chr},n}\right).$$

In particular,

$$\sup_{x\in\mathcal{X},\ 1\le t\le T}\left|\mathbb{P}(E\le t \mid X=x,\mathcal{F}_n) - p_t\right| = O_{\mathbb{P}}(\delta_n).$$

Proof. Fix $x \in \mathcal{X}$ and $t \in \{1, \dots, T\}$. If $t = T$, then $\{E \le T\}$ holds almost surely and $p_T = 1$, hence the difference is exactly 0. Therefore it suffices to treat $t \in \{1, \dots, T-1\}$.

By definition,

$$\{E \le t\} \iff \{\widetilde{R} \in [\tilde{\ell}_t(x), \tilde{u}_t(x)]\}.$$

By Lemma 2, on $\{\inf_x \hat{\sigma}(x) \ge \sigma_{\min}/2\}$,

$$\widetilde{R} = a_n(x) + b_n(x)\varepsilon, \quad a_n(x) := \frac{\mu(x) - \hat{\mu}(x)}{\hat{\sigma}(x)}, \quad b_n(x) := \frac{\sigma(x)}{\hat{\sigma}(x)}.$$

Moreover,

$$\sup_x |a_n(x)| \le \frac{2}{\sigma_{\min}} \|\hat{\mu} - \mu\|_\infty = O_{\mathbb{P}}(r_{\mu,n}),$$
$$\sup_x |b_n(x) - 1| \le \frac{2}{\sigma_{\min}} \|\hat{\sigma} - \sigma\|_\infty = O_{\mathbb{P}}(r_{\sigma,n}),$$

and $b_n(x) \ge 1/2$ uniformly with probability tending to one for n large enough.

Therefore, for $t \le T - 1$,

$$\mathbb{P}(E \le t \mid X = x, \mathcal{F}_n) = \mathbb{P}\left(\varepsilon \in \left[\frac{\tilde{\ell}_t(x) - a_n(x)}{b_n(x)}, \frac{\tilde{u}_t(x) - a_n(x)}{b_n(x)} \right] \;\middle|\; X = x, \mathcal{F}_n \right).$$

Define

$$\ell_{n,t}(x) := \frac{\tilde{\ell}_t(x) - a_n(x)}{b_n(x)}, \qquad u_{n,t}(x) := \frac{\tilde{u}_t(x) - a_n(x)}{b_n(x)}.$$

We now bound the endpoint perturbations. Using $b_n(x) \ge 1/2$,

$$\left| \frac{\ell^*(p_t)}{b_n(x)} - \ell^*(p_t) \right| \le 2|\ell^*(p_t)|\,|b_n(x) - 1| \le 2(B_T - 1) \sup_x |b_n(x) - 1|,$$

and similarly for $u^*(p_t)$. Hence, uniformly over x and $t \le T - 1$,

$$\left| \frac{\ell^*(p_t)}{b_n(x)} - \ell^*(p_t) \right| + \left| \frac{u^*(p_t)}{b_n(x)} - u^*(p_t) \right| = O_{\mathbb{P}}(B_T r_{\sigma,n}).$$

Using the decomposition

$$\ell_{n,t}(x) - \ell^*(p_t) = \frac{\tilde{\ell}_t(x) - \ell^*(p_t)}{b_n(x)} - \frac{a_n(x)}{b_n(x)} + \left(\frac{\ell^*(p_t)}{b_n(x)} - \ell^*(p_t) \right),$$

we have

$$\sup_{x, t \le T-1} \left| \frac{\tilde{\ell}_t(x) - \ell^*(p_t)}{b_n(x)} \right| \le 2 \sup_{x, t \le T-1} |\tilde{\ell}_t(x) - \ell^*(p_t)| = O_{\mathbb{P}}(r_{\mathrm{chr},n})$$

by (C2), and

$$\sup_x \left| \frac{a_n(x)}{b_n(x)} \right| \le 2 \sup_x |a_n(x)| = O_{\mathbb{P}}(r_{\mu,n}).$$

Combining these bounds with the preceding endpoint perturbation bound, we obtain

$$\sup_{x \in \mathcal{X}} \max_{1 \le t \le T-1} \left(|\ell_{n,t}(x) - \ell^*(p_t)| + |u_{n,t}(x) - u^*(p_t)| \right) = O_{\mathbb{P}}(r_{\mu,n} + B_T r_{\sigma,n} + r_{\mathrm{chr},n}).$$

By (A4a), for any intervals $[a, b]$ and $[a', b']$,

$$|\mathbb{P}(\varepsilon \in [a, b]) - \mathbb{P}(\varepsilon \in [a', b'])| \le \|f\|_\infty \, (|a - a'| + |b - b'|),$$

because $\mathbb{P}(\varepsilon \in [a, b]) = F(b) - F(a)$ and $|F(b) - F(b')| \le \|f\|_\infty \, |b - b'|$.

Applying this with $(a, b) = (\ell_{n,t}(x), u_{n,t}(x))$ and $(a', b') = (\ell^*(p_t), u^*(p_t))$, and then taking the supremum over x and $t \le T - 1$, yields

$$\sup_{x, t \le T-1} |\mathbb{P}(\varepsilon \in [\ell_{n,t}(x), u_{n,t}(x)]) - \mathbb{P}(\varepsilon \in [\ell^*(p_t), u^*(p_t)])| = O_{\mathbb{P}}(r_{\mu,n} + B_T r_{\sigma,n} + r_{\mathrm{chr},n}).$$

But $\mathbb{P}(\varepsilon \in [\ell^*(p_t), u^*(p_t)]) = p_t$ by definition of $S^*(p_t)$, hence for all $t \le T - 1$,

$$\sup_{x \in \mathcal{X}} |\mathbb{P}(E \le t \mid X = x, \mathcal{F}_n) - p_t| = O_{\mathbb{P}}(r_{\mu,n} + B_T r_{\sigma,n} + r_{\mathrm{chr},n}).$$

Combining this bound with the case $t = T$ completes the proof. □

Corollary 1. *Under the assumptions of Lemma 3,*

$$\eta_{Z,n} = O_{\mathbb{P}}(\delta_n), \qquad \gamma_n = O_{\mathbb{P}}(\delta_n).$$

Consequently, under the conditions for the location-scale model in Appendix A.4, Assumption 1 holds whenever $\delta_n \to 0$.

Proof. Let

$$\zeta_n := \sup_{x \in \mathcal{X}} \max_{0 \le t \le T} |F_x(t) - p_t|, \qquad p_t := t/T,$$

where $F_x(0) = p_0 = 0$. Lemma 3 and the definition of δ_n give $\zeta_n = O_{\mathbb{P}}(\delta_n)$. Then, uniformly over $r, j \in \mathcal{N}_r(J)$, and t,

$$\left|F_{X_j}(t) - F_{X_r}(t)\right| \le \left|F_{X_j}(t) - p_t\right| + |F_{X_r}(t) - p_t| \le 2\zeta_n.$$

Thus $\eta_{Z,n} \le 2\zeta_n = O_{\mathbb{P}}(\delta_n)$. Also, uniformly over x and $t = 1, \dots, T$,

$$\mathbb{P}(E = t \mid X = x, \mathcal{F}_n) = F_x(t) - F_x(t-1) \le p_t - p_{t-1} + 2\zeta_n = \frac{1}{T} + 2\zeta_n.$$

Since $1/T \le \delta_n$, $\gamma_n = O_{\mathbb{P}}(\delta_n)$. □

Lemma 4. *Assume (D1). Let (s) denote the index of the s-th smallest localization coordinate in J, so that $\widetilde{Z}_{(1)} < \cdots < \widetilde{Z}_{(m+1)}$. For each $s = 1, \dots, m+1$,*

$$\mathcal{N}_{(s)}(J) \subseteq \mathcal{W}_s := \{(r) : 1 \le r \le m+1,\ |r - s| \le k\}, \qquad |\mathcal{W}_s| \le 2k+1.$$

Neighborhoods whose centers are more than $2k$ ranks apart are disjoint. The test point belongs to at most $2k$ calibration neighborhoods, and each calibration point belongs to at most $2k+1$, including its own.

Proof. If $r > s + k$, the k points at ranks $s+1, \dots, s+k$ satisfy

$$|\widetilde{Z}_{(s+q)} - \widetilde{Z}_{(s)}| < |\widetilde{Z}_{(r)} - \widetilde{Z}_{(s)}|, \qquad q = 1, \dots, k.$$

Hence (r) cannot be among the k nearest neighbors of (s). The argument for $r < s - k$ is symmetric, so $\mathcal{N}_{(s)}(J) \subseteq \mathcal{W}_s$ and $|\mathcal{W}_s| \le 2k+1$.

If $|s - s'| > 2k$, then $\mathcal{W}_s \cap \mathcal{W}_{s'} = \varnothing$, and hence $\mathcal{N}_{(s)}(J) \cap \mathcal{N}_{(s')}(J) = \varnothing$.

For a fixed point (r), membership in $\mathcal{N}_{(s)}(J)$ requires $|s - r| \le k$. Counting the possible calibration centers gives

$$\left|\{(s) \in \mathcal{I}_{\mathrm{cal}} : (r) \in \mathcal{N}_{(s)}(J)\}\right| \le \begin{cases} 2k, & (r) = n+1, \\ 2k+1, & (r) \in \mathcal{I}_{\mathrm{cal}}. \end{cases}$$

The second case includes the neighborhood centered at (r). □

Lemma 5. *If $V_1, \dots, V_N$ are independent with $0 \le V_i \le 1$, then for $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N}(V_i - \mathbb{E}V_i)\right| > \epsilon\right) \le 2\exp(-2N\epsilon^2).$$

Proof. This is Hoeffding's inequality for bounded independent variables. □

Lemma 6. *Let $\mathcal{H}$ be a σ-field. Conditional on $\mathcal{H}$, let $E_1, \dots, E_k$ be independent random variables taking values in $\{1, \dots, T\}$ (not necessarily identically distributed). Define*

$$\widehat{F}_k(t) = \frac{1}{k}\sum_{j=1}^{k} \mathbf{1}\{E_j \le t\}, \qquad F_k(t) = \mathbb{E}[\widehat{F}_k(t) \mid \mathcal{H}].$$

Then for $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{1 \le t \le T} |\widehat{F}_k(t) - F_k(t)| > \epsilon \,\middle|\, \mathcal{H}\right) \le 2T\exp(-2k\epsilon^2),$$

almost surely. Consequently,

$$\mathbb{P}\left(\sup_{1 \le t \le T} |\widehat{F}_k(t) - F_k(t)| > \epsilon\right) \le 2T\exp(-2k\epsilon^2).$$

Proof. Fix $t \in \{1, \dots, T\}$. Conditional on $\mathcal{H}$, $\widehat{F}_k(t)$ is the average of k independent Bernoulli variables $\mathbf{1}\{E_j \le t\}$ with possibly different conditional means. Hoeffding's inequality therefore gives

$$\mathbb{P}\big(|\widehat{F}_k(t) - F_k(t)| > \epsilon \;\big|\; \mathcal{H}\big) \le 2\exp(-2k\epsilon^2).$$

Apply a union bound over $t = 1, \dots, T$ to obtain the conditional bound. Taking expectations yields the unconditional bound. □

Lemma 7. *If* $\sup_t |\widehat{F}(t) - F(t)| \le \epsilon$ *on* $\{1, \dots, T\}$ *and* $\widehat{Q}(p) = \min\{t : \widehat{F}(t) \ge p\}$*, then*

$$|F(\widehat{Q}(p)) - p| \le \epsilon + \Delta_F, \qquad \Delta_F := \max_{1 \le t \le T} (F(t) - F(t-1)),$$

with $F(0) = 0$*.*

Proof. Let $t^* = \widehat{Q}(p)$. Then $\widehat{F}(t^*) \ge p$ and $\widehat{F}(t^* - 1) < p$. Using $|\widehat{F} - F| \le \epsilon$,

$$F(t^*) \ge \widehat{F}(t^*) - \epsilon \ge p - \epsilon,$$

and

$$F(t^* - 1) \le \widehat{F}(t^* - 1) + \epsilon < p + \epsilon.$$

Hence

$$F(t^*) \le F(t^* - 1) + (F(t^*) - F(t^* - 1)) < p + \epsilon + \Delta_F.$$

Combining lower and upper bounds gives $|F(t^*) - p| \le \epsilon + \Delta_F$. □

Lemma 8. *Let* $W_0, W_1, \dots, W_k$ *be real numbers and let* $p \in (0, 1]$*. Define*

$$s := \lceil p(k+1) \rceil \in \{1, \dots, k+1\},$$

and let $\widehat{Q}_{k+1}(p)$ *be the left* p*-quantile of the multiset* $\{W_0, \dots, W_k\}$*, namely the* s*-th order statistic.*

If $s = k + 1$*, then* $\widehat{Q}_{k+1}(p) = \max_{0 \le j \le k} W_j$ *and therefore*

$$\mathbf{1}\{W_0 \le \widehat{Q}_{k+1}(p)\} = 1.$$

If $1 \le s \le k$*, let* $W_{(1)}^- \le \cdots \le W_{(k)}^-$ *denote the order statistics of* $\{W_1, \dots, W_k\}$*, and let* $\widehat{Q}_k(s/k)$ *denote its left* s/k*-quantile. Then*

$$\mathbf{1}\{W_0 \le \widehat{Q}_{k+1}(p)\} = \mathbf{1}\{W_0 \le W_{(s)}^-\} = \mathbf{1}\{W_0 \le \widehat{Q}_k(s/k)\}.$$

Proof. If $s = k + 1$, the quantile is the maximum and the indicator equals 1.

For $1 \le s \le k$, W_0 is no greater than the s-th order statistic exactly when at most $s - 1$ values are strictly smaller than W_0. Since

$$\sum_{j=0}^{k} \mathbf{1}\{W_j < W_0\} = \sum_{j=1}^{k} \mathbf{1}\{W_j < W_0\},$$

this criterion is the same for the full multiset and the multiset with W_0 removed. Hence the indicators are equal. □

Lemma 9. *Let* $U_1, \dots, U_m \in [0, 1]$ *be indexed by increasing* $\widetilde{Z}$*-rank among calibration points. Suppose that, conditional on* $\mathcal{G}_n$*, every subcollection* $\{U_{i_1}, \dots, U_{i_r}\}$ *whose indices are pairwise separated by more than* $2k$ *is jointly independent. Then, for* $\epsilon > 0$*,*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}(U_i - \mathbb{E}[U_i \mid \mathcal{G}_n])\right| > \epsilon \;\middle|\; \mathcal{G}_n\right) \le 2(2k+1)\exp\left(-2\left\lfloor \frac{m}{2k+1} \right\rfloor \epsilon^2\right).$$

In particular,

$$\left|\frac{1}{m}\sum_{i=1}^{m}(U_i - \mathbb{E}[U_i \mid \mathcal{G}_n])\right| = O_{\mathbb{P}}\left(\sqrt{\frac{k\log(m \vee 2)}{m}}\right).$$

Proof. Let $b := 2k+1$ and $q := \lfloor m/b \rfloor$. If $m < b$, then $q = 0$ and the concentration bound is trivial. For $m \geq b$, partition $\{1, \dots, m\}$ into the arithmetic progressions

$$\mathcal{J}_r := \{r,\, r+b,\, r+2b, \dots\} \cap \{1, \dots, m\}, \qquad r = 1, \dots, b.$$

The indices in each $\mathcal{J}_r$ are separated by at least $b > 2k$, so the corresponding variables are jointly independent conditional on $\mathcal{G}_n$. For each r, let

$$\overline{S}_r := \frac{1}{|\mathcal{J}_r|} \sum_{i \in \mathcal{J}_r} \Big(U_i - \mathbb{E}[U_i \mid \mathcal{G}_n] \Big).$$

Since $|\mathcal{J}_r| \geq q$, Lemma 5 gives

$$\mathbb{P}\left(|\overline{S}_r| > \epsilon \;\middle|\; \mathcal{G}_n \right) \leq 2 \exp\big(-2 |\mathcal{J}_r| \epsilon^2 \big) \leq 2 \exp\big(-2 q \epsilon^2 \big).$$

The full average is a convex combination of these averages:

$$\frac{1}{m} \sum_{i=1}^{m} (U_i - \mathbb{E}[U_i \mid \mathcal{G}_n]) = \sum_{r=1}^{b} \frac{|\mathcal{J}_r|}{m}\, \overline{S}_r.$$

Its absolute value is therefore at most $\max_{1 \leq r \leq b} |\overline{S}_r|$. A union bound gives

$$\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} (U_i - \mathbb{E}[U_i \mid \mathcal{G}_n]) \right| > \epsilon \;\middle|\; \mathcal{G}_n \right) \leq \sum_{r=1}^{b} \mathbb{P}\left(|\overline{S}_r| > \epsilon \;\middle|\; \mathcal{G}_n \right) \leq 2b \exp(-2q\epsilon^2).$$

For the rate statement, set $\epsilon = C\sqrt{k \log(m \vee 2)/m}$. When $m \geq b$, $q \asymp m/k$ and $b \leq m$, so the bound tends to zero for sufficiently large C. When $m < b$, the absolute average deviation is at most 1, which gives the same rate. □

Lemma 10. *Let $U \sim \mathrm{Unif}(0,1)$. Conditional on $U = u$, let $B \sim \mathrm{Bin}(k, u)$ and define $A := 1 + B \in \{1, \dots, k+1\}$. Then A is uniform on $\{1, \dots, k+1\}$; in particular, for all $a \in \{1, \dots, k+1\}$,*

$$\mathbb{P}(A \geq a) = \frac{k+2-a}{k+1}.$$

Proof. Write $b := a - 1 \in \{0, \dots, k\}$. The binomial probability mass function is

$$g_b(u) := \mathbb{P}(\mathrm{Bin}(k,u) = b) = \binom{k}{b} u^b (1-u)^{k-b}, \qquad u \in [0,1].$$

Integrating over the uniform distribution gives

$$\mathbb{P}(B = b) = \int_0^1 g_b(u)\, du = \binom{k}{b} \int_0^1 u^b (1-u)^{k-b}\, du,$$

where

$$\int_0^1 u^b (1-u)^{k-b}\, du = \mathrm{B}(b+1, k-b+1) = \frac{\Gamma(b+1)\Gamma(k-b+1)}{\Gamma(k+2)} = \frac{b!(k-b)!}{(k+1)!}.$$

Hence

$$\mathbb{P}(B = b) = \binom{k}{b} \frac{b!(k-b)!}{(k+1)!} = \frac{1}{k+1}.$$

Thus B is uniform on $\{0, \dots, k\}$ and $A = 1 + B$ is uniform on $\{1, \dots, k+1\}$. The survival probability follows by summing the uniform masses. □

Lemma 11. *Let U be a random variable taking values in $[0,1]$ with distribution function G_U satisfying $\mathbb{P}(U = 0) = 0$ and*

$$\sup_{0 \leq u \leq 1} |G_U(u) - u| \leq \epsilon.$$

Let $h : [0,1] \to [0,1]$ be nondecreasing and continuous. Then, for every $\eta \in [0,1]$,

$$\mathbb{E}\, h((U-\eta)_+) \geq \int_0^1 h(u)\, du - (\epsilon + \eta), \qquad \mathbb{E}\, h((U+\eta) \wedge 1) \leq \int_0^1 h(u)\, du + (\epsilon + \eta).$$

Proof. For any nondecreasing continuous $g : [0,1] \to [0,1]$, Stieltjes integration by parts gives

$$\mathbb{E}g(U) - \int_0^1 g(u)\,du = -\int_{(0,1]} (G_U(u) - u)\,dg(u),$$

because $G_U(0) = 0$ and $G_U(1) = 1$. Hence

$$\left|\mathbb{E}g(U) - \int_0^1 g(u)\,du\right| \le \epsilon\{g(1) - g(0)\} \le \epsilon.$$

Apply this bound to $g_-(u) = h((u-\eta)_+)$ and $g_+(u) = h((u+\eta) \wedge 1)$. Since h is nondecreasing and takes values in $[0,1]$,

$$0 \le \int_0^1 h(u)\,du - \int_0^1 g_-(u)\,du \le \eta, \qquad 0 \le \int_0^1 g_+(u)\,du - \int_0^1 h(u)\,du \le \eta.$$

Combining the two displays gives the result. □

Lemma 12. *Assume local score stability in Assumption 1. Fix $i \in \mathcal{I}_{\mathrm{cal}}$ such that $|\mathcal{N}_i(J)| = k+1$ and $n+1 \notin \mathcal{N}_i(J)$. Define*

$$A_i := 1 + \sum_{\ell \in \mathcal{N}_i(J)\setminus\{i\}} 1\{E_\ell \le E_i\}.$$

Then, uniformly over $a \in \{1, \dots, k+1\}$,

$$\left|\mathbb{P}(A_i \ge a \mid \mathcal{G}_n, X_{n+1}) - \frac{k+2-a}{k+1}\right| \le \eta_{Z,n} + \gamma_n.$$

Proof. Condition on $(\mathcal{G}_n, X_{n+1})$ throughout; all expectations below use this conditioning. The neighbor scores are independent by Lemma 1 and independence of the test observation. Let

$$F_i(t) := \mathbb{P}(E_i \le t \mid \mathcal{G}_n, X_{n+1}), \qquad F_\ell(t) := \mathbb{P}(E_\ell \le t \mid \mathcal{G}_n, X_{n+1}).$$

By the definition of $\eta_{Z,n}$,

$$\sup_{\ell \in \mathcal{N}_i(J)} \max_{0 \le t \le T} |F_\ell(t) - F_i(t)| \le \eta_{Z,n}.$$

Conditional on $E_i = t$, the indicators $1\{E_\ell \le t\}$ are independent Bernoulli variables with means $F_\ell(t)$. Thus their sum is stochastically bounded between binomial variables with parameters $(k, (F_i(t) - \eta_{Z,n})_+)$ and $(k, (F_i(t) + \eta_{Z,n}) \wedge 1)$. For

$$h_a(u) := \mathbb{P}\{1 + \mathrm{Bin}(k,u) \ge a\}, \qquad 0 \le u \le 1,$$

which is nondecreasing and continuous in u, this gives

$$\mathbb{E}h_a\{(F_i(E_i) - \eta_{Z,n})_+\} \le \mathbb{P}(A_i \ge a \mid \mathcal{G}_n, X_{n+1}) \le \mathbb{E}h_a\{(F_i(E_i) + \eta_{Z,n}) \wedge 1\}.$$

Set $U_i := F_i(E_i)$. The largest jump of F_i is at most γ_n. The distribution of U_i therefore has Kolmogorov distance at most γ_n from the uniform distribution on $[0,1]$. To see this, for any $u \in [0,1]$, let $s = \max\{t : F_i(t) \le u\}$ with the convention $F_i(0) = 0$. Then

$$\mathbb{P}(U_i \le u \mid \mathcal{G}_n, X_{n+1}) = F_i(s), \qquad 0 \le u - F_i(s) \le \begin{cases} F_i(s+1) - F_i(s), & s < T, \\ 0, & s = T, \end{cases} \le \gamma_n.$$

Lemma 11 therefore yields

$$\left|\mathbb{P}(A_i \ge a \mid \mathcal{G}_n, X_{n+1}) - \int_0^1 h_a(u)\,du\right| \le \eta_{Z,n} + \gamma_n$$

almost surely. Finally, if $U \sim \mathrm{Unif}(0,1)$ and $B \mid U \sim \mathrm{Bin}(k,U)$, then $1+B$ is uniform on $\{1, \dots, k+1\}$ by Lemma 10. Therefore

$$\int_0^1 h_a(u)\,du = \mathbb{P}(1 + B \ge a) = \frac{k+2-a}{k+1}.$$

Since $\eta_{Z,n} + \gamma_n$ does not depend on a, the error bound holds uniformly in a. □

Lemma 13. *Assume (D1). Fix the calibration coordinates $\{\widetilde{Z}_j\}_{j\in\mathcal{I}_{\mathrm{cal}}}$ and the variables used to break distance ties. For a deterministic $z\in\mathbb{R}$, set $\widetilde{Z}_{n+1}=z$ and $J(z)=\mathcal{I}_{\mathrm{cal}}\cup\{n+1\}$. For each $r\in J(z)$, let $\mathcal{N}_r(z)$ contain r and its k nearest neighbors. Define*

$$\mathcal{I}_{\mathrm{near}}(z):=\{r\in\mathcal{I}_{\mathrm{cal}}: n+1\in\mathcal{N}_r(z)\}.$$

As z varies over $\mathbb{R}$, the test neighborhood $\mathcal{N}_{n+1}(z)$, the set $\mathcal{I}_{\mathrm{near}}(z)$, and the full neighborhood system $\{\mathcal{N}_r(z)\}_{r\in J(z)}$ each take at most $C(m+1)(k+1)$ distinct values for a universal constant C. All three are constant on the open intervals between at most $C(m+1)(k+1)$ breakpoints. Moreover,

$$|\mathcal{I}_{\mathrm{near}}(z)|\le 2k+\mathbf{1}\{z\in\{\widetilde{Z}_j: j\in\mathcal{I}_{\mathrm{cal}}\}\}.$$

Proof. If $k=m$, every neighborhood equals $J(z)$ and the claims follow. Assume $k<m$, and order the calibration coordinates as $\widetilde{Z}_{(1)}<\cdots<\widetilde{Z}_{(m)}$.

First consider the test neighborhood. By Lemma 4, its k calibration neighbors lie among at most $2k$ indices adjacent in rank to the insertion position. For $s\in\{0,1,\dots,m\}$, consider

$$z\in(\widetilde{Z}_{(s)},\widetilde{Z}_{(s+1)}),\qquad \widetilde{Z}_{(0)}=-\infty,\quad \widetilde{Z}_{(m+1)}=+\infty.$$

The selected neighbors consist of ℓ points to the left and $k-\ell$ to the right for some $\ell\in\{0,1,\dots,k\}$. As z increases, distances to the left points increase and distances to the right points decrease, so ℓ is nonincreasing. Each change exchanges a selected left point for the next right candidate, at a midpoint $(\widetilde{Z}_{(a)}+\widetilde{Z}_{(b)})/2$. There are at most k changes in each insertion interval, giving at most $(m+1)k$ such breakpoints and $(m+1)(k+1)$ constant pieces over the open insertion intervals.

Next consider $\mathcal{I}_{\mathrm{near}}(z)$. For a calibration center r, let d_r be the distance to its k-th nearest calibration neighbor, with distance ties resolved using the fixed auxiliary variables. The test point is included when $|z-\widetilde{Z}_r|<d_r$ and excluded when $|z-\widetilde{Z}_r|>d_r$; at equality, the tie-breaking rule determines inclusion. Thus $\mathcal{I}_{\mathrm{near}}(z)$ can change only at the at most $2m$ points $\widetilde{Z}_r\pm d_r$. Including the configurations at these points gives $O(m)$ possibilities. At most k centers on either strict side of z can include the test point, and at most one center coincides with it, which proves the bound on $|\mathcal{I}_{\mathrm{near}}(z)|$.

For a center outside $\mathcal{I}_{\mathrm{near}}(z)$, the neighborhood consists entirely of calibration points and is fixed. For a center in $\mathcal{I}_{\mathrm{near}}(z)$, its neighborhood is obtained by adding the test point and removing its last selected calibration neighbor. The full system is therefore determined by $\mathcal{N}_{n+1}(z)$ and $\mathcal{I}_{\mathrm{near}}(z)$.

Let $\mathcal{B}$ be the union of the breakpoints for these two objects and the calibration coordinates. Then

$$|\mathcal{B}|\le(m+1)k+3m.$$

The full system is constant on each open interval between consecutive points of $\mathcal{B}$. Counting the intervals and the values at the breakpoints separately gives at most

$$2|\mathcal{B}|+1\le C(m+1)(k+1)$$

configurations for a universal C, as required. □

Corollary 2. *Assume (A1)–(A4), (B1)–(B2), (C2)–(C3), and (D1)–(D2). Then the following hold.*

(a) For RBC-CHR, the uniform conditional coverage error given $\mathcal{D}_n$ satisfies

$$\sup_{x\in\mathcal{X}}\left|\mathbb{P}\big(Y\in\widehat{C}^{\mathrm{RBC\text{-}CHR}}(x)\mid X=x,\mathcal{D}_n\big)-(1-\alpha)\right|=O_{\mathbb{P}}(\Delta^{\mathrm{RBC}}_{n,\mathrm{unif}}).$$

(b) For RLR-CHR, the uniform conditional coverage error given $\mathcal{D}_n$ satisfies

$$\sup_{x\in\mathcal{X}}\left|\mathbb{P}\big(Y\in\widehat{C}^{\mathrm{RLR\text{-}CHR}}(x)\mid X=x,\mathcal{D}_n\big)-(1-\alpha)\right|=O_{\mathbb{P}}(\Delta^{\mathrm{RLR}}_{n,\mathrm{unif}}).$$

(c) With high probability, the full response space is not selected:

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}\hat{t}(x)<T\right)\to 1,\qquad \mathbb{P}\left(\sup_{x\in\mathcal{X}}\hat{e}(x)<T\right)\to 1.$$

Proof. We first prove (a). Condition on $\mathcal{G}_n$.

By Lemma 13, the full neighborhood system has $S_n \le C(m+1)(k+1)$ distinct configurations as the test coordinate varies. Index them by $s \in \{1, \dots, S_n\}$.

Lemma 13 bounds the number of affected calibration neighborhoods by $2k+\mathbf{1}\{z \in \{\widetilde{Z}_j : j \in \mathcal{I}_{\mathrm{cal}}\}\}$. Every neighborhood not containing the pseudo-test point remains its calibration-only neighborhood, so the rank-window dependence bound is unchanged. If the pseudo-test value coincides with a calibration value, the additional near center contributes only one more normalized summand, so the same $O(k/m)$ error bound applies at these isolated configurations.

For each configuration s, let $\hat{t}_s$ and p_s^* be the RBC threshold and selected grid level obtained by running Algorithm 2 with that fixed configuration. The threshold $\hat{t}_s$ is the same for all pseudo-test covariates x whose $\widetilde{Z}(x)$ induces configuration s, but the conditional score CDF F_x may still depend on x. Under the stated conditions for the location-scale model, Lemma 3 gives

$$\sup_{x \in \mathcal{X}} \max_{0 \le t \le T} |F_x(t) - p_t| = O_{\mathbb{P}}(\delta_n).$$

The argument in the proof of Theorem 5, together with the bounds from Corollary 1, gives

$$|p_s^* - (1-\alpha)| = O_{\mathbb{P}}\left(\delta_n + \sqrt{\frac{\log(T \vee 2)}{k}} + \sqrt{\frac{k \log(m \vee 2)}{m}} + \frac{k}{m}\right),$$

and, uniformly over all x with configuration s,

$$|F_x(\hat{t}_s) - (1-\alpha)| = O_{\mathbb{P}}(\Delta_n^{\mathrm{RBC}}).$$

Taking a union bound over the S_n possible configurations only affects the concentration terms for the finite grid. Since $\log S_n = O(\log(m \vee 2))$, this replaces $\log(T \vee 2)$ by $\log((T \vee 2)(m \vee 2))$ inside the $1/\sqrt{k}$ term. In particular, for configuration s, let $\widehat{F}_s^+$ be the empirical CDF obtained by adding T to the k calibration scores in the test neighborhood. The same simultaneous bound gives

$$\min_{1 \le s \le S_n} \widehat{F}_s^+(T-1) \ge \frac{k}{k+1}(1-\gamma_n) - O_{\mathbb{P}}\left(\sqrt{\frac{\log((T \vee 2)(m \vee 2))}{k}}\right) = 1 - o_{\mathbb{P}}(1).$$

Hence $\bar{t}_s(p) < T$ simultaneously for all configurations and all $p \le 1 - \alpha/2$, with probability tending to one. The uniform bound on the acceptance fractions shows that there is an acceptable grid point within $O_{\mathbb{P}}(\Delta_{n,\mathrm{unif}}^{\mathrm{RBC}})$ of $1-\alpha$, so G2 selects a threshold below T. Therefore

$$\sup_{x \in \mathcal{X}} |F_x(\hat{t}(x)) - (1-\alpha)| = O_{\mathbb{P}}(\Delta_{n,\mathrm{unif}}^{\mathrm{RBC}}).$$

The conditional coverage identity in Appendix A.1 then gives (a). Applying the same union bound to the corresponding RLR bounds for each fixed configuration gives (b).

For (c), Lemma 3 and the uniform coverage bound just established imply

$$\sup_{x \in \mathcal{X}} |p_{\hat{t}(x)} - (1-\alpha)| = O_{\mathbb{P}}(\Delta_{n,\mathrm{unif}}^{\mathrm{RBC}}) = o_{\mathbb{P}}(1).$$

Since $p_T = 1$ and α is fixed, this implies $\sup_x \hat{t}(x) < T$ with probability tending to one. The same argument for RLR gives $\sup_x \hat{e}(x) < T$ with probability tending to one. □

Proposition 2. *Assume (A2)–(A4), (A5a), (B2), and (C2). Then conditions (E1)–(E4) hold with the approximation rate*

$$\rho_{\mathrm{chr},n} = r_{\mathrm{chr},n} + r_{\mu,n} + B_T r_{\sigma,n},$$

where $r_{\mathrm{chr},n}$ is the rate in (C2). In particular, $\delta_n' = \delta_n$.

Proof. Under (A2), the standardized residual satisfies

$$\widetilde{R} = a_n(x) + b_n(x)\varepsilon, \qquad a_n(x) = \{\mu(x) - \hat{\mu}(x)\}/\hat{\sigma}(x), \qquad b_n(x) = \sigma(x)/\hat{\sigma}(x).$$

Thus G_x is obtained from the distribution of ε by an increasing affine transformation, and $S_x^*(p) = a_n(x) + b_n(x)S^*(p)$. Conditions (E1)–(E3) then follow from (A3), (A4a), (A4b), (A5a) and (B2).

In particular, (B2) gives $\sup_x |a_n(x)| = O_{\mathbb{P}}(r_{\mu,n})$ and $\sup_x |b_n(x) - 1| = O_{\mathbb{P}}(r_{\sigma,n})$ on the event $\inf_x \hat{\sigma}(x) \geq \sigma_{\min}/2$.

Condition (E4) follows from (C2) together with

$$\left|\ell_x^*(p_t) - \ell^*(p_t)\right| \leq |a_n(x)| + |b_n(x) - 1|\, B_T,$$

and the same bound for the upper endpoint, where B_T is the endpoint bound of (A4c). □

Remark 3. *The rate of Theorem 3 reduces to $O_{\mathbb{P}}(\Delta_{n,\mathrm{unif}}^{\mathrm{RLR}})$ under the conditions of Proposition 2 and the theorem's sampling and localization conditions. Theorems 6 and 7 additionally require (C1) on the fixed window around $1 - \alpha$.*

Lemma 14. *Under (E2) and (E4),*

$$\sup_{x \in \mathcal{X}} \max_{0 \leq t \leq T} \left|F_x(t) - p_t\right| = O_{\mathbb{P}}(\rho_{\mathrm{chr},n}),$$

where $F_x(t) = \mathbb{P}(E \leq t \mid X = x, \mathcal{F}_n)$ and $p_0 := 0$, $p_T := 1$.

Proof. Fix x and $1 \leq t \leq T - 1$. By the definition of the score in (2), $\{E \leq t\} = \{\widetilde{R} \in \tilde{S}_t(x)\}$, so $F_x(t) = G_x(\tilde{S}_t(x))$. Both $\tilde{S}_t(x)$ and $S_x^*(p_t)$ are intervals, and the symmetric difference of two intervals is contained in the union of two intervals whose lengths are $|\tilde{\ell}_t(x) - \ell_x^*(p_t)|$ and $|\tilde{u}_t(x) - u_x^*(p_t)|$. By (E2) the G_x-measure of a set is at most κ times its Lebesgue measure, whence

$$\left|G_x(\tilde{S}_t(x)) - G_x(S_x^*(p_t))\right| \leq \kappa\Big(|\tilde{\ell}_t(x) - \ell_x^*(p_t)| + |\tilde{u}_t(x) - u_x^*(p_t)|\Big).$$

Since G_x is atomless by (E2), $G_x(S_x^*(p_t)) = p_t$ exactly. Taking the supremum over x and the maximum over $t \leq T-1$ and applying (E4) gives the bound. At the boundary levels, $F_x(0) = p_0 = 0$ and $F_x(T) = p_T = 1$. □

Corollary 3. *Under (E2), (E4), and $T \to \infty$, Assumption 1 holds with $\eta_{Z,n} = O_{\mathbb{P}}(\rho_{\mathrm{chr},n})$ and $\gamma_n = 1/T + O_{\mathbb{P}}(\rho_{\mathrm{chr},n})$, so both are $O_{\mathbb{P}}(\delta_n')$ with $\delta_n' = \rho_{\mathrm{chr},n} + 1/T$.*

Proof. Let $\zeta_n := \sup_{x \in \mathcal{X}} \max_{0 \leq t \leq T} |F_x(t) - p_t|$. Lemma 14 gives $\zeta_n = O_{\mathbb{P}}(\rho_{\mathrm{chr},n})$. For any r and any j in its neighborhood,

$$|F_{X_j}(t) - F_{X_r}(t)| \leq |F_{X_j}(t) - p_t| + |p_t - F_{X_r}(t)| \leq 2\zeta_n.$$

For the score atoms,

$$\mathbb{P}(E = t \mid X = x, \mathcal{F}_n) = F_x(t) - F_x(t-1) \leq p_t - p_{t-1} + 2\zeta_n = 1/T + 2\zeta_n.$$

Taking suprema gives the stated bounds. □

Remark 4. *In Lemma 14, condition (E2) bounds the G_x-measure of a symmetric difference by its length and ensures $G_x(S_x^*(p)) = p$ exactly. More generally, the proof can use the uniform bound $\sup_x G_x(A) \leq \omega(|A|)$ for every interval A, with $\omega(\epsilon) \to 0$. The bounded density condition in (E2) is just the case $\omega(\epsilon) = \kappa\epsilon$.*

A.8 Proofs of the asymptotic theorems

A.8.1 Proof of Theorem 5

Recall that, conditional on $\mathcal{D}_n$ and $X_{n+1} = x$,

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}^{\text{RBC-CHR}}(x) \mid X_{n+1} = x, \mathcal{D}_n\} = F_x(\hat{t}(x)).$$

It suffices to show that $F_x(\hat{t}(x))$ is close to $1 - \alpha$.

Fix $x = X_{n+1}$ and condition on $(\mathcal{G}_n, X_{n+1} = x)$, so all neighborhoods are deterministic. For $p \in \mathcal{G}$, let

$$\bar{t}_x(p) = Q_p\big(\{E_j : j \in \mathcal{N}_{n+1}(x) \setminus \{n+1\}\} \cup \{T\}\big)$$

be the G2 threshold at the test point. If $\bar{t}_x(p) = T$, the returned prediction interval is the full response space.

Step 1: quantile accuracy in the test neighborhood. Let $\widehat{F}_x^+(t)$ be the empirical CDF of the multiset in the previous display and let

$$F_x^+(t) := \mathbb{E}\{\widehat{F}_x^+(t) \mid \mathcal{G}_n, X_{n+1} = x\}$$

be its conditional expectation. The k random scores in the test neighborhood are conditionally independent, while the added value T is deterministic. By Lemma 6,

$$\sup_{0\le t\le T} |\widehat{F}_x^+(t) - F_x^+(t)| = O_{\mathbb{P}}\left(\sqrt{\frac{\log(T\vee 2)}{k}}\right).$$

The largest jump of F_x^+ is at most $\gamma_n + 1/(k+1)$: each calibration score distribution has atoms of mass at most γ_n, and the added value contributes mass $1/(k+1)$ at T. Lemma 7 gives, uniformly over $p \in \mathcal{G}$,

$$|F_x^+(\bar{t}_x(p)) - p| = O_{\mathbb{P}}\left(\sqrt{\frac{\log(T\vee 2)}{k}} + \gamma_n + \frac{1}{k}\right).$$

By local score stability, the average of the conditional CDFs of the k calibration scores differs from F_x by at most $\eta_{Z,n}$. Adding T changes this CDF by at most $1/(k+1)$. Hence

$$\sup_t |F_x^+(t) - F_x(t)| = O_{\mathbb{P}}\left(\eta_{Z,n} + \frac{1}{k}\right),$$

and therefore

$$|F_x(\bar{t}_x(p)) - p| = O_{\mathbb{P}}\left(\eta_{Z,n} + \gamma_n + \sqrt{\frac{\log(T\vee 2)}{k}} + \frac{1}{k}\right). \tag{A.8.RBC.1}$$

Step 2: concentration of the G2 acceptance fractions. For $j \in \{1,2\}$ and $p \in \mathcal{G}$, define

$$U_{r,j}^{(x)}(p) := \mathbf{1}\{E_r \le v_{r,j}^{(x)}(p)\}, \qquad r \in \mathcal{I}_{\mathrm{cal}},$$

so that $A_1^{(x)}(p)$ and $A_2^{(x)}(p)$ are normalized sums of these indicators, with the additional $+1$ term in $A_2^{(x)}(p)$. In one dimension, the kNN neighborhoods are contained in rank windows of width $2k+1$ (Lemma 4). The centers whose neighborhoods contain the pseudo-test index form the set $\mathcal{I}_{\mathrm{near}}(x)$, whose size is at most $2k$. On the complement, the indicators satisfy the conditional independence assumption of Lemma 9. Each indicator is a function of the scores indexed by its center's rank window. Centers pairwise separated by more than $2k$ in $\widetilde{Z}$-rank have disjoint windows (Lemma 4). Conditional on $(\mathcal{G}_n, X_{n+1} = x)$, the calibration scores are jointly independent, so indicators over disjoint windows are jointly conditionally independent. Lemma 9, together with a union bound over $p \in \mathcal{G}$, yields

$$\sup_{p\in\mathcal{G}} \left| A_j^{(x)}(p) - \mathbb{E}\{A_j^{(x)}(p) \mid \mathcal{G}_n, X_{n+1} = x\} \right| = O_{\mathbb{P}}\left(\sqrt{\frac{k\log(m\vee 2)}{m}} + \frac{k}{m}\right), \quad j = 1,2. \tag{A.8.RBC.2}$$

The contribution from the near centers is bounded by $2k/(m+1) = O(k/m)$.

Step 3: expectation of the G2 acceptance fractions. First consider a far center $r \notin \mathcal{I}_{\mathrm{near}}(x)$. Then $n+1 \notin \mathcal{N}_r(x)$, so

$$v_{r,1}^{(x)}(p) = v_{r,2}^{(x)}(p) = Q_p(\{E_j : j \in \mathcal{N}_r(x)\}).$$

This multiset contains E_r itself. To handle this self-inclusion, we use Lemma 8. Except for the boundary case in which the empirical quantile is the maximum, the event $\{E_r \le Q_p(\{E_j : j \in \mathcal{N}_r\})\}$ is equivalent to comparing E_r with a leave-one-out quantile computed from $\mathcal{N}_r \setminus \{r\}$. This quantile is conditionally independent of E_r. Lemmas 6 and 7, together with local score stability and the atom bound, then give

$$\left| \mathbb{E}\{U_{r,j}^{(x)}(p) \mid \mathcal{G}_n, X_{n+1} = x\} - p \right| = O_{\mathbb{P}}\left(\eta_{Z,n} + \gamma_n + \sqrt{\frac{\log(T\vee 2)}{k}} + \frac{1}{k}\right), \quad j = 1,2, \tag{A.8.RBC.3}$$

uniformly in p.

For a near center, $v_{r,2}^{(x)}(p)$ is computed with the test score set to 0; the same leave-one-out argument applies after shifting the index of the order statistic, with an additional $O(1/k)$ boundary error. For $v_{r,1}^{(x)}(p)$, the added value $\bar{t}_x(p)$ may depend on the local scores, so we bound the indicator by 1; the total effect on the normalized average is $O(k/m)$ because there are at most $2k$ near centers. Combining these cases gives

$$\sup_{p\in\mathcal{G}}\left|\mathbb{E}\{A_j^{(x)}(p)\mid \mathcal{G}_n, X_{n+1}=x\}-p\right| = O_{\mathbb{P}}\left(\eta_{Z,n}+\gamma_n+\sqrt{\frac{\log(T\vee 2)}{k}}+\frac{1}{k}+\frac{k}{m}\right), \quad j=1,2. \tag{A.8.RBC.4}$$

Combining (A.8.RBC.2)–(A.8.RBC.4), and using $1/k = o\{\sqrt{\log(T\vee 2)/k}\}$, gives

$$\sup_{p\in\mathcal{G}}|A_j^{(x)}(p)-p| = O_{\mathbb{P}}(\Delta_{n,Z}^{\mathrm{RBC}}), \qquad j=1,2. \tag{A.8.RBC.5}$$

Step 4: locating the selected grid point and handling fallback. The simultaneous bound for the empirical CDF in Step 1 and the atom bound give

$$\widehat{F}_x^{+}(T-1) \geq \frac{k}{k+1}(1-\gamma_n) - O_{\mathbb{P}}\left(\sqrt{\frac{\log(T\vee 2)}{k}}\right) = 1-o_{\mathbb{P}}(1).$$

Thus, with probability tending to one, $\bar{t}_x(p) < T$ for every $p \leq 1-\alpha/2$. By (A.8.RBC.5), both G2 inequalities hold at a grid point $p = 1-\alpha+O_{\mathbb{P}}(\Delta_{n,Z}^{\mathrm{RBC}})$, which is below $1-\alpha/2$ with probability tending to one. G2 therefore stops at the first grid point satisfying both inequalities. The bound in (A.8.RBC.5) and monotonicity of $A_1^{(x)}(p)$ and $A_2^{(x)}(p)$ then give

$$|p^*(x)-(1-\alpha)| = O_{\mathbb{P}}(\Delta_{n,Z}^{\mathrm{RBC}}). \tag{A.8.RBC.6}$$

Finally, evaluate (A.8.RBC.1) at $p=p^*(x)$ and combine with (A.8.RBC.6):

$$|F_x(\hat{t}(x))-(1-\alpha)| = O_{\mathbb{P}}(\Delta_{n,Z}^{\mathrm{RBC}}).$$

This proves the theorem.

A.8.2 Proof of Theorem 2

Fix $x = X_{n+1}$ and condition on $(\mathcal{G}_n, X_{n+1}=x)$. It suffices to prove

$$|F_x(\hat{e}(x))-(1-\alpha)| = O_{\mathbb{P}}(\Delta_{n,Z}^{\mathrm{RLR}}).$$

Step 1: empirical CDF in the test neighborhood. For a candidate score e,

$$A_{n+1}(e) = 1+\sum_{\ell\in\mathcal{N}_{n+1}(J)\setminus\{n+1\}}\mathbf{1}\{E_\ell\leq e\}.$$

Define

$$\widehat{F}_{n+1}(e) := \frac{A_{n+1}(e)-1}{k}.$$

By Lemma 6 and local score stability,

$$\sup_{1\leq e\leq T}|\widehat{F}_{n+1}(e)-F_x(e)| = O_{\mathbb{P}}\left(\eta_{Z,n}+\sqrt{\frac{\log(T\vee 2)}{k}}\right). \tag{A.8.RLR.1}$$

Step 2: remove the calibration centers whose ranks depend on the pseudo-test point. Let

$$\mathcal{I}_{\mathrm{near}} := \{i\in\mathcal{I}_{\mathrm{cal}} : n+1\in\mathcal{N}_i(J)\}, \qquad \mathcal{I}_{\mathrm{far}} := \mathcal{I}_{\mathrm{cal}}\setminus\mathcal{I}_{\mathrm{near}}.$$

By Lemma 4, $|\mathcal{I}_{\mathrm{near}}| \le 2k$. Define

$$\widetilde{p}(e) := \frac{1 + \sum_{i \in \mathcal{I}_{\mathrm{far}}} \mathbf{1}\{A_i \ge A_{n+1}(e)\}}{m+1}.$$

Then

$$\sup_{1 \le e \le T} |p(e) - \widetilde{p}(e)| \le \frac{2k}{m+1}. \tag{A.8.RLR.2}$$

Step 3: approximation of the conformal p-value. For $i \in \mathcal{I}_{\mathrm{far}}$, the local rank A_i does not depend on the candidate score e. Lemma 12 gives, uniformly over $a \in \{1, \dots, k+1\}$,

$$\left| \mathbb{P}(A_i \ge a \mid \mathcal{G}_n, X_{n+1}) - \frac{k+2-a}{k+1} \right| \le \eta_{Z,n} + \gamma_n. \tag{A.8.RLR.3}$$

The indicators $\mathbf{1}\{A_i \ge a\}$, indexed by $\widetilde{Z}$-rank, satisfy the conditional independence assumption of Lemma 9. Each A_i is a function of the scores in the rank window of its center. Far centers pairwise separated by more than $2k$ ranks have disjoint windows (Lemma 4). The calibration scores are jointly independent conditional on $(\mathcal{G}_n, X_{n+1})$, so any such subcollection of indicators is jointly conditionally independent. Lemma 9, followed by a union bound over a, yields

$$\begin{aligned} \sup_{1 \le a \le k+1} \left| \frac{1}{|\mathcal{I}_{\mathrm{far}}|} \sum_{i \in \mathcal{I}_{\mathrm{far}}} \mathbf{1}\{A_i \ge a\} - \mathbb{E}\left[\frac{1}{|\mathcal{I}_{\mathrm{far}}|} \sum_{i \in \mathcal{I}_{\mathrm{far}}} \mathbf{1}\{A_i \ge a\} \mid \mathcal{G}_n, X_{n+1} \right] \right| \\ = O_{\mathbb{P}}\left(\sqrt{\frac{k \log(m \vee 2)}{m}} \right). \end{aligned} \tag{A.8.RLR.4}$$

Combining (A.8.RLR.2)–(A.8.RLR.4) with

$$\frac{k+2-A_{n+1}(e)}{k+1} = 1 - \widehat{F}_{n+1}(e) + O\left(\frac{1}{k}\right)$$

gives, uniformly over e,

$$\left| p(e) - \{1 - \widehat{F}_{n+1}(e)\} \right| = O_{\mathbb{P}}\left(\eta_{Z,n} + \gamma_n + \sqrt{\frac{k \log(m \vee 2)}{m}} + \frac{k}{m} + \frac{1}{k} \right).$$

Using (A.8.RLR.1) and absorbing $1/k$ into the displayed rate, we obtain

$$\sup_{1 \le e \le T} |p(e) - \{1 - F_x(e)\}| = O_{\mathbb{P}}(\Delta_{n,Z}^{\mathrm{RLR}}). \tag{A.8.RLR.5}$$

Step 4: non-emptiness and boundary behavior. Set

$$\varepsilon_n := \sup_{1 \le e \le T} |p(e) - \{1 - F_x(e)\}|.$$

By (A.8.RLR.5), $\varepsilon_n = O_{\mathbb{P}}(\Delta_{n,Z}^{\mathrm{RLR}}) = o_{\mathbb{P}}(1)$. For the remainder of the proof, work on the event $\{\varepsilon_n + \gamma_n < \alpha \wedge (1-\alpha)\}$, whose probability tends to one.

Because $E \in \{1, \dots, T\}$, $F_x(0) = 0$ and $F_x(T) = 1$. Also $F_x(1) \le \gamma_n$. Hence

$$p(1) \ge 1 - F_x(1) - \varepsilon_n \ge 1 - \gamma_n - \varepsilon_n > \alpha$$

and the acceptance set $\{e : p(e) > \alpha\}$ is nonempty. Similarly,

$$p(T) \le 1 - F_x(T) + \varepsilon_n = \varepsilon_n < \alpha$$

and the selected index satisfies $\hat{e}(x) < T$.

Step 5: locate the selected index. By definition, $\hat{e}(x)$ is accepted, so $p(\hat{e}(x)) > \alpha$. Using (A.8.RLR.5),

$$1 - F_x(\hat{e}(x)) \ge \alpha - \varepsilon_n, \qquad \text{hence} \qquad F_x(\hat{e}(x)) \le 1 - \alpha + \varepsilon_n. \tag{A.8.RLR.6}$$

Since $\hat{e}(x) < T$ and $\hat{e}(x)$ is the largest accepted index, $\hat{e}(x)+1$ is rejected: $p(\hat{e}(x)+1) \le \alpha$. Again by (A.8.RLR.5),

$$1 - F_x(\hat{e}(x)+1) \le \alpha + \varepsilon_n, \qquad \text{so} \qquad F_x(\hat{e}(x)+1) \ge 1 - \alpha - \varepsilon_n.$$

The atom bound gives $F_x(\hat{e}(x)+1) - F_x(\hat{e}(x)) \le \gamma_n$. Therefore

$$F_x(\hat{e}(x)) \ge 1 - \alpha - \varepsilon_n - \gamma_n. \tag{A.8.RLR.7}$$

Combining (A.8.RLR.6) and (A.8.RLR.7), and using that γ_n is included in $\Delta_{n,Z}^{\mathrm{RLR}}$, gives

$$|F_x(\hat{e}(x)) - (1-\alpha)| = O_{\mathbb{P}}(\Delta_{n,Z}^{\mathrm{RLR}}).$$

This proves Theorem 2.

A.8.3 Proof of Theorem 3

Proof. Corollary 3 gives the stated bounds on $\eta_{Z,n}$ and γ_n, so Theorem 2 applies under the theorem's sampling and localization conditions. Under (E2) and (E4), Lemma 14 gives the uniform approximation $\sup_x \max_t |F_x(t) - p_t| = O_{\mathbb{P}}(\rho_{\mathrm{chr},n})$. The configuration and concentration argument in the proof of Corollary 2 therefore applies with $\delta'_n = \rho_{\mathrm{chr},n} + 1/T$ in place of δ_n. Write $\Delta'_{n,\mathrm{unif}}$ for $\Delta^{\mathrm{RLR}}_{n,\mathrm{unif}}$ with this substitution. The level $\hat{e}(x)$ selected by local rank calibration then satisfies

$$\sup_{x\in\mathcal{X}} \big|p_{\hat{e}(x)} - (1-\alpha)\big| = O_{\mathbb{P}}\big(\Delta'_{n,\mathrm{unif}}\big). \tag{A.8.1}$$

Next we identify the oracle. Fix x and let $C^*(x)$ be a shortest interval with $\mathbb{P}(Y \in C^*(x) \mid X = x) = 1-\alpha$, so that $L^*(x) = |C^*(x)|$. The map $\phi_x(y) = (y - \hat{\mu}(x))/\hat{\sigma}(x)$ is affine and strictly increasing by (E1), hence a bijection from intervals to intervals that multiplies Lebesgue measure by $1/\hat{\sigma}(x)$ and leaves conditional probabilities unchanged. Consequently ϕ_x maps the shortest $(1-\alpha)$-interval for Y given $X = x$ onto the shortest $(1-\alpha)$-interval for $\widetilde{R}$ given $X = x$, and

$$L^*(x) = \hat{\sigma}(x)\, w^*_x(1-\alpha). \tag{A.8.2}$$

The interval returned by RLR-CHR is $\hat{\mu}(x) + \hat{\sigma}(x)\tilde{S}_{\hat{e}(x)}(x)$, of length $\widehat{L}^{\text{RLR-CHR}}(x) = \hat{\sigma}(x)\,\tilde{w}_{\hat{e}(x)}(x)$ with $\tilde{w}_t(x) = \tilde{u}_t(x) - \tilde{\ell}_t(x)$. Combining this expression with (A.8.2) gives

$$\big|\widehat{L}^{\text{RLR-CHR}}(x) - L^*(x)\big| \le \hat{\sigma}_{\max}\Big(\underbrace{\big|\tilde{w}_{\hat{e}(x)}(x) - w^*_x(p_{\hat{e}(x)})\big|}_{\text{(I)}} + \underbrace{\big|w^*_x(p_{\hat{e}(x)}) - w^*_x(1-\alpha)\big|}_{\text{(II)}}\Big).$$

Term (I) is at most $|\tilde{u}_{\hat{e}}(x) - u^*_x(p_{\hat{e}})| + |\tilde{\ell}_{\hat{e}}(x) - \ell^*_x(p_{\hat{e}})|$, which is $O_{\mathbb{P}}(\rho_{\mathrm{chr},n})$ uniformly in x by (E4) on the event $\sup_{x\in\mathcal{X}} \hat{e}(x) \le T-1$. This event has probability tending to one by (A.8.1), since $p_T = 1$.

Term (II) is at most $L_w|p_{\hat{e}(x)} - (1-\alpha)|$ by (E3), once (A.8.1) places $p_{\hat{e}(x)}$ in the neighborhood $[1-\alpha-\eta_0, 1-\alpha+\eta_0]$ with probability tending to one. It is therefore $O_{\mathbb{P}}(\Delta'_{n,\mathrm{unif}})$ uniformly in x. Adding the two bounds and using $\hat{\sigma}_{\max} < \infty$ gives the stated rate. □

A.8.4 Proof of Theorem 7

Because $\tilde{S}_t(x)$ are intervals, $\widehat{C}^{\text{RBC-CHR}}(x)$ and $\widehat{C}^{\text{RLR-CHR}}(x)$ have the forms

$$\widehat{C}^{\text{RBC-CHR}}(x) = \big[\hat{\mu}(x) + \hat{\sigma}(x)\tilde{\ell}_{\hat{t}(x)}(x),\ \hat{\mu}(x) + \hat{\sigma}(x)\tilde{u}_{\hat{t}(x)}(x)\big],$$

$$\widehat{C}^{\text{RLR-CHR}}(x) = \big[\hat{\mu}(x) + \hat{\sigma}(x)\tilde{\ell}_{\hat{e}(x)}(x),\ \hat{\mu}(x) + \hat{\sigma}(x)\tilde{u}_{\hat{e}(x)}(x)\big].$$

For two closed intervals $[a,b]$ and $[c,d]$, the Hausdorff distance is

$$d_H([a,b],[c,d]) = \max\{|a-c|,\ |b-d|\}. \tag{A.8.3}$$

By (E1), $\sup_x \hat{\sigma}(x) = O_{\mathbb{P}}(1)$. By (C1), for any $1 \le s < t \le T-1$ with $p_s, p_t \in [1-\alpha-\eta_0, 1-\alpha+\eta_0]$,

$$|\tilde{\ell}_t(x) - \tilde{\ell}_s(x)| \le \sum_{r=s}^{t-1} |\tilde{\ell}_{r+1}(x) - \tilde{\ell}_r(x)| \le \frac{L_S}{T}(t-s) = L_S|p_t - p_s|,$$

and similarly,

$$|\tilde{u}_t(x) - \tilde{u}_s(x)| \le L_S|p_t - p_s|. \tag{A.8.4}$$

Control of $\sup_x|p_{\hat{t}(x)} - p_{\hat{e}(x)}|$. As in the proof of Theorem 3, Lemma 14 gives the uniform score approximation under (E2) and (E4), so the proof of Corollary 2 applies with δ_n replaced by δ'_n; write $\Delta'_{n,\mathrm{unif}}$ for the resulting rate. From Corollary 2(a) together with Lemma 14, $\sup_x|p_{\hat{t}(x)} - (1-\alpha)| = O_{\mathbb{P}}(\Delta'_{n,\mathrm{unif}})$; from Corollary 2(b) together with the same lemma, $\sup_x|p_{\hat{e}(x)} - (1-\alpha)| = O_{\mathbb{P}}(\Delta'_{n,\mathrm{unif}})$. Therefore,

$$\sup_{x\in\mathcal{X}}|p_{\hat{t}(x)} - p_{\hat{e}(x)}| \le \sup_x|p_{\hat{t}(x)} - (1-\alpha)| + \sup_x|p_{\hat{e}(x)} - (1-\alpha)| = O_{\mathbb{P}}(\Delta'_{n,\mathrm{unif}}).$$

Since $\Delta'_{n,\mathrm{unif}} \to 0$, both selected probability levels lie in $[1-\alpha-\eta_0/2, 1-\alpha+\eta_0/2]$ uniformly in x with probability tending to one. On this event both indices are below T, and every grid level between them lies in the window of (C1). The endpoint bounds above therefore apply to the selected indices.

Hausdorff distance bound. The absolute difference between the left endpoints satisfies

$$|\hat{\mu}(x)+\hat{\sigma}(x)\tilde{\ell}_{\hat{t}(x)}(x)-\big(\hat{\mu}(x)+\hat{\sigma}(x)\tilde{\ell}_{\hat{e}(x)}(x)\big)| = \hat{\sigma}(x)|\tilde{\ell}_{\hat{t}(x)}(x)-\tilde{\ell}_{\hat{e}(x)}(x)| \le \hat{\sigma}(x)L_S|p_{\hat{t}(x)}-p_{\hat{e}(x)}|.$$

Taking the supremum and using $\sup_x \hat{\sigma}(x) = O_{\mathbb{P}}(1)$ gives

$$\sup_x |\mathrm{left}^{\mathrm{RBC}}(x) - \mathrm{left}^{\mathrm{RLR}}(x)| = O_{\mathbb{P}}(\Delta'_{n,\mathrm{unif}}).$$

Applying (A.8.4) to the right endpoints similarly gives

$$\sup_x |\mathrm{right}^{\mathrm{RBC}}(x) - \mathrm{right}^{\mathrm{RLR}}(x)| = O_{\mathbb{P}}(\Delta'_{n,\mathrm{unif}}).$$

Combining the two endpoint bounds through the identity (A.8.3) gives the Hausdorff bound.

Length difference bound. Recall $\widehat{L}^{\text{RBC-CHR}}(x) = \hat{\sigma}(x)\tilde{w}_{\hat{t}(x)}(x)$ and $\widehat{L}^{\text{RLR-CHR}}(x) = \hat{\sigma}(x)\tilde{w}_{\hat{e}(x)}(x)$. Then

$$\begin{aligned}|\widehat{L}^{\text{RBC-CHR}}(x) - \widehat{L}^{\text{RLR-CHR}}(x)| = \hat{\sigma}(x)|\tilde{w}_{\hat{t}(x)}(x) - \tilde{w}_{\hat{e}(x)}(x)| &\le \hat{\sigma}(x)\Big(|\tilde{u}_{\hat{t}} - \tilde{u}_{\hat{e}}| + |\tilde{\ell}_{\hat{t}} - \tilde{\ell}_{\hat{e}}|\Big) \\ &\le 2\hat{\sigma}(x)L_S|p_{\hat{t}(x)} - p_{\hat{e}(x)}|.\end{aligned}$$

Taking the supremum and using $\sup_x \hat{\sigma}(x) = O_{\mathbb{P}}(1)$ and (A.8.4) gives the length difference bound. □

A.8.5 Proof of Theorem 6

Proof. By the triangle inequality, $\sup_x|\widehat{L}^{\text{RBC-CHR}}(x) - L^*(x)| \le \sup_x|\widehat{L}^{\text{RBC-CHR}}(x) - \widehat{L}^{\text{RLR-CHR}}(x)| + \sup_x|\widehat{L}^{\text{RLR-CHR}}(x) - L^*(x)|$. The first supremum is bounded by Theorem 7 and the second by Theorem 3, whose hypotheses are those assumed here. □

Remark 5. *The localization terms in the conditional coverage bounds of Theorems 2 and 5 are $\sqrt{\log(T\vee 2)/k}$ and $\sqrt{k\log(m\vee 2)/m}$. Balancing them suggests choosing $k \asymp \sqrt{\frac{m\,\log(T\vee 2)}{\log(m\vee 2)}}$, up to integer constraints, which yields*

$$\sqrt{\frac{\log(T\vee 2)}{k}} + \sqrt{\frac{k\log(m\vee 2)}{m}} = O\left(\left(\frac{\log(T\vee 2)\,\log(m\vee 2)}{m}\right)^{1/4}\right).$$

With this choice, k/m is of smaller order. The grid size T should grow so that the grid error $1/T$ is negligible relative to the other error terms. Its choice also affects the logarithmic terms and the approximation error of the fitted CHR intervals.

B Additional experimental details and complete numerical results

Unless otherwise stated, experiments use 100 repetitions, and we report means across repetitions with standard errors (SEs) in parentheses. These SEs reflect variability from data sampling, model fitting, and evaluation. CHR-based methods on real data use 1,000 equal-width histogram bins and $T = 1{,}000$ score levels by default. Relative length is computed within each repetition as the ratio of a method's mean interval length to CHR's mean interval length, and these ratios are then averaged across repetitions.

B.1 Datasets and implementation

Datasets. The main experiments use eight regression datasets. We predict the first torque in SARCOS and exclude the remaining torques. For SGEMM, the response is the mean of four run times, and we use a fixed random subsample of 50,000 observations. Tetouan and Parking Birmingham use Zone 1 power consumption and parking occupancy as their responses, respectively. In both datasets, timestamps are converted to calendar features. We exclude the simultaneous Zone 2 and Zone 3 power consumption variables from Tetouan and the car park identifier from Parking Birmingham. For Video Transcoding, we predict transcoding user time and exclude the memory outcome from the same run and the video identifier. For Naval, we predict the compressor decay coefficient and exclude the turbine decay coefficient. MV uses the given response y and numeric covariates only. French Load uses the electricity load response from OpenML dataset 46340, retaining the source table's row index column as a covariate.

We also consider eight datasets from the CQR and CHR studies [Romano et al., 2019, Sesia and Romano, 2021]: Bike Sharing Demand [Cukierski, 2014, Fanaee-T and Gama, 2014], CASP Protein [Rana, 2013], MEPS panels 19–21 [Agency for Healthcare Research and Quality, 2017, 2018], Facebook Comment Volume variants 1 and 2 [Singh, 2015], and BlogFeedback [Buza, 2014]. The responses are bike rental counts for Bike, root mean square deviation (RMSD) for CASP, healthcare utilization for MEPS, and comment counts for Facebook and BlogFeedback. For Bike, we use the Kaggle training file and follow the CQR preprocessing: we convert its timestamp to calendar features, encode season and weather as indicator variables, and remove the casual and registered rental counts from the covariates. For MEPS, we use the preprocessed data files provided in the SLCP code repository. These files are derived from AHRQ public-use data. Data preparation for CASP, Facebook, and BlogFeedback follows CQR.

SGEMM, Tetouan, Parking Birmingham, Video Transcoding, Naval, CASP, Facebook, and BlogFeedback are from the UCI Machine Learning Repository under CC BY 4.0. MV and French Load are from OpenML, where both are listed as "Public". SARCOS is from the data page of *Gaussian Processes for Machine Learning*, which does not list an explicit license. Bike is subject to the Kaggle competition rules. MEPS is subject to the AHRQ data use agreement.

Real data fitting. In each real data repetition, 2,000 observations are used for testing, $\lfloor 0.25N \rfloor$ for calibration, and the remainder for training. Methods use the same data split within each repetition. Numeric covariates are min-max scaled and responses are standardized using the training split only. Interval endpoints are transformed back to the response scale. The neural network regressors have three hidden layers with 64 ReLU units per layer and dropout rate 0.1. They are trained with Adam using learning rate 5×10^{-4} and batch size 128. These network settings are held fixed across the real datasets. A 10% validation split is used for early stopping based on validation loss, with patience 10 and at most 200 epochs. Cross-fitting uses five folds. Split-CP, Norm-SC and MCR share a point network $\hat{f}$ fitted on the full training split; Norm-SC's scale is fitted to that network's in-sample absolute residuals.

The scale models for G2-Residual, RBC-CHR and RLR-CHR are fitted to out-of-fold absolute training residuals. We apply clipping and median rescaling to the scale predictions, with the bounds and rescaling factor estimated from training data only. Calibration and test predictions are obtained from a refit on the full training split and undergo the same transformations.

Histogram construction. For CHR-based methods, the real data experiments follow the released configuration of Sesia and Romano [2021], with 99 quantile heads at $0.01, \ldots, 0.99$. Bin edges cover the training range of the standardized target, extended by 1% on each side. We make the

fitted quantiles nondecreasing using a running maximum, then convert them into histogram bin probabilities. The downward nesting step clips the randomized interval by at most one boundary bin to preserve nesting. Response standardization and this nesting clip are departures from the released implementation.

Calibration. For G2-Residual, RBC-CHR and RLR-CHR, the neighborhood size k is

$$k = \mathrm{round}\left[\min\left\{300, \lfloor m/10 \rfloor, \max\left\{k_{\min}, \sqrt{\frac{m \log T}{\log m}}\right\}\right\}\right], \qquad k_{\min} = \begin{cases} 20, & m < 300, \\ 30, & m \geq 300. \end{cases}$$

Ties in the localization coordinate are broken by independent jitter as in Appendix A.1, while ties in the conformity scores are retained.

CQR uses its own quantile network to estimate the conditional quantiles at levels 0.05 and 0.95 when $\alpha = 0.1$. A CQR interval with reversed endpoints is empty and has length zero. MCR uses `crepes` [Boström, 2024]. In the real data implementation, its difficulty estimate averages the absolute training residuals over 25 nearest neighbors; calibration difficulties define 20 Mondrian groups, within which the residuals are calibrated.

As in the released CHR implementation, RLR-CHR restricts $\tilde{S}_T(x)$ to the range of the histogram, and residuals outside this interval receive score $T + 1$. The search stops at level T, giving marginal coverage of at least $1 - \alpha - \mathbb{P}(E_{n+1} = T + 1)$. In the main runs, the largest observed fraction of calibration scores exceeding T was about 0.1%.

Synthetic experiment. In the synthetic experiment in Section 4.1, each generated sample is split evenly into training and calibration sets and evaluated on 1,000 independent test points. The synthetic CHR implementation uses 197 quantile heads and $\Delta_\alpha = 0.001$, following the released CHR configuration, with 1,000 response grid points. We use full-batch training for the CHR quantile networks, with a learning rate of 5×10^{-4} and a dropout rate of 0.1. The number of epochs is selected using the median validation loss over ten 80/20 splits of the training data, with a maximum of 2,000 epochs. The nuisance fits use 500 epochs. The scale fit uses out-of-fold absolute residuals. The localized procedures use $k = \max\{20, \mathrm{round}\,\sqrt{m}\}$.

Software. The real data experiments use R and Rcpp with TensorFlow and Keras, and the synthetic experiments use Python with PyTorch. The code and its documentation provide the full numerical settings, software versions, seed rules, environment setup, and execution instructions. CQR's loaders use the MIT license, SLCP and `crepes` use BSD-3-Clause, and the upstream AIF360 preparation scripts use Apache 2.0. The CHR and RLCP source repositories contain no explicit license statement.

B.2 Complete results for the main datasets

Tables 3–10 report the complete real data results for the methods in Table 1. Length and SD(Len.) are the mean and standard deviation across repetitions of the mean interval length in each test set. For RLCP, these summaries use the median interval width within each repetition, following the convention in Appendix B.9.

For methods other than RLCP, we flag a repetition when its mean interval length is non-finite or exceeds five times CHR's mean length in the same repetition. We also flag CQR repetitions in which more than 1% of intervals are empty. Daggered entries summarize only unflagged repetitions, with the same repetitions used for coverage, WSC, length, and their SEs. SE denotes the standard deviation across repetitions divided by the square root of the number of repetitions included in the summary. Summaries over all repetitions are given alongside the affected studies.

Table 3: Complete results for SARCOS.

Method	Coverage	SE	Length	SD(Len.)	Rel. length	WSC	SE(WSC)
Split-CP	0.9004	0.0008	11.90	0.4491	1.0614	0.8605	0.0043
Norm-SC	0.8989	0.0007	10.78	0.3869	0.9617	0.8840	0.0025
CQR	0.9000	0.0008	11.07	0.4234	0.9877	0.8872	0.0028
CHR	0.9001	0.0007	11.22	0.4221	1.0000	0.8843	0.0030
MCR	0.9007	0.0008	10.20	0.3212	0.9100	0.9024	0.0026
G2-Residual	0.9062	0.0007	10.38	0.2342	0.9265	0.8946	0.0029
RBC-CHR	0.9058	0.0007	8.860	0.2491	0.7909	0.8975	0.0028
RLR-CHR	0.9046	0.0007	8.833	0.2480	0.7884	0.8960	0.0028
RLCP	0.9011	0.0007	10.85	0.4113	0.9679	0.8929	0.0025

Table 4: Complete results for SGEMM GPU.

Method	Coverage	SE	Length	SD(Len.)	Rel. length	WSC	SE(WSC)
Split-CP	0.9009	0.0007	86.40	9.287	1.3335	0.7960	0.0057
Norm-SC	0.8999	0.0007	68.32	7.332	1.0542	0.8934	0.0029
CQR	0.9010	0.0007	81.03	5.596	1.2511	0.8811	0.0034
CHR	0.9008	0.0007	64.96	4.155	1.0000	0.8919	0.0028
MCR	0.9019	0.0007	82.25	8.293	1.2696	0.8964	0.0025
G2-Residual	0.9079	0.0006	68.14	3.962	1.0528	0.9028	0.0027
RBC-CHR	0.9083	0.0006	47.46	4.142	0.7333	0.8983	0.0025
RLR-CHR	0.9076	0.0006	47.40	4.514	0.7324	0.8966	0.0027
RLCP	0.9006	0.0007	78.28	10.02	1.2089	0.8666	0.0037

Table 5: Complete results for Tetouan City Power Consumption Zone 1.

Method	Coverage	SE	Length	SD(Len.)	Rel. length	WSC	SE(WSC)
Split-CP	0.8998	0.0008	4728	182.2	1.0701	0.8718	0.0035
Norm-SC	0.8998	0.0008	4435	163.9	1.0036	0.8867	0.0027
CQR	0.9001	0.0007	4481	147.4	1.0143	0.8884	0.0026
CHR	0.9003	0.0007	4424	173.1	1.0000	0.8895	0.0027
MCR	0.9010	0.0007	4385	145.8	0.9926	0.8919	0.0027
G2-Residual	0.9073	0.0008	4420	115.3	1.0004	0.8929	0.0028
RBC-CHR	0.9064	0.0008	3816	102.3	0.8638	0.8936	0.0030
RLR-CHR	0.9079	0.0007	3837	85.17	0.8687	0.8986	0.0028
RLCP	0.8994	0.0007	4619	153.9	1.0457	0.8985	0.0027

Table 6: Complete results for Parking Birmingham.

Method	Coverage	SE	Length	SD(Len.)	Rel. length	WSC	SE(WSC)
Split-CP	0.9008	0.0007	612.0	35.03	1.2001	0.7657	0.0069
Norm-SC	0.9002	0.0007	573.7	35.38	1.1245	0.8340	0.0063
CQR	0.8994	0.0007	548.5	23.51	1.0752	0.8763	0.0032
CHR	0.9010	0.0006	511.6	28.77	1.0000	0.8900	0.0029
MCR	0.9024	0.0007	587.7	24.49	1.1524	0.8854	0.0030
G2-Residual	0.9049	0.0007	572.5	19.24	1.1225	0.8677	0.0043
RBC-CHR	0.9043	0.0007	453.5	18.76	0.8891	0.8863	0.0032
RLR-CHR	0.9032	0.0007	452.2	18.67	0.8865	0.8814	0.0032
RLCP	0.9001	0.0007	574.6	29.96	1.1264	0.8771	0.0032

Table 7: Complete results for Naval Propulsion Compressor Decay.

Method	Coverage	SE	Length	SD(Len.)	Rel. length	WSC	SE(WSC)
Split-CP	0.9001	0.0008	0.005951	0.0006871	1.0874	0.7264	0.0066
Norm-SC	0.9010	0.0009	0.005545	0.0006467	1.0135	0.7337	0.0076
CQR	0.9010	0.0009	0.005864	0.0009194	1.0661	0.7584	0.0066
CHR	0.9008	0.0008	0.005567	0.0007463	1.0000	0.7339	0.0069
MCR	0.9058	0.0008	0.004263	0.0004726	0.7789	0.7991	0.0051
G2-Residual	0.9077	0.0008	0.005857	0.0004963	1.0704	0.8108	0.0055
RBC-CHR	0.9099	0.0008	0.003592	0.0003356	0.6554	0.8342	0.0045
RLR-CHR	0.9083	0.0008	0.003586	0.0003367	0.6541	0.8319	0.0044
RLCP	0.9003	0.0007	0.005580	0.0007600	1.0177	0.8246	0.0050

Table 8: Complete results for MV.

Method	Coverage	SE	Length	SD(Len.)	Rel. length	WSC	SE(WSC)
Split-CP	0.8998	0.0007	2.896	0.5664	1.2689	0.7264	0.0097
Norm-SC	0.8990	0.0007	2.212	0.4064	0.9682	0.7900	0.0061
CQR	0.9000	0.0007	2.587	0.2902	1.1364	0.7004	0.0087
CHR	0.9000	0.0007	2.317	0.3119	1.0000	0.7071	0.0084
MCR	0.9014	0.0007	2.278	0.4446	0.9958	0.8823	0.0038
G2-Residual	0.9099	0.0007	2.067	0.1553	0.9059	0.8780	0.0037
RBC-CHR	0.9090	0.0007	1.165	0.1533	0.5106	0.8653	0.0053
RLR-CHR	0.9083	0.0007	1.158	0.1521	0.5077	0.8688	0.0052
RLCP	0.9004	0.0007	2.655	0.4669	1.1610	0.8726	0.0045

Table 9: Complete results for Video Time.

Method	Coverage	SE	Length	SD(Len.)	Rel. length	WSC	SE(WSC)
Split-CP	0.9004	0.0007	5.730	0.4294	1.0389	0.6113	0.0065
Norm-SC	0.9009	0.0007	4.969	0.3102	0.9004	0.8250	0.0045
CQR	0.8994	0.0007	5.770	0.3568	1.0447	0.7297	0.0092
CHR	0.9017	0.0006	5.567	0.5462	1.0000	0.7677	0.0101
MCR	0.9004	0.0007	5.723	0.3476	1.0370	0.8731	0.0031
G2-Residual	0.9023	0.0007	5.113	0.2361	0.9266	0.8693	0.0035
RBC-CHR	0.9037†	0.0012†	2.872†	0.6943†	0.5204†	0.8840†	0.0034†
RLR-CHR	0.9057†	0.0013†	2.906†	0.9314†	0.5245†	0.8853†	0.0036†
RLCP	0.9001	0.0007	3.968	0.2829	0.7191	0.8760	0.0035

† The results for Video Time exclude 1 RBC-CHR and 2 RLR-CHR repetitions; see Appendix B.3 for values computed over all repetitions.

Table 10: Complete results for French Load Consumption.

Method	Coverage	SE	Length	SD(Len.)	Rel. length	WSC	SE(WSC)
Split-CP	0.9001	0.0007	2168000	152100	1.1004	0.8418	0.0040
Norm-SC	0.9004	0.0007	1974000	128300	1.0018	0.8898	0.0029
CQR	0.8986	0.0008	2060000	111900	1.0451	0.8546	0.0041
CHR	0.8999	0.0006	1979000	127100	1.0000	0.8802	0.0033
MCR	0.9003	0.0007	1940000	132000	0.9845	0.8964	0.0024
G2-Residual	0.9027	0.0007	1862000	62210	0.9449	0.8968	0.0027
RBC-CHR	0.9037	0.0007	1249000	50600	0.6335	0.8922	0.0030
RLR-CHR	0.9031	0.0007	1246000	50400	0.6319	0.8918	0.0030
RLCP	0.8997	0.0007	2078000	130300	1.0547	0.8898	0.0028

B.3 Score ties and degenerate repetitions

On Video Time, RLR-CHR returns the full range covered by the residual histogram at every test point in two of the 100 repetitions, with coverage and WSC equal to one. RBC-CHR returns a single histogram bin in the same repetitions, only one of which has an RBC-CHR mean length more than five times CHR's.

Over all 100 repetitions, RLR-CHR has WSC 0.888, marginal coverage 0.9076, mean length $1{,}190$ (SD 1.107×10^4), and mean relative length 206.2 (median 0.502). RBC-CHR has WSC 0.885, marginal coverage 0.9046, mean length 3.95 (SD 10.78), and mean relative length 0.705 (median 0.499).

The independent refits for the density study described later in Appendix B.5 show the same behavior on Video Time in three of the 100 repetitions. In the 100 repetitions of the BlogFeedback bin study, RLR-CHR has three degenerate repetitions with 100 bins and one with 1,000 bins (Table 15).

An overly wide histogram support can place most standardized residuals in one bin, creating a large atom at the lowest score level and violating the small-atom condition in Assumption 1. Since local ranks count scores less than or equal to the center score, ties can produce many maximal local ranks and cause the rank test to accept $e = T$, returning the whole histogram support. This score concentration can arise without large response atoms: the most frequent response value in Video Time accounts for only 0.09% of observations.

After fitting, this mechanism can be diagnosed by inspecting the frequencies of calibration scores, especially the share at level one, and comparing the histogram support width with the residual interquartile range. For a new covariate x, the rank test at $e = T$ provides a direct check: if $p(T) > \alpha$, RLR-CHR returns the whole histogram support. This check uses calibration scores and the new covariate, without requiring the test response.

Remark 6. *The effect of these score atoms can be removed by randomized score tie-breaking together with the corresponding randomized boundary inclusion when constructing prediction intervals, following the randomization principle of CHR [Sesia and Romano, 2021].*

B.4 Results on additional regression datasets

We also evaluate the eight methods on datasets from the CQR and CHR studies [Romano et al., 2019, Sesia and Romano, 2021], using the configuration in Appendix B.1 without dataset-specific tuning. Community, Concrete, and Star are omitted because they are too small for the 2,000-point test split. The main results use $\delta = 0.1$. For these additional regression datasets, we evaluate WSC at $\delta \in \{0.05, 0.10, 0.20\}$ with the same splits and search directions across slab masses.

Table 11 reports WSC and relative length, with datasets ordered by the fraction of responses equal to zero. RLR-CHR maintains WSC near the nominal level, while its relative length ranges from 1.08 to 1.84. The largest length costs occur on datasets with many zero responses.

Zero inflation can lead to longer intervals for RLR-CHR by making the residual histogram harder to estimate, which runs counter to the goal of progressive simplification. Standardization maps an atom at a fixed response value y_0 to $(y_0 - \hat{\mu}(x))/\hat{\sigma}(x)$, whose location can vary with x. The residual

histogram must then track both this location and the atom's conditional probability, whereas CHR models the atom at a fixed response value. Such distributions also violate condition (E2) for length optimality, which requires a uniformly bounded conditional density (Appendix A.4). The presence of large response atoms can be easily checked before fitting.

Table 11: WSC and relative interval length on the additional datasets at $\alpha = \delta = 0.1$. Zeros is the percentage of responses equal to zero in the full dataset. WSC values closest to 0.9 and the smallest relative lengths are bold.

	Bike Sharing		**Protein (CASP)**		**MEPS 20**		**MEPS 21**	
Zeros (%)	0.0		0.6		26.5		27.3	
Method	WSC	Rel. length	WSC	Rel. length	WSC	Rel. length	WSC	Rel. length
Split-CP	0.831 (.004)	1.267 (.009)	0.881 (.003)	1.319 (.003)	0.569 (.007)	1.461 (.016)	0.544 (.007)	1.524 (.017)
Norm-SC	0.886 (.003)	1.103 (.007)	0.891 (.003)	1.099 (.003)	0.861 (.005)	1.152 (.013)	0.864 (.004)	1.177 (.014)
CQR	0.846 (.005)	1.096 (.007)	0.885 (.003)	1.137 (.002)	0.882 (.003)	1.254 (.008)	0.883 (.004)	1.249 (.008)
CHR	0.847 (.005)	**1.000** (.000)	0.894 (.003)	**1.000** (.000)	0.890 (.003)	**1.000** (.000)	0.889 (.003)	**1.000** (.000)
MCR	0.894 (.003)	1.205 (.008)	0.891 (.003)	1.075 (.002)	0.889 (.003)	1.557 (.016)	0.883 (.003)	1.534 (.019)
G2-Residual	**0.900** (.003)	1.209 (.009)	0.894 (.002)	1.119 (.002)	0.887 (.003)	1.496 (.016)	**0.906** (.003)	1.731 (.017)
RBC-CHR	0.897 (.003)	1.188 (.010)	**0.898** (.003)	1.085 (.003)	0.898 (.003)	1.422 (.014)	0.914 (.002)	1.739 (.027)
RLR-CHR	0.901 (.003)	1.202 (.009)	0.892 (.003)	1.081 (.003)	**0.899** (.003)	1.416 (.014)	0.910 (.002)	1.606 (.017)

	MEPS 19		**Facebook 1**		**Facebook 2**		**BlogFeedback**	
Zeros (%)	29.6		55.1		55.4		64.0	
Method	WSC	Rel. length	WSC	Rel. length	WSC	Rel. length	WSC	Rel. length
Split-CP	0.534 (.007)	1.495 (.019)	0.578 (.006)	1.306 (.017)	0.562 (.006)	1.229 (.012)	0.651 (.006)	1.369 (.023)
Norm-SC	0.863 (.005)	1.158 (.013)	0.759 (.006)	**0.985** (.011)	0.784 (.006)	**0.930** (.010)	0.820 (.005)	1.092 (.018)
CQR	0.884 (.003)	1.233 (.009)	0.850 (.006)	1.249 (.010)	0.864 (.004)	1.155 (.007)	0.870 (.004)	1.204 (.015)
CHR	**0.894** (.003)	**1.000** (.000)	0.889 (.003)	1.000 (.000)	0.891 (.003)	1.000 (.000)	0.883 (.003)	**1.000** (.000)
MCR	0.882 (.003)	1.556 (.017)	0.840 (.006)	1.765 (.018)	0.854 (.005)	1.598 (.013)	0.872 (.003)	2.201 (.024)
G2-Residual	0.908 (.002)	1.703 (.018)	0.863 (.004)	1.650 (.014)	0.866 (.004)	1.458 (.011)	0.881 (.003)	1.870 (.020)
RBC-CHR	0.913 (.002)	1.798 (.032)	**0.900** (.003)	1.455 (.016)	0.895 (.003)	1.257 (.013)	**0.900** (.002)	1.849 (.031)
RLR-CHR	0.911 (.002)	1.578 (.016)	0.901 (.003)	1.459 (.014)	**0.903** (.003)	1.278 (.013)	0.905 (.003)	1.844 (.030)

Table 12: WSC on the additional datasets at slab masses $\delta = 0.05$ and 0.20, with $\alpha = 0.1$. WSC values closest to 0.9 are bold.

	WSC at $\delta = 0.05$							
Method	**Bike Sharing**	**Protein (CASP)**	**MEPS 20**	**MEPS 21**	**MEPS 19**	**Facebook 1**	**Facebook 2**	**BlogFeedback**
Split-CP	0.837 (.006)	0.883 (.005)	0.481 (.011)	0.407 (.009)	0.426 (.010)	0.487 (.009)	0.479 (.009)	0.541 (.011)
Norm-SC	0.874 (.005)	0.882 (.004)	0.854 (.006)	0.869 (.006)	0.862 (.005)	0.742 (.008)	0.759 (.008)	0.808 (.007)
CQR	0.842 (.006)	0.888 (.004)	0.884 (.005)	0.885 (.004)	0.884 (.004)	0.827 (.010)	0.851 (.008)	0.873 (.005)
CHR	0.835 (.007)	**0.894 (.004)**	0.888 (.004)	0.885 (.004)	0.888 (.004)	0.888 (.004)	**0.901 (.004)**	0.885 (.005)
MCR	0.888 (.004)	**0.894 (.004)**	0.886 (.004)	0.883 (.005)	0.879 (.004)	0.828 (.008)	0.846 (.005)	0.865 (.004)
G2-Residual	**0.902 (.003)**	**0.894 (.004)**	0.888 (.004)	**0.899 (.004)**	**0.901 (.004)**	0.863 (.005)	0.850 (.006)	0.879 (.005)
RBC-CHR	0.897 (.004)	0.893 (.004)	0.905 (.004)	0.913 (.003)	0.911 (.004)	0.895 (.004)	0.894 (.003)	0.897 (.004)
RLR-CHR	**0.902 (.004)**	**0.894 (.004)**	**0.904 (.004)**	0.910 (.004)	0.907 (.004)	**0.900 (.004)**	0.898 (.004)	**0.899 (.004)**

	WSC at $\delta = 0.20$							
Method	**Bike Sharing**	**Protein (CASP)**	**MEPS 20**	**MEPS 21**	**MEPS 19**	**Facebook 1**	**Facebook 2**	**BlogFeedback**
Split-CP	0.838 (.003)	0.881 (.002)	0.705 (.004)	0.696 (.004)	0.691 (.004)	0.687 (.004)	0.682 (.004)	0.733 (.003)
Norm-SC	0.888 (.002)	0.890 (.002)	0.871 (.003)	0.874 (.003)	0.875 (.003)	0.793 (.005)	0.806 (.005)	0.840 (.004)
CQR	0.871 (.003)	0.889 (.002)	0.883 (.002)	0.884 (.003)	0.881 (.002)	0.866 (.003)	0.872 (.003)	0.872 (.003)
CHR	0.861 (.003)	0.894 (.002)	0.890 (.002)	0.890 (.002)	**0.894 (.002)**	0.889 (.002)	0.894 (.002)	0.886 (.002)
MCR	**0.899 (.002)**	0.893 (.002)	0.893 (.002)	0.888 (.002)	0.889 (.002)	0.856 (.003)	0.867 (.003)	0.879 (.003)
G2-Residual	**0.901 (.002)**	0.893 (.002)	0.892 (.002)	**0.906 (.002)**	0.907 (.002)	0.874 (.003)	0.870 (.003)	0.885 (.002)
RBC-CHR	0.904 (.002)	**0.897 (.002)**	**0.901 (.002)**	0.912 (.002)	0.912 (.002)	0.902 (.002)	0.895 (.002)	**0.906 (.002)**
RLR-CHR	0.907 (.002)	0.895 (.002)	**0.901 (.002)**	0.910 (.002)	0.910 (.002)	**0.901 (.002)**	**0.899 (.002)**	**0.906 (.002)**

B.5 Coverage by covariate density

We examine whether RLR-CHR maintains coverage near the nominal level across regions of different covariate density, particularly where observations are sparse. We compare CHR and RLR-CHR across five strata of 400 test observations, ordered from D1 (sparsest) to D5 (densest). Strata are defined using each test observation's distance to the 20th nearest training observation in standardized covariate space. This density proxy uses the full covariates and differs from the coordinate Z based on the fitted scale. The networks are independently refitted on the same splits as Table 1.

Compared with CHR, RLR-CHR has coverage closer to 0.9 in the sparsest stratum on all eight datasets and generally less variation in coverage across density strata (Table 13). This reduction

in variation is especially clear on Video Time, where CHR's lowest coverage occurs in the densest stratum. Larger variation in coverage remains on Naval and MV, and RLR-CHR still undercovers in D2 on Parking and the three sparsest strata on Naval. The Video Time RLR-CHR cells exclude three flagged repetitions; including them gives D1–D5 coverage 0.902, 0.899, 0.923, 0.914, and 0.901.

Table 13: Coverage across strata defined by covariate density. D1 is sparsest and D5 densest.

Stratum	Method	SARCOS	SGEMM GPU	Tetouan	Parking	Naval	MV	Video Time	French Load
D1 (sparsest)	RLR-CHR	0.897 (.002)	0.896 (.002)	0.911 (.002)	0.903 (.002)	0.877 (.003)	0.879 (.002)	0.899 (.002)†	0.897 (.002)
	CHR	0.875 (.002)	0.890 (.002)	0.885 (.002)	0.929 (.002)	0.848 (.004)	0.866 (.002)	0.935 (.002)	0.881 (.002)
D2	RLR-CHR	0.909 (.001)	0.907 (.002)	0.905 (.002)	0.872 (.002)	0.865 (.003)	0.901 (.001)	0.896 (.002)†	0.896 (.001)
	CHR	0.899 (.001)	0.906 (.002)	0.893 (.002)	0.885 (.002)	0.886 (.003)	0.895 (.002)	0.906 (.002)	0.888 (.002)
D3	RLR-CHR	0.915 (.001)	0.906 (.001)	0.903 (.002)	0.914 (.002)	0.880 (.002)	0.912 (.002)	0.921 (.002)†	0.898 (.001)
	CHR	0.910 (.002)	0.905 (.001)	0.900 (.002)	0.902 (.002)	0.889 (.003)	0.911 (.001)	0.912 (.002)	0.896 (.002)
D4	RLR-CHR	0.912 (.001)	0.912 (.001)	0.907 (.001)	0.914 (.002)	0.945 (.002)	0.918 (.002)	0.911 (.002)†	0.906 (.002)
	CHR	0.902 (.001)	0.901 (.001)	0.913 (.001)	0.886 (.002)	0.924 (.002)	0.919 (.002)	0.893 (.002)	0.907 (.001)
D5 (densest)	RLR-CHR	0.895 (.002)	0.910 (.001)	0.911 (.002)	0.914 (.002)	0.967 (.001)	0.931 (.001)	0.898 (.002)†	0.918 (.001)
	CHR	0.916 (.001)	0.901 (.002)	0.915 (.001)	0.897 (.002)	0.962 (.002)	0.912 (.002)	0.855 (.003)	0.926 (.001)

† The RLR-CHR results for Video Time exclude 3 of the 100 repetitions; values computed over all repetitions are given in the text.

B.6 Score diagnostic by fitted scale

Assumption 1 requires residual CHR score distributions to vary little within localization neighborhoods, even when the standardized residual distribution varies with X. We examine how differences in residual shape across groups defined by the fitted scale are reflected in score distributions and coverage. On SGEMM, we group test observations by tertiles of $Z = \log \hat{\sigma}(X)$ and compare the empirical CDFs of standardized residuals and normalized residual CHR scores. The models are independently refitted on the main splits, and relative length is computed within each scale group.

Figure 2 shows that the residual CHR score CDFs are similar across scale groups despite differences in standardized residual distributions. Mean coverage remains close to 0.9, ranging from 0.906 to 0.908. Relative length rises from 0.641 in the lowest scale group to 0.771 in the highest, so the length advantage over CHR is smaller at higher fitted scales.

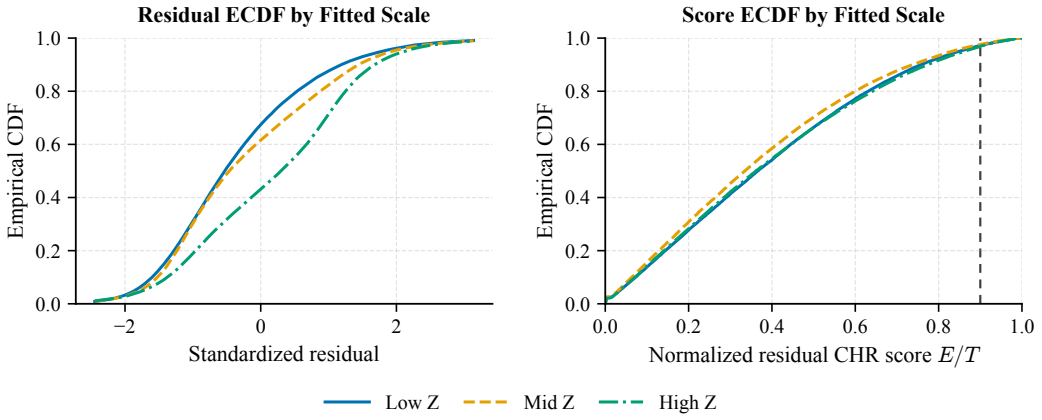


Figure 2: Pooled empirical CDFs of standardized residuals (left) and normalized residual CHR scores (right) within tertiles of the fitted scale on SGEMM. The vertical line marks $E/T = 0.9$.

B.7 Sensitivity to calibration choices

We examine how RLR-CHR's coverage and interval length change with the neighborhood size, score resolution, histogram bin count, and miscoverage level, and evaluate WSC at different slab masses.

B.7.1 Neighborhood size

For RLR-CHR on SGEMM, we vary the neighborhood size using $k = \mathrm{round}(ck_0)$, where k_0 is the unrounded balancing rule in Appendix B.1 and c ranges from 0.5 to 2. We use independent refits on the main splits, holding the fitted models and scores fixed as c varies within each repetition.

Figure 3 shows that RLR-CHR's performance is generally stable across neighborhood sizes, with a tradeoff between WSC and interval length. The smallest neighborhood ($c = 0.5$) gives higher WSC and longer intervals, while for $c \geq 0.75$ WSC remains near 0.9 and relative length changes little. Mean marginal coverage is 0.9183 at $c = 0.5$ and ranges from 0.9027 to 0.9075 for $c \geq 0.75$.

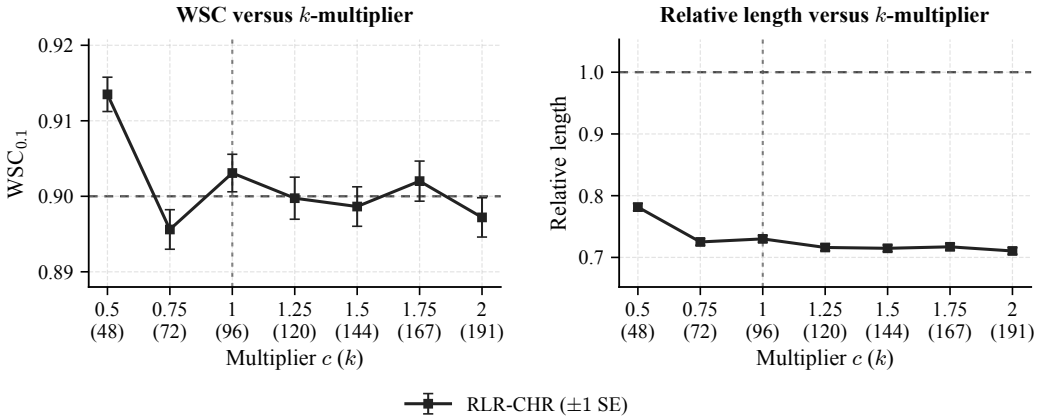


Figure 3: RLR-CHR WSC and relative length on SGEMM as the neighborhood multiplier varies. Error bars show ± 1 SE; the vertical line marks the default $c = 1$.

B.7.2 Score levels and histogram bins

We use independent refits to examine sensitivity to the number of score levels on SGEMM and to the number of histogram bins on BlogFeedback.

On SGEMM, we vary T from 25 to 1,000 while keeping the fitted models and the number of histogram bins fixed within each repetition. The default neighborhood rule changes k with T, giving $k = 65, 72, 78, 84, 96$ at the five settings in Table 14. WSC remains near 0.9, ranging from 0.8960 to 0.9036, while mean interval length changes little, ranging from 47.24 to 49.08.

On BlogFeedback, we hold T and the fitted quantile networks fixed within each repetition and increase the number of bins from 100 to 1,000. This reduces mean interval length for both CHR and RLR-CHR, with a larger reduction for CHR, while WSC changes little (Table 15).

The RLR-CHR means exclude three flagged repetitions at 100 bins and one at 1,000 bins. With all repetitions included, the 100-bin row has coverage 0.9152, WSC 0.9104, and length 120.56 (SE 88.75); the 1,000-bin row has coverage 0.9099, WSC 0.9033, and length 25.96 (SE 0.66).

Table 14: RLR-CHR sensitivity to T on SGEMM. The default neighborhood size varies with T.

T	Coverage	$\mathrm{WSC}_{0.1}$	Average length
25	0.9110 (0.0007)	0.9003 (0.0028)	49.08 (0.34)
50	0.9104 (0.0007)	0.9036 (0.0028)	48.79 (0.34)
100	0.9075 (0.0007)	0.8971 (0.0026)	48.06 (0.34)
200	0.9032 (0.0007)	0.8967 (0.0027)	47.24 (0.34)
1,000	0.9057 (0.0007)	0.8960 (0.0028)	47.62 (0.34)

B.7.3 Miscoverage level and slab mass

We assess sensitivity to the nominal coverage level on SGEMM and Bike using $\alpha = 0.05$ and 0.20, and evaluate WSC at $\delta \in \{0.05, 0.10, 0.20\}$. The comparison includes CHR and RLR-CHR at $\alpha = 0.05$ (Table 16) and all eight methods at $\alpha = 0.20$ (Table 17).

Table 15: Sensitivity to the number of histogram bins on BlogFeedback.

Method	Bins	Coverage	$\text{WSC}_{0.1}$	Average length
CQR	—	0.8989 (0.0008)	0.8711 (0.0034)	17.13 (0.22)
CHR	100	0.9005 (0.0007)	0.8843 (0.0036)	31.44 (0.25)
CHR	1,000	0.9001 (0.0008)	0.8840 (0.0028)	14.11 (0.14)
RLR-CHR	100	0.9130 (0.0013)†	0.9079 (0.0025)†	28.28 (0.85)†
RLR-CHR	1,000	0.9094 (0.0008)†	0.9027 (0.0028)†	25.37 (0.29)†

† The RLR-CHR results exclude 3 of the 100 repetitions at 100 bins and 1 at 1,000 bins; summaries over all repetitions are in the text.

Across these slab masses, RLR-CHR's WSC is close to the nominal level at both values of α. Its relative length shows the same pattern in both studies: intervals are shorter than CHR's on SGEMM and longer on Bike. At $\alpha = 0.20$, RBC-CHR and G2-Residual have WSC similar to RLR-CHR, while the two residual CHR methods give shorter intervals than G2-Residual on SGEMM.

Table 16: CHR and RLR-CHR at $\alpha = 0.05$. WSC is reported at $\delta = 0.05, 0.10, 0.20$.

		CHR WSC			RLR-CHR WSC			RLR-CHR
α	Dataset	$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.20$	$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.20$	rel. length
0.05	SGEMM GPU	0.943 (.003)	0.948 (.002)	0.947 (.002)	0.957 (.002)	0.955 (.002)	0.956 (.001)	0.798 (.012)
0.05	Bike Sharing	0.907 (.005)	0.912 (.004)	0.927 (.002)	0.955 (.003)	0.957 (.002)	0.955 (.001)	1.385 (.014)

Table 17: Comparison of the eight methods at $\alpha = 0.20$. WSC is reported at $\delta = 0.05, 0.10, 0.20$; WSC values closest to 0.8 at each slab mass and the smallest relative lengths in each dataset are bold.

SGEMM GPU	WSC				
Method	$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.20$	Rel. length	Coverage
Split-CP	0.624 (.009)	0.635 (.007)	0.659 (.005)	1.137 (.015)	0.800 (.001)
Norm-SC	0.790 (.005)	0.792 (.003)	0.795 (.003)	1.102 (.014)	0.802 (.001)
CQR	0.775 (.006)	0.773 (.005)	0.767 (.004)	1.216 (.011)	0.801 (.001)
CHR	0.783 (.005)	0.778 (.004)	0.784 (.003)	1.000 (.000)	0.801 (.001)
MCR	0.789 (.005)	0.789 (.004)	0.793 (.003)	1.209 (.015)	0.801 (.001)
G2-Residual	0.804 (.005)	**0.797 (.004)**	**0.801 (.002)**	1.092 (.009)	0.804 (.001)
RBC-CHR	0.799 (.005)	0.794 (.004)	**0.799 (.002)**	**0.696 (.007)**	0.805 (.001)
RLR-CHR	**0.800 (.005)**	0.796 (.003)	**0.801 (.002)**	0.698 (.009)	0.805 (.001)
Bike Sharing	WSC				
Method	$\delta = 0.05$	$\delta = 0.10$	$\delta = 0.20$	Rel. length	Coverage
Split-CP	0.678 (.008)	0.677 (.006)	0.686 (.004)	1.116 (.008)	0.799 (.001)
Norm-SC	0.765 (.005)	0.781 (.004)	0.786 (.003)	1.051 (.007)	0.799 (.001)
CQR	0.734 (.007)	0.752 (.006)	0.761 (.005)	1.094 (.007)	0.801 (.001)
CHR	0.734 (.008)	0.745 (.005)	0.753 (.004)	**1.000 (.000)**	0.801 (.001)
MCR	0.779 (.006)	0.791 (.004)	0.793 (.003)	1.122 (.008)	0.803 (.001)
G2-Residual	0.786 (.006)	0.794 (.004)	0.794 (.003)	1.127 (.007)	0.810 (.002)
RBC-CHR	0.796 (.005)	0.793 (.004)	0.794 (.003)	1.103 (.007)	0.808 (.001)
RLR-CHR	**0.799 (.005)**	**0.797 (.004)**	**0.798 (.003)**	1.111 (.007)	0.810 (.002)

B.8 Experiment with varying skewness

We assess RLR-CHR when the standardized error distribution varies with the covariate, so the location-scale assumption (A2) does not hold. We retain the location and scale functions of Section 4.1 and replace the error by a centered skew-normal random variable with unit variance and shape parameter $+4$ for $x < 0.5$ and -4 otherwise.

As shown in Table 18, RLR-CHR retains WSC near the nominal level and its length advantage over G2-Residual and MCR. CHR produces shorter intervals but still exhibits overcoverage for $x < 0.5$ and undercoverage for $x \geq 0.5$.

B.9 RLCP baseline

We implement the randomized localization method of Hore and Barber [2025] using the absolute residual score $|Y - \hat{f}(X)|$, where $\hat{f}$ is fitted on the full training split. The comparison uses the

Table 18: Design with varying skewness at $n = 3{,}000$, split equally into training and calibration. WSC values closest to 0.9 and the smallest mean interval lengths are bold.

Method	$\text{WSC}_{0.1}$	Average length	Coverage, $x < 0.5$	Coverage, $x \geq 0.5$
CHR	0.7079 (0.0087)	**1.668 (0.006)**	0.9440	0.8586
Norm-SC	0.7358 (0.0127)	2.594 (0.022)	0.8685	0.9296
G2-Residual	0.8813 (0.0049)	2.431 (0.012)	0.9105	0.9041
MCR	0.8791 (0.0059)	2.454 (0.011)	0.9169	0.9003
RLR-CHR	**0.8928 (0.0040)**	2.041 (0.015)	0.9023	0.9167

same point network architecture, data splits, and seeds as the main experiments, with independently refitted networks. Localization uses a Gaussian kernel in the full covariate space, with covariates standardized using training means and standard deviations.

We select the bandwidth using the authors' effective-size criterion with target 50. The criterion is evaluated on a random training subsample of the same size as the calibration split, using 500 randomly selected centers with Gaussian perturbations and normalized weights computed over the entire subsample. We search bandwidths $0.02, 0.04, \ldots, 6$ in increasing order and return the first that reaches the target, or 6 if none does.

RLCP produces unbounded intervals on 0.9–2.0% of test points across these datasets. Its mean length is therefore not finite. Following the authors' convention, we summarize RLCP by the median interval width within each repetition. In Table 1, this median width is divided by CHR's mean length from the main experiment in the same repetition.

B.10 Computational cost

RBC-CHR applies the G2 rule of Guan [2019, Theorem 3.2 and Algorithm 1]. The quantile in the test neighborhood and the two G2 acceptance fractions are nondecreasing in the adjusted probability level, so the implementation uses bisection to obtain the same returned threshold as the scan in Algorithm 2.

Experiments ran on Apple M4 Max (16 CPU cores, 64 GB memory) and M4 Pro (12 CPU cores, 48 GB memory) machines, using up to ten CPU workers per run and one thread per worker. Table 19 reports approximately 1,450 hours summed across CPU workers.

Table 19: Approximate compute for the reported experiments, measured as elapsed time summed across CPU workers.

Study	CPU	Compute (h)	Study	CPU	Compute (h)
Synthetic grid and skewness	Both	149	Covariate density	Both	352
Main comparison	M4 Max	395	SGEMM diagnostics	M4 Max	28
RLCP comparison	M4 Pro	229	Calibration sensitivity	Both	90
Additional datasets	Both	204	WSC estimator calibration	M4 Max	0.05

Guan [2019, Corollary 3.1] introduced the G1 rule and noted that its direct computation was too expensive for practical use. This motivated the practical G2 construction in that paper's Theorem 3.2 and Algorithm 1, which RBC-CHR uses. Guan [2023, Section 3.2] later gave an $O(m \log m)$ algorithm for the closure of the G1 score acceptance set, given precomputed localizer values and cumulative probabilities.

A G1 version of RBC-CHR would use the same fitted models, residual CHR scores, and localizer as RLR-CHR. After preprocessing the calibration data, both require $O(m \log m)$ per test point for neighborhood construction and calibration when T is fixed. For the one-dimensional k-nearest-neighbor localizer used here, RLR-CHR and the G2 version of RBC-CHR also admit implementations with the same complexity bound per test point. We reexamined these computational relationships while preparing the camera-ready version.

B.11 Worked example of local-rank calibration

The following example illustrates how RLR-CHR constructs local ranks and uses them to select the calibrated score threshold. Consider $m = 5$ calibration points, $T = 4$ levels, $k = 2$, and $\alpha = 0.4$.

Suppose the localization coordinates and residual CHR scores are

$$(\widetilde{Z}_1, \ldots, \widetilde{Z}_5) = (0.10, 0.30, 0.50, 0.85, 1.40), \qquad (E_1, \ldots, E_5) = (2, 1, 3, 2, 4).$$

For the first point, take $\widetilde{R}_1 \notin \widetilde{S}_1(X_1)$ and $\widetilde{R}_1 \in \widetilde{S}_2(X_1)$; then $E_1 = \min\{t : \widetilde{R}_1 \in \widetilde{S}_t(X_1)\} = 2$. The other scores are obtained in the same way. The test point has $\widetilde{Z}_6 = 0.60$. The self-including neighborhoods are $\mathcal{N}_1 = \{1, 2, 3\}$, $\mathcal{N}_2 = \{2, 1, 3\}$, $\mathcal{N}_3 = \{3, 6, 2\}$, $\mathcal{N}_4 = \{4, 6, 3\}$, $\mathcal{N}_5 = \{5, 4, 6\}$, and $\mathcal{N}_6 = \{6, 3, 4\}$. Only calibration ranks A_3, A_4, A_5 depend on the candidate test score. Table 20 gives the two levels visited by the downward scan.

Table 20: The two levels visited by the downward scan in the worked example.

e	A_1	A_2	A_3	A_4	A_5	A_6	$p(e)$
4	2	1	2	1	3	3	$2/6$
3	2	1	3	1	3	3	$3/6$

At $e = 4$, only A_5 reaches the test rank $A_6 = 3$, giving $p(4) = 2/6 \leq 0.4$. At $e = 3$, both A_3 and A_5 reach that rank, so $p(3) = 3/6 > 0.4$ and the scan returns $e^\star = 3$. The prediction interval is $\hat{\mu}(X_6) + \hat{\sigma}(X_6)\tilde{S}_3(X_6)$. At the remaining levels, $p(2) = 5/6$ and $p(1) = 1$. Figure 4 shows the test neighborhood and the full acceptance path. The final test compares all local ranks, including those from neighborhoods distant from the test point; locality enters through the construction of each rank.

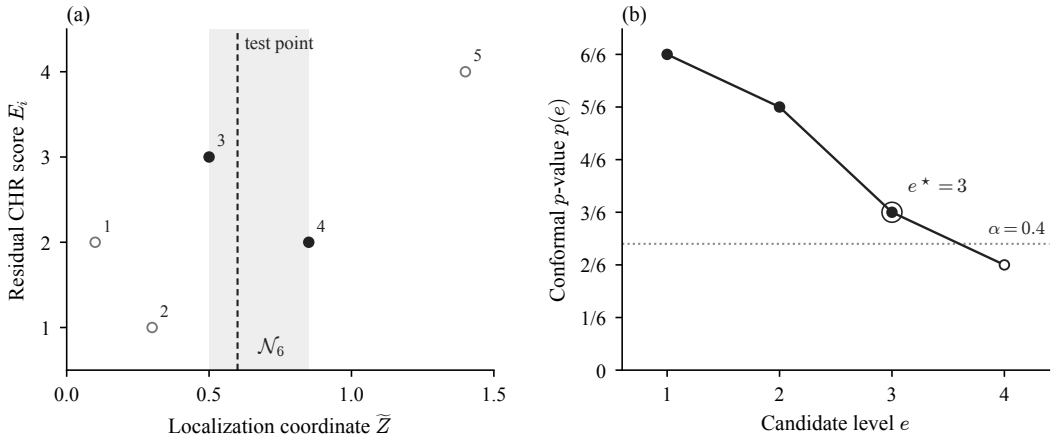


Figure 4: Worked example. Left: calibration and test points in the localization coordinate, with the test neighborhood indicated. Right: the acceptance path $p(e)$, the level $\alpha = 0.4$, and the selected score $e^\star = 3$.

B.12 Guarantees and further related work

The methods in Table 21 have finite-sample marginal validity under their respective exchangeability and construction conditions. Their conditional or local guarantees differ in target and assumptions.

Other approaches adapt quantile estimation, conformity scores, or the calibration target. Orthogonal quantile regression and uncertainty-aware CQR modify quantile estimation [Feldman et al., 2021, Rossellini et al., 2024], while boosted conformity scores directly target conditional coverage or length [Xie et al., 2024]. Learned partitions and group-based calibration provide alternatives to a scalar scale coordinate [Cabezas et al., 2025, Jung et al., 2023, Gibbs et al., 2025]; self-calibrating conformal prediction conditions on a calibrated model output [van der Laan and Alaa, 2024]. These targets differ from conditioning on the full covariate at a realized test point.

Classical conformal density regions [Lei et al., 2013] and recent methods for volume minimization [Le Bars and Humbert, 2025] also clarify why efficiency depends on the permitted set class. A disconnected region may have smaller volume than the shortest single interval.

Concurrent work gives finite-sample bounds for localized conformal prediction [Conrad et al., 2026] and a general decomposition of conditional miscoverage [Min et al., 2026]. Our bound on conditional

Table 21: Comparison of coverage and length guarantees. Asymptotic conditional coverage is subject to the stated consistency or stability assumptions; local guarantees use the averaging or partition target specified by the method.

Method	Marginal	Conditional or local	Length or set efficiency
CQR [Romano et al., 2019]	Finite-sample	—	—
CHR [Sesia and Romano, 2021]	Finite-sample	Asymptotic conditional under histogram consistency	Asymptotic oracle interval length
Local bands [Lei and Wasserman, 2014]	Finite-sample	Local; asymptotic conditional under regularity	Asymptotic oracle efficiency
Guan LCP [Guan, 2019, 2023]	Finite-sample	Localized coverage in a bandwidth limit	—
RLCP [Hore and Barber, 2025]	Finite-sample	Randomized neighborhood-averaged guarantees	—
CD/HPD-split [Izbicki et al., 2022]	Finite-sample	Local and asymptotic conditional	Oracle efficiency for density regions
RBC-CHR	Finite-sample	Asymptotic conditional under Assumption 1	Oracle interval length under (E1)–(E4), (C1), $T \to \infty$ and localization conditions
RLR-CHR	Finite-sample	Theorem 2: local score stability	Theorem 3: oracle interval length; Theorem 7: equivalence to RBC-CHR

coverage makes the terms for score atoms and overlapping neighborhoods explicit for the local-rank rule, and the equivalence result concerns its weighted counterpart on the same residual scores.

B.13 Worst-slab coverage estimation and calibration study

Our WSC estimator uses 25% of the test points to search over slabs along random unit directions and the remaining 75% to evaluate the selected slab. Among slabs containing at least a fraction δ of the search subsample, the search selects the one with the lowest empirical coverage. We use 100 directions for the synthetic experiments and 1,000 for the real data experiments.

We use this split to avoid the selection bias that can arise when a slab is selected and evaluated on the same sample. Even when coverage is nominal in every slab, some empirical coverage estimates can be low because of sampling variability. Selecting and reporting the lowest estimate on the same sample can therefore produce a downward bias.

To check the effect of the split, we simulate independent Bernoulli(0.9) coverage indicators, so that every slab has true coverage 0.9. We compare our 25/75 split estimator with an alternative that selects and evaluates the slab on the same sample. The two Gaussian designs have dimensions $p = 1$ and $p = 10$, and test sizes 1,000 and 2,000, respectively.

Table 22: WSC estimator calibration under independent Bernoulli(0.9) coverage, with $\delta = 0.1$. The Gaussian rows use $X \sim N(0, I_p)$, 200 simulations, and 100 random unit directions; the toy example with disjoint slabs uses 20,000 simulations.

Design	Estimator	Mean	5% quantile	95% quantile
Ten disjoint slabs, 100 points each	Minimum	0.8516	0.8200	0.8800
$p = 1$, $n_{\text{test}} = 1{,}000$	Same-sample	0.8278	0.7905	0.8558
$p = 1$, $n_{\text{test}} = 1{,}000$	Split 25/75	0.9008	0.8461	0.9535
$p = 10$, $n_{\text{test}} = 2{,}000$	Same-sample	0.8152	0.7960	0.8326
$p = 10$, $n_{\text{test}} = 2{,}000$	Split 25/75	0.9029	0.8665	0.9419

The calibration study in Table 22 shows a substantial downward bias when a slab is selected and evaluated on the same sample, whereas independent evaluation is centered near the nominal coverage of 0.9. A persistent deficit in split-sample WSC therefore reflects local undercoverage. For example, CHR's synthetic WSC is 0.741 (0.007) at $n = 8{,}000$, well below nominal.

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 main claims in the abstract and Introduction are supported by the methodology, theoretical results, and experiments in Sections 2–4.

   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: Sections 4 and 5 discuss the limitations of the localizer, the length cost on additional datasets, and failures caused by score ties. Appendix B provides further diagnostics, sensitivity analyses, and computational details.

   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: The assumptions are stated or referenced with each theoretical result. Appendix A provides complete proofs, with the regularity conditions collected in Appendix A.4.

Guidelines:

- 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: Algorithm 1, Section 4, and Appendix B describe the method, data generation and preprocessing, model fitting, calibration, and evaluation procedures used in the main experiments.

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: Code and instructions for reproducing the main experiments are available at `https://github.com/Xiangshi-L/rlr-chr`. Data sources and preprocessing details are provided in Appendix B.1.

   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: Section 4 and Appendices B.1 and B.9 describe the data splits, model architectures, training settings, and choices of calibration parameters.

   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 are reported with standard errors across repetitions. Figures 1 and 3 show ± 1 SE, and Appendix B explains the calculation and sources of variability.

   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 B.10 reports hardware, memory, worker counts, and compute estimates for the reported experiments, including an overall total. Preliminary and other unreported experiments required additional compute.

   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 research follows the NeurIPS Code of Ethics. The experiments use simulated and public benchmark data, with no participant recruitment (Section 4 and Appendix B.1).

   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: The paper develops a conformal prediction method and does not study a deployed application or a use case with direct societal impact.

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 work introduces no trained model or dataset with high misuse risk.

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: Section 4 and Appendix B.1 cite the data sources and existing methods. Appendix B.1 describes the available licenses and terms of use.

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 code package linked in Section 4 includes documentation of the methods, software environment, data preparation, and reproduction commands, together with an MIT 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: No participants were recruited or contacted, and there was no crowdsourcing; MEPS is used only for secondary analysis of public de-identified data.

    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 study uses simulated data and public benchmark datasets, including de-identified MEPS public use files. It involves no participant recruitment or linkage to identifiable records.

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 used for manuscript editing and formatting, literature searches, and code generation. They are not a component of the proposed statistical method.

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.