CRPS-Optimal Binning for Univariate Conformal Regression

Paolo Toccaceli

Centre for Reliable Machine Learning, Royal Holloway, University of London

Abstract

We propose a method for non-parametric conditional distribution estimation based on partitioning covariate-sorted observations into contiguous bins and using the within-bin empirical CDF as the predictive distribution. Bin boundaries are chosen to minimise the total leave-one-out Continuous Ranked Probability Score (LOO-CRPS), which admits a closed-form cost function with $O(n^2 \log n)$ precomputation and $O(n^2)$ storage; the globally optimal K-partition is recovered by a dynamic programme in $O(n^2 K)$ time. Minimisation of within-sample LOO-CRPS turns out to be inappropriate for selecting K as it results in in-sample optimism. We instead select K by K-fold cross-validation of test CRPS, which yields a U-shaped criterion with a well-defined minimum. Having selected K^* and fitted the full-data partition, we form two complementary predictive objects: the Venn prediction band and a conformal prediction set based on CRPS as the nonconformity score, which carries a finite-sample marginal coverage guarantee at any prescribed level ε. On real benchmarks against split-conformal competitors (Gaussian split conformal, CQR, CQR-QRF, and conformalized isotonic distributional regression), the method produces substantially narrower prediction intervals while maintaining near-nominal coverage.

Keywords: conformal prediction, conformal regression, distribution-free, non-parametric regression, optimal binning

1 Introduction

A fundamental problem in supervised learning is to estimate not just the conditional mean $\mathbb{E}[Y \mid X = x]$ but the full conditional distribution $P(Y \mid X = x)$. A distributional forecast communicates uncertainty in a way that a point prediction cannot: it is necessary for decision-making under risk, hypothesis testing at a test point, and the construction of valid prediction sets.

The simplest non-parametric approach is to collect nearby training observations and use their empirical CDF as a stand-in for $P(Y \mid X = x)$. Nearest-neighbour and kernel density methods implement this idea but require tuning a bandwidth that trades off local fidelity against statistical efficiency. When the covariate is one-dimensional and the data are sorted by x, a natural alternative is to partition the sorted sequence into contiguous bins and predict with the within-bin ECDF. Binning is interpretable, fast, and directly connected to the Venn predictor framework of Vovk, Gammerman, and Shafer [12]: the within-bin ECDF is the reference class predictor for a test point that falls in a given bin.

The central question is how to place bin boundaries optimally. Informal choices, such as equal-width or equal-count bins, are oblivious to the heteroscedasticity of the response and

to the local density of the covariate. We argue that bin boundaries should minimise a proper scoring rule evaluated on the training data, so that the binning criterion and the predictive goal are aligned. Among proper scoring rules, the CRPS is particularly well suited to this purpose: it targets the full distribution (not a single functional), it admits a closed-form leave-one-out formula that depends only on pairwise absolute differences, and it is the same score used in the conformal prediction step, ensuring coherence between bin selection and prediction-set construction.

Contributions.

1. **Closed-form LOO-CRPS cost.** We derive the total leave-one-out CRPS of a bin of size m as $\text{cost}(S) = mW/(m-1)^2$, where $W = \sum_{\ell<r} |y_\ell - y_r|$ is the pairwise dispersion (Proposition 1). This scalar is updated in $O(\log n)$ per extension step, giving $O(n^2 \log n)$ precomputation and $O(n^2)$ storage for all subintervals.

2. **Exact optimal partitioning via dynamic programming.** The additive cost structure satisfies the optimal substructure property, so a DP recovers the globally optimal K-partition in $O(n^2K)$ time (Section 5). Unlike greedy binary segmentation [19], which fixes cut points sequentially and can miss the global optimum, the DP evaluates all continuations simultaneously and is exact — and vastly more efficient than exhaustive search, which would enumerate all $\binom{n-1}{K-1}$ candidate K-partitions.

3. **Cross-validated K selection.** We identify in-sample optimism as the reason within-sample LOO-CRPS fails as a model selection criterion and propose a K-fold cross-validated test-CRPS criterion that yields a U-shaped curve and a sensible K^* (Section 7).

4. **Conformal prediction sets and Venn connection.** We construct CRPS-based conformal prediction sets with finite-sample marginal coverage guarantees. In the split-conformal case, convexity of the score guarantees that prediction sets are always connected intervals; in our transductive setting, single-interval structure is observed empirically in all experiments. We also formalise the regression analog of the Venn Predictor, namely a constant-width family of augmented ECDFs, as the theoretical underpinning that motivates the LOO scoring construction (Section 8).

A distinguishing feature of the method is that it is *transductive*: all n observations are used for both partitioning and conformal calibration, with no data held out. Split-conformal competitors must reserve a calibration set, effectively halving the sample available for model fitting. This data-efficiency advantage is particularly pronounced in small-sample settings.

Organisation. Section 2 surveys related work. Section 3 fixes notation and defines the binning problem. Section 4 derives the closed-form LOO-CRPS cost. Section 5 presents the DP recurrence. Section 6 describes precomputation and complexity. Section 7 analyses K selection, diagnosing the failure of within-sample LOO-CRPS, and proposes the cross-validated criterion. Section 8 covers the Venn band, conformal prediction, and the approximate-exchangeability trade-off that governs the choice of K. Section 9 gives the synthetic numerical illustration. Section 10 presents real-data experiments on Old Faithful and

the motorcycle benchmark. Section 11 discusses the role of CRPS. Section 12 concludes and identifies open problems.

2 Related Work

Conformal prediction for regression. Conformal prediction [12] wraps any nonconformity score in a distribution-free prediction set with finite-sample marginal coverage; Lei et al. [24] provide a comprehensive treatment for the regression setting, covering a range of nonconformity scores and establishing finite-sample guarantees. The inductive (split-conformal) variant [13] divides training data into a fitting half and a calibration half; the quantile of calibration nonconformity scores determines the prediction set for a new point. Conformalized Quantile Regression (CQR) [14] improves efficiency in heteroscedastic settings by using conditional quantile estimates as the base predictor and measuring residuals relative to the estimated interval. Mondrian conformal prediction [12] stratifies the calibration set into categories (taxonomies) so that coverage holds conditionally within each stratum; our method is a Mondrian predictor with data-adaptive bins as strata and CRPS as the nonconformity score. Sesia and Romano [25] propose conformal prediction based on conditional histograms, stratifying calibration scores by the value of a one-dimensional predictor (such as an estimated quantile); our method instead partitions on the covariate x directly and selects bin boundaries to minimise LOO-CRPS rather than conditioning on a downstream predictor score.

Venn predictors. Venn predictors [12] output a set of probability distributions, one for each candidate label of the test point. Vovk and Petej [15] develop Venn-ABERS predictors for binary classification: isotonic regression on an underlying scoring function implicitly identifies an optimal partition of the score range into reference classes, each with finite-sample calibration guarantees. Our method mirrors this structure on the covariate axis: the DP selects CRPS-optimal bins, defining reference classes whose within-bin ECDF best describes the local conditional distribution. Our Venn prediction band (Section 8) then follows: for each hypothesised response value y^*, the within-bin ECDF is augmented with y^*, yielding a family of distributions parametrised by the test label.

Conformal predictive systems and distributional conformal prediction. Allen et al. [26] establish a unifying framework: any distributional regression procedure that is in-sample calibrated, when conformalized, yields a conformal predictive system with out-of-sample calibration guarantees under exchangeability. Their framework covers conformal binning, that is, partitioning calibration data into covariate-similar groups and predicting with within-group ECDFs, and conformal isotonic distributional regression (IDR) as canonical instances, but leaves the bin-selection criterion unspecified. The present paper fills this gap: we identify CRPS-optimal contiguous bins via dynamic programming and provide a closed-form cost formula, a CV criterion for K, and structural results on the resulting prediction set. Since the within-bin ECDF is in-sample auto-calibrated by construction, our conformal binning step is an instance of their framework and inherits the corresponding calibration guarantee. Chernozhukov et al. [27] construct (approximately) conditionally valid prediction intervals by permuting PIT residuals; their approach targets conditional validity

via a rank-based argument rather than CRPS-optimal predictive distributions. Randahl et al. [28] enforce coverage within pre-specified *outcome* bins to promote coverage equity across label subgroups; their partition is on the response y rather than on the covariate x, and the objective is coverage equity rather than predictive sharpness.

Conditional distribution estimation. Quantile Regression Forests [16] estimate the full conditional CDF by aggregating leaf-node empirical CDFs from an ensemble. NGBoost [17] fits parametric conditional distributions via natural-gradient boosting with proper scoring rule objectives. Parametric distributional regression, from classical quantile regression [18] to modern deep architectures, can be efficient when the distributional family is correctly specified but degrades under misspecification. Our method is model-free: it requires only observations sorted by covariate, makes no distributional assumption, and adapts automatically to heteroscedasticity and multimodality.

Optimal partitioning and change-point detection. The DP recurrence of Section 5 is structurally identical to exact segment neighbourhood algorithms [19, 20] that minimise an additive segmented cost over all K-partitions. PELT [20] achieves linear expected complexity via a pruning rule; the $O(n^2K)$ cost of our algorithm reflects the absence of a comparable pruning condition for the LOO-CRPS cost function. Greedy binary segmentation offers an $O(n \log n)$ approximation but is susceptible to local optima (Section 5). The key distinction from change-point detection is objective: classical methods seek structurally homogeneous segments (constant mean, variance, or full distribution), whereas we minimise predictive LOO-CRPS to optimise the calibration of the within-bin ECDF as a forecasting distribution, irrespective of within-segment homogeneity.

3 Setup

Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$ be a training sample. Sort observations by covariate value so that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, and write y_i for the response paired with $x_{(i)}$.

A *K-partition* of the sorted observations is a sequence of indices $0 = b_0 < b_1 < \cdots < b_K = n$ defining K contiguous bins $B_k = \{b_{k-1} + 1, \ldots, b_k\}$ for $k = 1, \ldots, K$. Within bin B_k, the predictive distribution for a new $x \in [x_{(b_{k-1}+1)}, x_{(b_k)}]$ is taken to be the empirical CDF of $\{y_i : i \in B_k\}$.

4 LOO-CRPS Cost of a Bin

The CRPS: definition and geometric interpretation

For a predictive CDF F and a scalar outcome $y \in \mathbb{R}$, the *Continuous Ranked Probability Score* [1, 2, 3, 4] is the integrated squared difference between F and the CDF of a point mass at y:

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} \big(F(t) - \mathbf{1}[t \geq y]\big)^2 \, \mathrm{d}t. \tag{1}$$

Here $\mathbf{1}[t \geq y]$ is the right-continuous CDF of a Dirac mass at y; some references write $\mathbf{1}[t < y]$ (the left-continuous version), which agrees everywhere except at the single point

$t = y$ and leaves the integral unchanged. Geometrically, the integrand is the squared vertical gap between F and the step at y (Figure 1). CRPS is zero if and only if F is a point mass at y; diffuse or mis-centred forecasts accumulate a larger integrated gap and hence a larger score. An equivalent energy-score form [10, Eq. (21)], convenient for computation, is

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F|X - y| - \tfrac{1}{2}\,\mathbb{E}_F|X - X'|, \tag{2}$$

where X, X' are independent draws from F. The first term penalises mis-location; the second subtracts a self-dispersion penalty that rewards sharpness, preventing the score from being minimised by a diffuse prior. CRPS is a *strictly proper* scoring rule for the class of all distributions with finite first moment [10]: its expected value under P is uniquely minimised by the forecast $\hat{F} = P$.

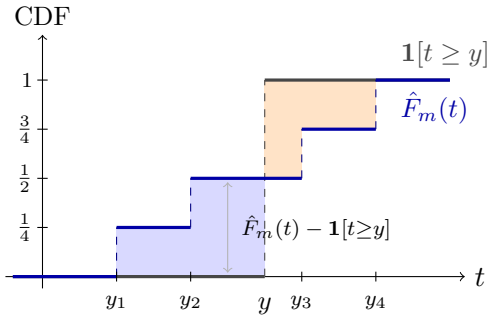

Figure 1: Geometric interpretation of $\mathrm{CRPS}(\hat{F}_m, y)$ as the integral of the squared vertical gap between the predictive CDF $\hat{F}_m$ (blue step function, $m = 4$ atoms) and the step $\mathbf{1}[t \geq y]$ (grey) at the observed outcome y. Blue shading marks intervals where $\hat{F}_m(t) > \mathbf{1}[t \geq y]$ (too little forecast mass above t); orange marks intervals where $\hat{F}_m(t) < \mathbf{1}[t \geq y]$ (too little mass below t). $\mathrm{CRPS} = \int (\hat{F}_m(t) - \mathbf{1}[t \geq y])^2\,\mathrm{d}t$ is large when the forecast is mis-centred or over-dispersed relative to y.

Empirical CDF and the LOO cost

For a predictive CDF $\hat{F}$ consisting of m equally weighted atoms and outcome y, applying (2) gives

$$\mathrm{CRPS}(\hat{F}, y) = \frac{1}{m}\sum_{i=1}^{m}|y_i - y| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|y_i - y_j|.$$

Let S be a bin of size m with response values $y_1, \ldots, y_m$. Write

$$d_k = \sum_{\ell \neq k}|y_\ell - y_k|, \qquad D = \sum_{\ell \neq r}|y_\ell - y_r| = 2\sum_{\ell < r}|y_\ell - y_r|.$$

The leave-one-out predictive distribution for observation k is the ECDF of $\{y_\ell : \ell \in S,\ \ell \neq k\}$. Applying the CRPS formula with $m - 1$ atoms:

$$\mathrm{CRPS}(\hat{F}_{S\setminus\{k\}},\, y_k) = \frac{d_k}{m-1} - \frac{D - 2d_k}{2(m-1)^2}.$$

Proposition 1. *The total leave-one-out CRPS of bin S is*

$$\text{cost}(S) \;=\; \sum_{k\in S} \text{CRPS}(\hat{F}_{S\setminus\{k\}},\, y_k) \;=\; \frac{m}{2(m-1)^2}\, D \;=\; \frac{m}{(m-1)^2} \sum_{\ell<r,\; \ell,r\in S} |y_\ell - y_r|.$$

Proof. Sum the per-observation expression over $k \in S$. For the first term,

$$\sum_{k=1}^{m} \frac{d_k}{m-1} = \frac{1}{m-1} \sum_{k=1}^{m} d_k = \frac{D}{m-1},$$

since $\sum_k d_k = \sum_k \sum_{\ell\neq k} |y_\ell - y_k| = D$. For the second term, note that removing k from the pairwise sum gives $\sum_{\ell\neq r,\, \ell,r\neq k} |y_\ell - y_r| = D - 2d_k$, so

$$\sum_{k=1}^{m} \frac{D-2d_k}{2(m-1)^2} = \frac{mD-2D}{2(m-1)^2} = \frac{D(m-2)}{2(m-1)^2}.$$

Subtracting and simplifying the numerator,

$$\text{cost}(S) = \frac{D}{m-1} - \frac{D(m-2)}{2(m-1)^2} = \frac{D\big[2(m-1)-(m-2)\big]}{2(m-1)^2} = \frac{mD}{2(m-1)^2}.$$

Since $D = 2\sum_{\ell<r} |y_\ell - y_r| = 2W$, this equals $mW/(m-1)^2$. □

For a bin spanning sorted indices i through j (with $m = j - i + 1$), write

$$W(i,j) = \sum_{\substack{\ell<r \\ \ell,r\in\{i,\ldots,j\}}} |y_\ell - y_r|, \qquad c(i,j) = \frac{j-i+1}{(j-i)^2}\, W(i,j).$$

5 Dynamic Programme

Define $\text{dp}[k][j]$ as the minimum total LOO-CRPS achievable by partitioning observations $1,\ldots,j$ into exactly k contiguous bins.

Minimum bin size. The LOO cost $c(i,j)$ requires at least two observations: for a singleton bin the leave-one-out distribution is empty, so the LOO-CRPS is undefined. We set $c(i,i) = +\infty$ by convention, which causes the DP to exclude singleton bins automatically. The index ranges below implicitly assume $j - i \geq 1$ (i.e. $m \geq 2$).

Base case.

$$\text{dp}[1][j] = c(1,j), \qquad j = 2,\ldots,n.$$

Recurrence. If $\text{dp}[k-1][i]$ gives the cost of the optimal partition of the first i observations into $k-1$ bins, then to partition $1,\ldots,j$ into k bins optimally it suffices to try every candidate last-bin start $i+1$: the last bin covers observations $i+1,\ldots,j$ with cost $c(i+1,j)$, and the preceding i observations are already optimally split at cost $\text{dp}[k-1][i]$. Taking the minimum over i gives the recurrence ($k \geq 2$, $j \geq k$):

$$\text{dp}[k][j] = \min_{k-1\,\leq\, i\,<\, j} \Big\{ \text{dp}[k-1][i] + c(i+1,\, j) \Big\}.$$

Solution. The optimal K-partition cost is $\mathrm{dp}[K][n]$. Bin boundaries are recovered by backtracking the argmin at each step.

Algorithm 1 gives the complete procedure.

Algorithm 1 CRPS-optimal K-partition via dynamic programming

Require: Sorted observations $(x_{(1)}, y_1), \ldots, (x_{(n)}, y_n)$; number of bins K
Ensure: Optimal bin boundaries $0 = b_0 < b_1 < \cdots < b_K = n$

Phase 1: Precompute cost matrix

for $i = 1$ **to** n **do**
  Initialise Fenwick trees T_{cnt}, T_{sum}; running total $\Sigma \leftarrow 0$; $W \leftarrow 0$
  for $j = i$ **to** n **do**
    Insert y_j into T_{cnt} and T_{sum}; update $\Sigma \leftarrow \Sigma + y_j$
    $r \leftarrow T_{\mathrm{cnt}}.\mathrm{prefixQuery}(y_j)$; $\quad S_{\le} \leftarrow T_{\mathrm{sum}}.\mathrm{prefixQuery}(y_j)$; $\quad S_{>} \leftarrow \Sigma - S_{\le}$
    $W \leftarrow W + y_j \cdot r - S_{\le} + S_{>} - y_j \cdot (j - i + 1 - r)$
    $m \leftarrow j - i + 1$
    $c[i][j] \leftarrow \begin{cases} m\,W/(m-1)^2 & \text{if } m \ge 2 \\ +\infty & \text{if } m = 1 \end{cases}$
  end for
end for

Phase 2: Fill DP table

for $j = 2$ **to** n **do**
  $\mathrm{dp}[1][j] \leftarrow c[1][j]$; $\quad \mathrm{split}[1][j] \leftarrow 0$
end for
for $k = 2$ **to** K **do**
  for $j = k$ **to** n **do**
    $\mathrm{dp}[k][j] \leftarrow +\infty$
    for $i = k - 1$ **to** $j - 1$ **do**
      $v \leftarrow \mathrm{dp}[k-1][i] + c[i+1][j]$
      if $v < \mathrm{dp}[k][j]$ **then**
        $\mathrm{dp}[k][j] \leftarrow v$; $\quad \mathrm{split}[k][j] \leftarrow i$
      end if
    end for
  end for
end for

Phase 3: Backtrack boundaries

$b_K \leftarrow n$; $\quad k \leftarrow K$; $\quad j \leftarrow n$
while $k \ge 1$ **do**
  $b_{k-1} \leftarrow \mathrm{split}[k][j]$; $\quad j \leftarrow b_{k-1}$; $\quad k \leftarrow k - 1$
end while
return $(b_0, b_1, \ldots, b_K)$

6 Precomputation and Complexity

Precomputing $W(i,j)$. Fix left endpoint i and scan $j = i, i+1, \ldots, n$. Let $m = j - i + 1$ denote the current number of elements in the bin $[i,j]$. Adding y_{j+1} to the set $\{y_i, \ldots, y_j\}$ increases the pairwise dispersion W by $\Delta W = \sum_{\ell=i}^{j} |y_\ell - y_{j+1}|$. To evaluate this sum in $O(\log n)$ without iterating over all m elements, split by sign. Let r denote the number of existing values $\leq y_{j+1}$ (the *rank* of y_{j+1} in the current set), $S_{\leq} = \sum_{\ell:\, y_\ell \leq y_{j+1}} y_\ell$, and $S_{>} = \sum_{\ell:\, y_\ell > y_{j+1}} y_\ell$. Then

$$\begin{aligned}\Delta W &= \sum_{\ell:\, y_\ell \leq y_{j+1}} (y_{j+1} - y_\ell) \;+ \sum_{\ell:\, y_\ell > y_{j+1}} (y_\ell - y_{j+1}) \\ &= y_{j+1} \cdot r - S_{\leq} \;+\; S_{>} - y_{j+1} \cdot (m - r).\end{aligned}$$

The quantities r and $S_{\leq}$ are *prefix queries*: a count and a sum over all stored values up to a query point. Maintaining a Fenwick tree (Binary Indexed Tree) indexed by value rank provides both in $O(\log n)$ per insertion; $S_{>}$ follows as the running total minus $S_{\leq}$ (see next paragraph for details). Since there are $O(n)$ left endpoints and $O(n)$ extensions for each, the total precomputation is $O(n^2 \log n)$ with $O(n^2)$ storage.

The choice of the Fenwick tree. Two Fenwick trees [5] suffice: one storing counts (to obtain r) and one storing values (to obtain $S_{\leq}$). A sorted array would answer prefix queries in $O(1)$ but requires $O(m)$ per insertion, raising the total precomputation to $O(n^3)$. A balanced binary search tree (e.g. an order-statistics tree) achieves the same $O(\log n)$ per operation but with higher constant factors and greater implementation complexity. The Fenwick tree is therefore the natural choice: simple, cache-friendly, and sufficient to attain the $O(n^2 \log n)$ bound.

DP. The recurrence evaluates in $O(n^2 K)$ time. There are K layers and $O(n)$ entries per layer; filling each entry $\mathrm{dp}[k][j]$ requires scanning $O(n)$ candidate split points i, with each lookup of $c(i+1, j)$ taking $O(1)$ from the precomputed table.

Quadrangle inequality and tightness of $O(n^2 K)$. If the cost function $c(i,j)$ satisfied the *quadrangle inequality* (QI), $c(a,c) + c(b,d) \leq c(a,d) + c(b,c)$ for all $a \leq b \leq c \leq d$, then by the Knuth–Yao theorem [6, 7] the optimal split points would be monotone in j, enabling a divide-and-conquer fill of each DP row in $O(n)$ rather than $O(n^2)$, reducing the total to $O(nK)$. However, the LOO-CRPS cost *violates* the QI already at $n = 4$: for $\mathbf{y} = (0,0,0,0,0,1,1,1,1,1)$ and the quadruple $(a,b,c,d) = (0,3,4,5)$ the inequality fails by a gap of 0.30. The mechanism is the $m/(m-1)^2$ prefactor, which is concave in m for small m, making the cost disproportionately sensitive to small bins. The $O(n^2 K)$ complexity is therefore tight for the LOO-CRPS cost.

7 Selecting K

The DP finds the optimal partition, i.e. the one that minimises total LOO-CRPS, for a *fixed* K. A natural approach to selecting K is to treat the total LOO-CRPS $\mathrm{dp}[K][n]$ as a function

of K and pick its minimum[1]. When taking this approach, one must be careful not to optimise jointly the partition and its LOO evaluation on the same data, as this would lead to a biased underestimate of generalisation error (see in-sample optimism in model selection [8, 9]).

Running example. To illustrate the methods of this and subsequent sections, we use a heteroscedastic synthetic dataset with $n = 1000$ observations: $X_i \sim \text{Uniform}(0, 3)$, $Y_i \mid X_i = x \sim \mathcal{N}(3x, (1 + x)^2)$. The conditional mean grows from 0 to 9 and the conditional standard deviation from 1 to 4 over $[0, 3]$, giving a $16\times$ increase in variance. Observations are sorted by x before all subsequent steps. A full analysis is given in Section 9.

7.1 K-fold cross-validated K selection

The correct remedy is to separate partition optimisation from model selection via K-fold cross-validation (we use F for the number of folds to avoid confusion with the number of bins K). The sorted observations are assigned to folds by interleaving: observation i (in the x-sorted order) goes to fold $i \bmod F$, so that every fold inherits the x-sorted structure. For each fold $f \in \{0, \ldots, F-1\}$, let $\mathcal{T}_f$ and $\mathcal{V}_f$ denote the training and test subsets. For each candidate K, find the optimal K-partition $\hat{\mathcal{P}}_K^{(f)}$ on $\mathcal{T}_f$ using the DP, then evaluate the test CRPS on $\mathcal{V}_f$:

$$\text{TestCRPS}_f(K) = \frac{1}{|\mathcal{V}_f|} \sum_{(x_i, y_i) \in \mathcal{V}_f} \text{CRPS}\Big(\hat{F}^{(f)}_{b(x_i)},\, y_i\Big)\,,$$

where $b(x_i)$ is the bin in $\hat{\mathcal{P}}_K^{(f)}$ that contains x_i and $\hat{F}^{(f)}_{b(x_i)}$ is the ECDF of the training y-values in that bin. The final criterion averages across folds:

$$\overline{\text{TestCRPS}}(K) = \frac{1}{F} \sum_{f=0}^{F-1} \text{TestCRPS}_f(K), \qquad K^* = \operatorname*{arg\,min}_{K=1,\ldots,K_{\max}} \overline{\text{TestCRPS}}(K).$$

We use $F = 5$ folds and set $K_{\max} = \lfloor n/10 \rfloor$ throughout, ensuring the minimum expected bin size on each training fold is at least 8; values larger than $\lfloor n/10 \rfloor$ typically produce nearly empty bins and are penalised heavily by test CRPS regardless.

Cross-validation guards against in-sample optimism: a partition that overfits the training bins will incur high test CRPS, producing a genuine U-shaped curve in K and a data-driven K^*. In small datasets the variance of $\text{TestCRPS}(K)$ is high, and the selected K^* may still imply bins with few observations; the CV criterion is choosing the best available tradeoff, not guaranteeing any minimum bin size. The resulting small-bin coarseness manifests as loss of interval efficiency rather than loss of coverage validity, since Proposition 2 holds for any $m \geq 2$.

[1] An as alternative, one could adopt the nonparametric Bayesian approach to selecting K is to place a Dirichlet process prior over the partition; the number of clusters is then determined by the concentration parameter rather than cross-validation. While principled, this approach does not in general minimise CRPS, and the choice of concentration parameter introduces a separate hyperparameter. We use CV as a simpler, assumption-free alternative.

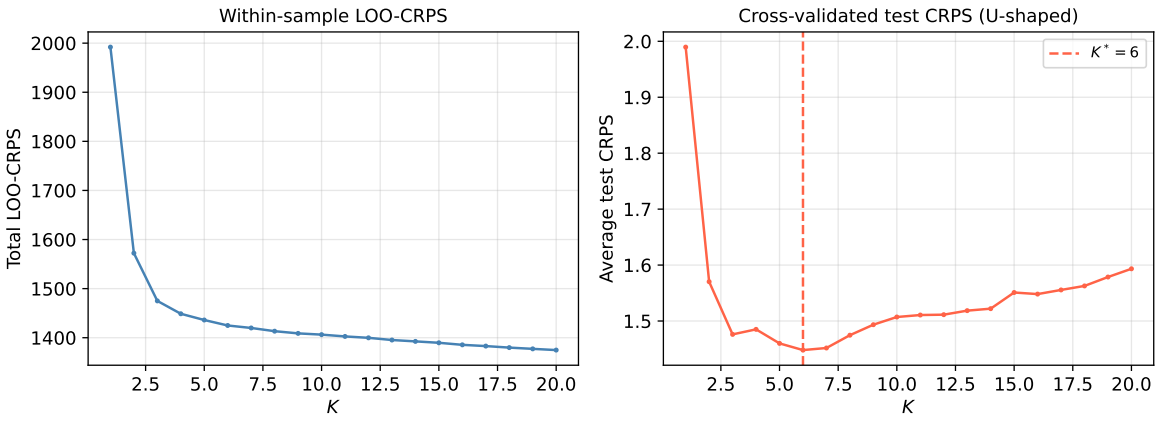

Figure 2: Within-sample LOO-CRPS (left) and cross-validated test CRPS (right) as functions of K on the running example. The within-sample criterion is nearly monotone decreasing, confirming its susceptibility to in-sample optimism. The test CRPS has a clear U-shape with minimum at $K^* = 6$. Note that the left and right y-axis scales differ: the left shows a total (sum over all observations), the right an average (per held-out test point).

8 Predictive Distributions: Venn Prediction and Conformal Inference

Having selected K^* and fitted the partition on all data, each test point x^* falls in some bin B_k with m training observations $y_1, \ldots, y_m$. We describe two complementary ways to form a predictive distribution (or set) for the response y^*, both motivated by the Venn predictor framework.

8.1 The Venn Prediction Band

Following Vovk et al., the *Venn prediction* for y^* is the family of augmented empirical CDFs indexed by the hypothetical label $y_h \in \mathbb{R}$:

$$F_h(t) = \frac{1}{m+1}\Big(\#\{i : y_i \le t\} + \mathbf{1}[y_h \le t]\Big).$$

Each F_h is a valid CDF; $\{F_h : y_h \in \mathbb{R}\}$ is the set of all distributional predictions consistent with one additional observation in the bin.

At each t, the band spans

$$\underline{F}(t) = \frac{\#\{i : y_i \le t\}}{m+1}, \qquad \overline{F}(t) = \frac{\#\{i : y_i \le t\} + 1}{m+1},$$

a constant width of $1/(m+1)$ everywhere. In terms of the training ECDF $\hat{F}_m(t) = m^{-1}\#\{i : y_i \le t\}$:

$$\underline{F}(t) = \frac{m}{m+1}\,\hat{F}_m(t), \qquad \overline{F}(t) = \underline{F}(t) + \frac{1}{m+1}.$$

The width $1/(m+1)$ vanishes as m grows, reflecting that the hypothetical label contributes negligible uncertainty relative to the training ECDF in large bins. Under exchangeability of

$(y_1, \ldots, y_m, y^*)$, the true F_{y^*} lies in the band by construction, making this a valid multiprobabilistic prediction in the sense of Vovk et al.; it is the direct set-valued analog of the Venn predictor interval in binary classification.

Limitation. Figure 3 shows the Venn band for each bin of the synthetic example. The band width $1/(m+1)$ is determined entirely by bin size m and carries no information about the local density or the position of y^* within the bin. A more informative object is the conformal prediction set below.

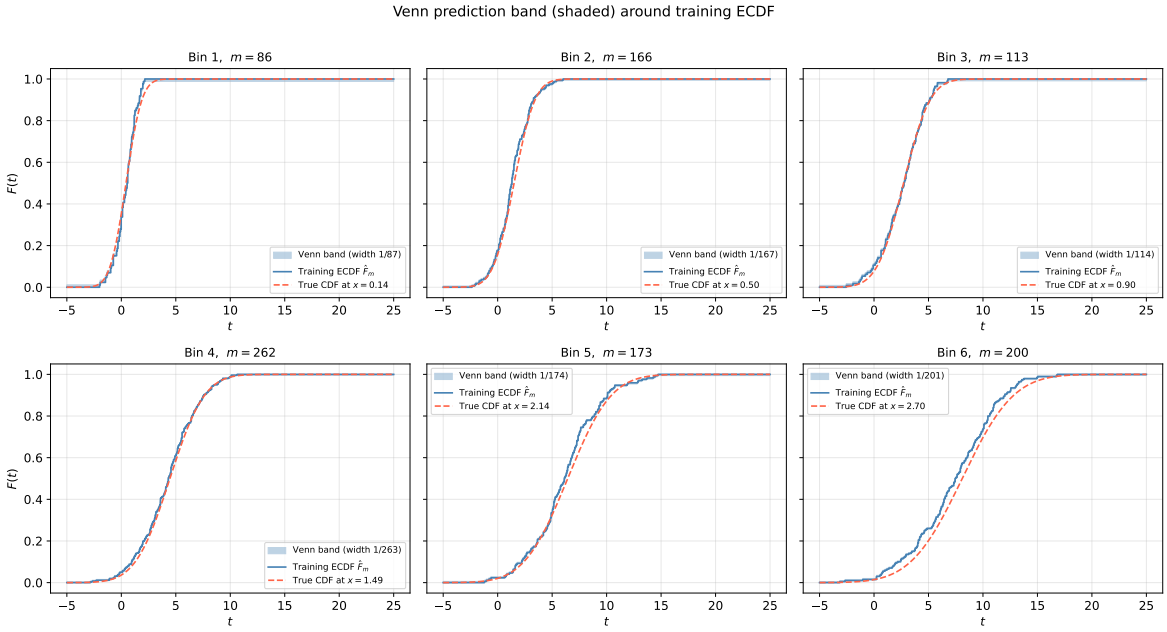

Figure 3: Venn prediction band (shaded) and training ECDF $\hat{F}_m$ (step function) for each of the six bins on the running example, alongside the true conditional CDF at the bin midpoint (dashed red). The band width $1/(m+1)$ ranges from 0.004 to 0.012 across bins, invisible at these bin sizes.

8.2 Conformal Prediction Set

We briefly recall the key concepts of conformal prediction [12] in the context of our setting and we show how to obtain valid prediction sets.

Nonconformity score. For test candidate y_h, define

$$\alpha(y_h) = \mathrm{CRPS}(\hat{F}_m,\, y_h) = \frac{1}{m}\sum_{i=1}^{m} |y_i - y_h| - \frac{W}{m^2},$$

where $W = \sum_{i<j} |y_i - y_j|$ is the pairwise dispersion of the training bin (independent of y_h). For each training observation y_j $(j = 1, \ldots, m)$ in the augmented set $\{y_1, \ldots, y_m, y_h\}$, define

the leave-one-out nonconformity score

$$\alpha_j(y_h) = \mathrm{CRPS}(F_{m+1}^{(-j)},\, y_j),$$

where $F_{m+1}^{(-j)}$ is the ECDF of $\{y_1, \ldots, y_m, y_h\} \setminus \{y_j\}$. Write $\alpha_{m+1}(y_h) \equiv \alpha(y_h)$ for the test score.

Conformal p-value.

$$p(y_h) = \frac{1}{m+1}\, \#\big\{j \in \{1, \ldots, m+1\} : \alpha_j(y_h) \geq \alpha(y_h)\big\}.$$

Prediction set.

$$\Gamma^{\varepsilon} = \{y_h \in \mathbb{R} : p(y_h) > \varepsilon\}.$$

Proposition 2. *Under exchangeability of* $(y_1, \ldots, y_m, y^*)$, $\Pr(y^* \in \Gamma^{\varepsilon}) \geq 1 - \varepsilon$.

Proof. Under exchangeability, any permutation of $(y_1, \ldots, y_m, y^*)$ produces the same multiset of nonconformity scores $\{\alpha_j(y^*)\}_{j=1}^{m+1}$. Hence the rank $R$of $\alpha(y^*) = \alpha_{m+1}(y^*)$ from largest to smallest is uniformly distributed on $\{1, \ldots, m+1\}$ (with ties broken uniformly). By definition, $p(y^*) = R/(m+1)$, so

$$\Pr\big(y^* \notin \Gamma^{\varepsilon}\big) = \Pr\big(p(y^*) \leq \varepsilon\big) = \Pr\big(R \leq \lfloor \varepsilon(m+1) \rfloor\big) = \frac{\lfloor \varepsilon(m+1) \rfloor}{m+1} \;\leq\; \varepsilon. \qquad \square$$

The proof establishes $\Pr(p(y^*) \leq \varepsilon) \leq \varepsilon$ for every $\varepsilon \in (0, 1)$: the p-value is *super-uniform*, meaning its CDF is bounded above by the CDF of a Uniform$(0, 1)$ random variable. Equivalently, $\Pr(p(y^*) > \varepsilon) \geq 1 - \varepsilon$, which is the coverage guarantee. The inequality is strict whenever $\varepsilon(m+1)$ is not an integer, reflecting the discreteness of the p-value on the grid $\{1/(m+1), \ldots, 1\}$.

Remark on exchangeability. The proof of Proposition 2 relies on two ingredients:

1. **Score symmetry.** Given a fixed set of $m+1$ values in a bin, any permutation of $(y_1, \ldots, y_m, y^*)$ produces the same multiset of LOO nonconformity scores. This is a property of the LOO-CRPS computation, which depends only on the multiset, not on which element is designated as the test point. Score symmetry always holds.

2. **Statistical exchangeability of bin members.** Under i.i.d. sampling, the $m+1$ values in the bin must be exchangeable *as random variables*. This requires that the event "these particular observations share a bin" does not introduce dependence among their y-values.

It is condition (2) that is violated by a data-dependent partition. The DP uses all y-values to determine bin boundaries, so conditioning on the event that a particular set of observations shares a bin acts as a y-dependent filter: the observations ended up together partly *because* their y-values were mutually compatible with the DP's cost criterion. This selection effect biases the within-bin y-distribution away from the unconditional i.i.d. model.

A concrete thought experiment makes this vivid. Consider an observation near a bin boundary. If its y-value were very different from its bin-mates, the DP might have placed the boundary on the other side, assigning it to the adjacent bin. Conditioning on it being in *this* bin therefore biases its y-distribution toward values compatible with the present companions.

The violation is mild in practice. The violation is not in the score computation (which is permutation-invariant for any fixed bin) but in the probability model over which observations end up together. For the partition to change, a single observation must alter the DP cost matrix substantially enough to shift a boundary — an event that becomes increasingly unlikely as bin sizes grow, because one observation has diminishing influence on the cost of a bin with $m \gg 1$ members. The residual approximation error is governed by the smoothness of the conditional distribution relative to the bin width; the bias–variance analysis in Section 8.5 quantifies the trade-off.

Theoretical support. Allen et al. [26] offer a complementary perspective: their Theorem 2.1 shows that any conformal binning procedure whose within-bin predictor is in-sample auto-calibrated yields valid out-of-sample calibration guarantees under exchangeability, without requiring the partition to be fixed independently of the data. Because our LOO-CRPS scores are exactly calibrated within each fixed bin (Proposition 2), this result provides formal support for the approximate validity observed in our experiments.

Structure of Γ^ε. The score $\alpha(y_h)$ is convex and piecewise linear in y_h, with breakpoints at $y_1, \ldots, y_m$ and slopes ± 1; hence $\{y_h : \alpha(y_h) \leq c\}$ is a closed interval for any $c \geq 0$. The training scores $\alpha_j(y_h)$ also depend on y_h, but each $F_{m+1}^{(-j)}$ differs from $\hat{F}_m$ by $O(1/m)$, so for large m they are approximately constant. Because $\alpha(y_h)$ is convex in y_h (Section 8.4), Γ^ε is a connected interval centred near the empirical median of $\hat{F}_m$, with width shrinking as m grows.

The suitability of LOO-CRPS as a nonconformity score Several properties make the LOO-CRPS a natural and principled nonconformity measure in this setting.

Properness implies sensitivity to genuine nonconformity. Because CRPS is strictly proper, the expected score $\mathbb{E}_P[\mathrm{CRPS}(\hat{F}, Y)]$ is uniquely minimised when $\hat{F} = P$. Consequently, $\alpha(y_h) = \mathrm{CRPS}(\hat{F}_m, y_h)$ is large precisely when y_h is surprising under the within-bin training distribution, not merely when it is far from some summary statistic such as the mean or median. Nonconformity scores based on absolute residuals $|y_h - \hat{\mu}|$ (or quantile residuals) implicitly assume a parametric location or quantile model for the bin; CRPS makes no such assumption, which matters when the within-bin distribution is skewed or multimodal.

The LOO structure guarantees exchangeability. Conformal validity requires that the scores $\alpha_1(y_h), \ldots, \alpha_m(y_h), \alpha(y_h)$ be *exchangeable* under exchangeability of $(y_1, \ldots, y_m, y^*)$. This holds here because every score is computed in exactly the same way: each α_j is the CRPS of the m-atom ECDF of the remaining m elements of $\{y_1, \ldots, y_m, y_h\}$ evaluated at y_j, and $\alpha(y_h)$ is the same quantity for the test element. If instead we used the full-sample $\hat{F}_m$ (without the LOO adjustment) to score both training and test points, the training scores $\mathrm{CRPS}(\hat{F}_m, y_j)$ and the test score $\mathrm{CRPS}(\hat{F}_m, y^*)$ would not be on equal footing: the training ECDF was fitted using y_j, but not using y^*, breaking the symmetry on which conformal coverage rests.

Coherence with the binning criterion. The DP selects bin boundaries by minimising total LOO-CRPS on the training set. The conformal step then evaluates the test point by the same

quantity: the marginal LOO-CRPS of adding y^* to the selected bin. This coherence means the bin is already optimised for the task of distributional prediction, and the prediction set Γ^ε has the natural interpretation as the set of test values whose inclusion would not increase the bin's per-observation LOO-CRPS beyond the level seen at the training points (see also Section 8.3). Figure 4 illustrates the p-value curve at three test locations.

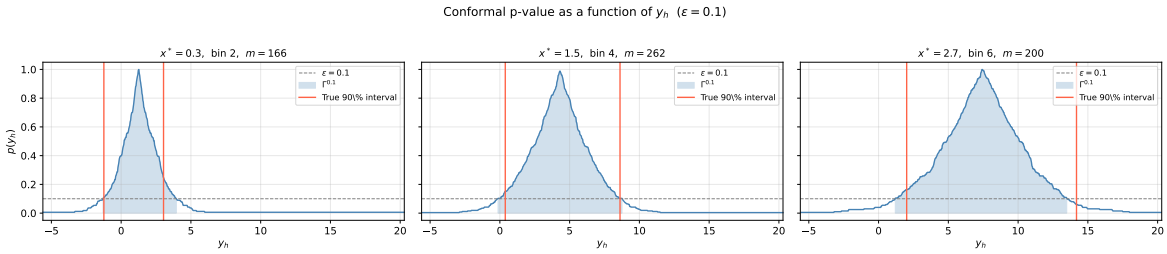

Figure 4: Conformal p-value $p(y_h)$ as a function of the candidate response y_h at three test points $x^* \in \{0.3, 1.5, 2.7\}$ for $\varepsilon = 0.10$. The shaded region is the prediction set $\Gamma^{0.10}$; vertical red lines mark the true 90% interval under $\mathcal{N}(3x^*, (1+x^*)^2)$. The p-value curve is unimodal (each monotone piece is convex, since the underlying nonconformity score $\alpha(y_h)$ is convex), yielding a single connected prediction set in each case.

8.3 Connection to the DP Cost

The sum of all $m+1$ nonconformity scores in the augmented set equals the total LOO-CRPS of that set:

$$\sum_{j=1}^{m+1} \alpha_j(y_h) = \text{cost}\big(\{y_1, \ldots, y_m, y_h\}\big) = \frac{m+1}{m^2}\, W\big(\{y_1, \ldots, y_m, y_h\}\big).$$

The test score $\alpha(y_h) = \text{CRPS}(\hat{F}_m, y_h)$ measures how much y_h increases the mean absolute deviation from the bin; it is large when y_h is far from the bulk of $y_1, \ldots, y_m$. The prediction set Γ^ε therefore consists of values whose addition would not dramatically inflate the DP cost of bin B_k.

The Venn band and conformal prediction set are thus both consistent with the DP's cost structure: the band provides a marginal-distribution guarantee with trivially computable constant width $1/(m+1)$, while Γ^ε provides a finite-sample coverage guarantee at the chosen level ε with data-adaptive width. Figure 5 shows the resulting prediction intervals across the covariate range.

8.4 Interval Structure and Non-Convex Extensions

Γ^ε is approximately a single interval. The convexity of the CRPS nonconformity score is not incidental; it is a structural consequence of measuring average distance to the training distribution. Writing

$$\alpha(y_h) = \frac{1}{m}\sum_{i=1}^{m} |y_i - y_h| - \frac{W}{m^2},$$

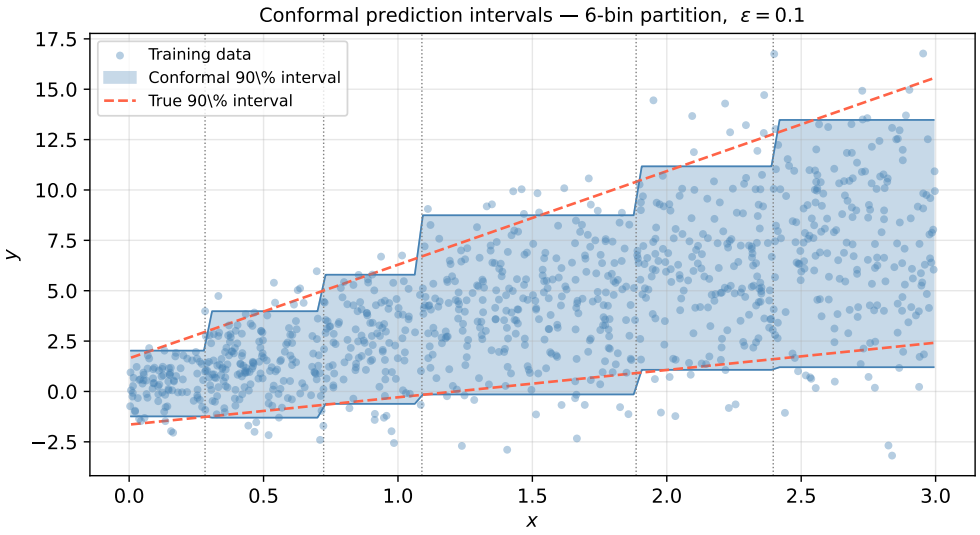

Figure 5: Conformal 90% prediction intervals (shaded blue) and true 90% intervals (dashed red) across the covariate range, with bin boundaries marked by dotted vertical lines. The interval width adapts to the within-bin spread, tracking the increasing conditional variance of the heteroscedastic data-generating process.

the second term is constant in y_h, and the first is a sum of convex functions $|y_i - y_h|$, hence convex with a unique minimum at the empirical median of $\{y_1, \ldots, y_m\}$. More generally, any score of the form $\mathbb{E}_{\hat{F}_m}[\ell(y_h, Y)]$ with ℓ convex in its first argument (e.g. squared error, absolute error, CRPS) inherits this property. Consequently, the sublevel set $\{y_h : \alpha(y_h) \leq c\}$ is a closed interval for every $c \geq 0$.

In a split-conformal predictor the training scores do not depend on y_h, so $p(y_h)$ is non-increasing in $\alpha(y_h)$ and convexity of α directly implies that Γ^ε is a connected interval. In the present transductive (full-data) setting all $m+1$ scores depend on y_h, and their relative ranking can in principle change as y_h varies, so the split-conformal argument does not apply directly. We conjecture that Γ^ε remains a connected interval whenever $\alpha(y_h)$ is convex; in all our experiments (synthetic, Old Faithful, motorcycle; m ranging from 22 to 262) the prediction set is a single interval without exception.

The consequence for multimodal bins is concrete. Suppose the training responses in B_k are bimodal with well-separated modes at $\mu_1 \ll \mu_2$. The empirical median lies in the inter-modal gap; $\alpha(y_h)$ attains its minimum there. The resulting Γ^ε spans both modes and the low-density gap between them: it is valid (marginal coverage is guaranteed) and correctly wide, but informationally wasteful as it assigns coverage to a region of near-zero probability mass. The Venn ECDF $\hat{F}_m$ represents the bimodal shape correctly as a distributional object, but inverting the convex CRPS score to form a prediction *set* discards this information.

A bandwidth-free non-convex alternative: the k-NN score. Non-contiguous prediction sets require a nonconformity score that is non-convex in y_h, i.e. one that is simultaneously small near each mode and large in the inter-modal gap. Density-based scores such as $-\log \hat{f}(y_h)$ achieve this but require a bandwidth. In one dimension, the k-nearest-neighbour

(k-NN) distance provides a bandwidth-free alternative. Define

$$\alpha^{(k)}(y_h) = d_{(k)}\big(y_h,\ \{y_1, \ldots, y_m\}\big),$$

the k-th smallest value of $|y_h - y_i|$ over $i = 1, \ldots, m$. For $k = 1$ this is simply $\min_i |y_h - y_i|$, which has a local minimum at every training point and is large precisely where the training distribution is sparse.

The LOO version for the conformal construction is defined symmetrically in the augmented set $\{y_1, \ldots, y_m, y_h\}$: for each training observation y_j,

$$\alpha_j^{(1)}(y_h) = \min\Big(\min_{i \neq j} |y_j - y_i|,\ |y_j - y_h|\Big),$$

and $\alpha_{m+1}^{(1)}(y_h) \equiv \alpha^{(1)}(y_h)$. Under exchangeability of $(y_1, \ldots, y_m, y^*)$, the conformal p-value

$$p^{(1)}(y_h) = \frac{1}{m+1} \#\big\{j : \alpha_j^{(1)}(y_h) \geq \alpha^{(1)}(y_h)\big\}$$

is super-uniform, and Proposition 2 holds without modification.

The prediction set $\Gamma^{\varepsilon,(1)} = \{y_h : p^{(1)}(y_h) > \varepsilon\}$ is approximately the union

$$\bigcup_{i=1}^{m} \big[y_i - c_\varepsilon,\ y_i + c_\varepsilon\big],$$

where c_ε is the $(1-\varepsilon)$-quantile of the training LOO scores (exact in the limit $m \to \infty$ where the y_h-dependence of $\alpha_j(y_h)$ vanishes). For a bimodal distribution, training points cluster near both modes, so c_ε is set by the intra-cluster spacing; if this is smaller than half the inter-modal gap, the prediction set consists of two disjoint intervals.

The effective resolution is determined purely by the data; no bandwidth is specified by the user. This adaptivity is rooted in a classical result from one-dimensional order statistics: the 1-NN distance is a consistent density estimator without smoothing, since $m \cdot \min_i |y_h - Y_i|$ converges in distribution to $\mathrm{Exp}(f(y_h))$ when $Y_1, \ldots, Y_m \sim f$ [11]. The conformal threshold c_ε therefore plays the role of a density threshold, automatically calibrated for $1-\varepsilon$ coverage. The resulting $\Gamma^{\varepsilon,(1)}$ is a non-parametric highest-density region (HDR) of the within-bin empirical distribution.

Binning remains LOO-CRPS optimal. The k-NN modification applies only to the conformal step; the DP binning continues to use the LOO-CRPS cost of Section 4. This separation is principled: the binning objective is to group covariate-sorted observations so that the within-bin ECDF is a good predictive distribution for the local conditional $P(Y \mid X = x)$, which is a question of predictive accuracy for which LOO-CRPS is the correct criterion regardless of whether the within-bin distribution is unimodal or multimodal.

Using $\sum_{j \in B_k} \alpha_j^{(1)}$ as the DP cost would be computationally feasible (the cost is additive over bins, so optimal substructure is preserved), but it would measure local y-density rather than predictive quality, and it would exhibit the same in-sample optimism as within-sample LOO-CRPS: a size-2 bin with $|y_1 - y_2| \approx 0$ has near-zero 1-NN cost, driving the DP to over-partition. The two steps therefore use different criteria for different purposes: LOO-CRPS for reference-class selection, and either CRPS or k-NN for conformal evaluation.

Synthetic illustration. Figure 6 illustrates the contrast on a synthetic bimodal bin. A single bin of $m = 50$ training observations is drawn from $0.5\mathcal{N}(-3, 0.5^2) + 0.5\mathcal{N}(3, 0.5^2)$. The left panel shows the CRPS nonconformity score $\alpha(y_h)$: it is convex with a minimum at the inter-modal gap, so Γ^{ε} is a single wide interval $[-5.5, 5.5]$ that spans both modes and the low-density region between them. The right panel shows the 1-NN score $\alpha^{(1)}(y_h)$: it has local minima near each cluster of training points and is large in the gap, so $\Gamma^{\varepsilon,(1)}$ decomposes into two disjoint intervals, one around each mode. Both prediction sets carry the super-uniform coverage guarantee; empirical coverage and set-size comparisons over many realisations are given in Table 1 below.

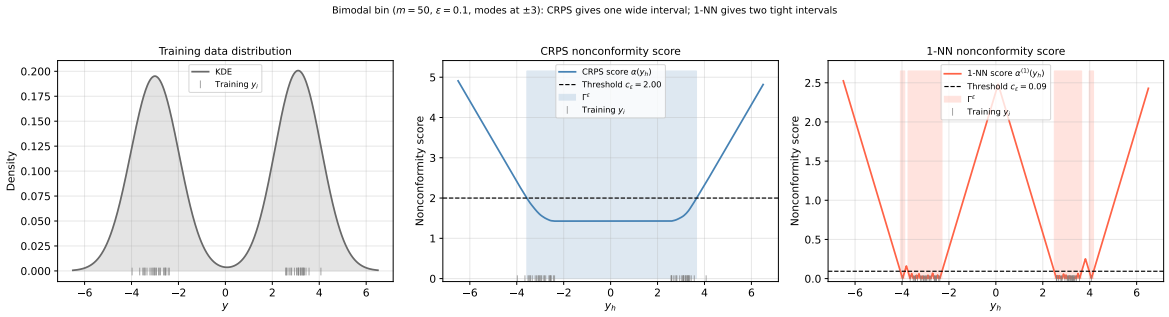

Figure 6: Synthetic bimodal bin ($m = 50$, $\varepsilon = 0.10$, two modes at ± 3). *Left:* training data distribution (KDE + rug), showing the two clearly separated modes. *Centre:* the CRPS nonconformity score is convex; the prediction set Γ^{ε} is a single interval spanning both modes. *Right:* the 1-NN score has two local minima near the data clusters; the prediction set $\Gamma^{\varepsilon,(1)}$ consists of two tight intervals, one near each mode, with the inter-modal gap excluded. Rug marks indicate the $m = 50$ training observations.

Empirical coverage. Table 1 summarises empirical coverage and mean set measure (total Lebesgue measure of the prediction set) averaged over $R = 500$ independent realisations of the bimodal Data Generating Process, each with $m = 50$ training and $m_{\text{test}} = 500$ test observations drawn from the same mixture. Both scores achieve valid coverage above the nominal $1 - \varepsilon = 0.90$, confirming the super-uniform guarantee. The 1-NN prediction set is on average 1.84× smaller in Lebesgue measure than the CRPS set, quantifying the efficiency gain from excluding the low-density inter-modal gap.

Score	Coverage (%)	Mean set measure
CRPS	92.6 ± 0.2	7.51 ± 0.01
1-NN	91.0 ± 0.3	4.09 ± 0.03

Table 1: Bimodal bin ($m = 50$, $\varepsilon = 0.10$): empirical coverage and mean Lebesgue measure of the prediction set, averaged over $R = 500$ seeds ($\pm$ one standard error). Both scores satisfy the super-uniform guarantee; the 1-NN set is 1.84× smaller by excluding the low-density inter-modal region.

8.5 Approximate Exchangeability and the Bias–Variance Trade-off in K

The conformal coverage guarantee (Proposition 2) requires the test point y^* to be exchangeable with the m training responses in its bin. This holds exactly when $P(Y \mid X = x)$ is constant across the bin's x-range, and is violated whenever the DGP is "non-stationary" within a bin, which is the generic situation. Two competing effects determine the severity of the violation.

Wide bins (small K). When a bin spans a broad range of x-values, the within-bin ECDF $\hat{F}_m$ averages over a region where $P(Y \mid X = x)$ varies. Exchangeability is most severely violated, and the resulting conformal intervals may be systematically mis-sized at a specific x^*: conservative where the local conditional spread is smaller than the bin average, anti-conservative where it is larger.

Narrow bins (large K). Narrower bins improve exchangeability but at two distinct costs. First, the conformal p-value takes values on the grid $\{1/(m+1), 2/(m+1), \ldots, 1\}$, so the number of calibration scores sets a hard floor on achievable interval widths. More sharply: for $m < \lceil 1/\varepsilon \rceil - 1$, no point can be excluded at level ε and the prediction set is the entire real line. At $\varepsilon = 0.10$ this floor falls at $m < 9$; bins approaching this threshold produce intervals that are valid but practically useless. Second, even well above the floor, with small m the conformal quantile is estimated from few nonconformity scores, so interval widths are highly variable across test instances. Proposition 2 holds for any $m \geq 2$, but the efficiency loss is severe.

Cross-validation as the trade-off criterion. The test CRPS of Section 7.1 penalises both failure modes jointly. A partition whose within-sample score benefits from accidental y-homogeneity will incur high CRPS on the held-out half, whose y-values are not atypically clustered. A partition with bins too narrow to represent the local conditional distribution will also incur high test CRPS, because the within-bin ECDF has too few atoms or assigns mass in the wrong region. The U-shaped minimum of TestCRPS(K) is therefore a genuine empirical optimum for predictive accuracy, and its finite minimum implies that further splitting is penalised more by the granularity cost than it gains from improved exchangeability.

9 Numerical Illustration

We illustrate the full pipeline on the running example (Section 7: $n = 1000$, $Y \mid X = x \sim \mathcal{N}(3x, (1+x)^2)$, $X \sim \text{Uniform}(0, 3)$).

Cross-validated K selection. The 5-fold CV procedure (Section 7.1) is applied with $K_{\max} = 20$ (see Figure 2). The within-sample LOO-CRPS is nearly monotone decreasing in K, exhibiting a behaviour not dissimilar to that commonly seen for training loss. The test

CRPS is U-shaped and attains its minimum at $K^* = 6$, yielding bins

$$B_1 : x \in [0.00,\ 0.28],\ \ m_1 = 86, \qquad B_2 : x \in [0.28,\ 0.72],\ \ m_2 = 166,$$
$$B_3 : x \in [0.72,\ 1.09],\ \ m_3 = 113, \qquad B_4 : x \in [1.09,\ 1.88],\ \ m_4 = 262,$$
$$B_5 : x \in [1.89,\ 2.39],\ \ m_5 = 173, \qquad B_6 : x \in [2.40,\ 3.00],\ \ m_6 = 200.$$

Figure 7 shows the resulting partition. It adapts to the jointly growing mean and variance: the narrow bins B_1 and B_3 isolate transition regions where the mean slope and rising σ interact most strongly, while the wider bins are supported by larger sample counts.

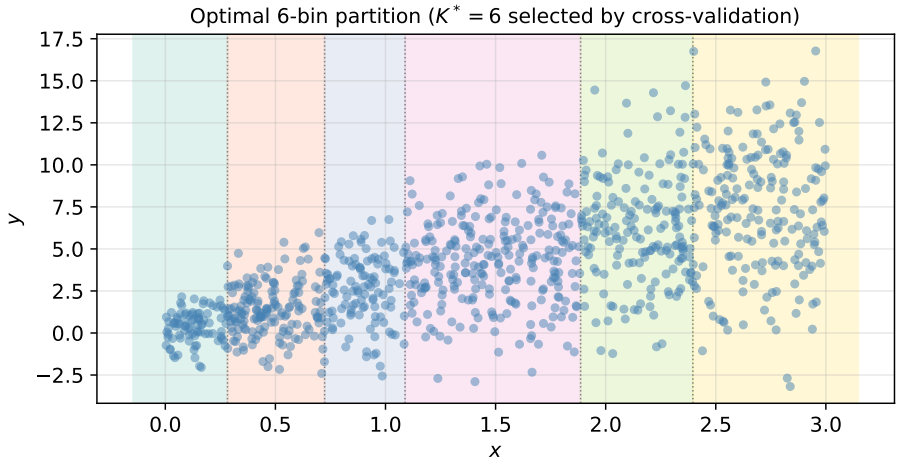

Figure 7: Training scatter with the optimal 6-bin partition. Shaded regions correspond to the six bins; dotted vertical lines mark the bin boundaries (midpoints between adjacent training observations).

Venn prediction band. For each bin the band width is $1/(m_k + 1)$, ranging from approximately 0.004 (bin 4) to 0.012 (bin 1) across the six bins. At these bin sizes the Venn band is invisible against the step function of the training ECDF, confirming that the band carries no practically useful information beyond the ECDF itself.

Conformal prediction intervals. Table 2 reports the 90% conformal prediction intervals ($\varepsilon = 0.10$) at three representative test points, alongside the true 90% intervals under the generating distribution.

The conformal intervals are slightly conservative at $x^* = 0.3$ and 1.5 (widths 5.28 vs 4.28; 8.89 vs 8.22), reflecting the residual within-bin heterogeneity, and nearly exact at $x^* = 2.7$ (width 12.27 vs 12.17). The near-exact match at the widest bin illustrates the method's adaptation: the high-variance region receives a bin large enough that the ECDF well represents the local spread.

Empirical coverage. On a fresh test set of 2,000 observations drawn from the same distribution, the empirical coverage is 94.6%, 89.1%, and 79.7% at $\varepsilon = 0.05$, 0.10, and 0.20 respectively. All three values meet or lie within the nominal guarantees' sampling uncertainty:

x^*	Bin	$\Gamma^{0.1}$	Width	True 90% width
0.3	2	$[-1.30,\ 3.98]$	5.28	4.28
1.5	4	$[-0.15,\ 8.74]$	8.89	8.22
2.7	6	$[\ 1.20,\ 13.48]$	12.27	12.17

Table 2: Conformal 90% prediction intervals at three representative x-values (one per non-consecutive bin). Widths grow monotonically with x, closely tracking the oracle under the true distribution.

at $\varepsilon = 0.10$ the shortfall of 0.9 pp is well within the ± 0.7 pp standard error. With 2,000 test points, the standard error on a coverage estimate near p is $\sqrt{p(1-p)/2000} \lesssim 0.7$ pp, so a single random draw is informative to within ± 1.4 pp at 95% confidence. Figure 8 shows the distribution of conformal p-values $p(y^*)$ on this test set; the near-uniform distribution is consistent with the super-uniform guarantee, with slight conservatism reflecting the approximate exchangeability within bins.

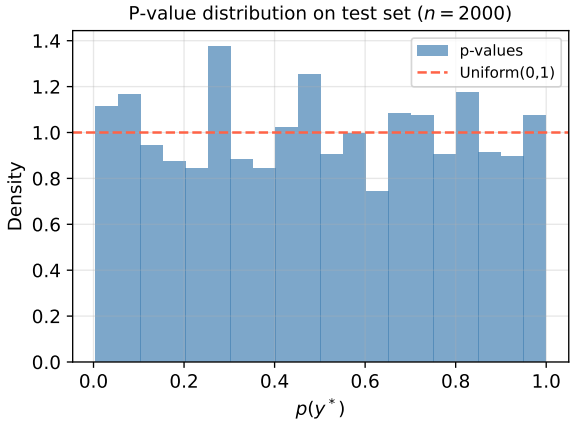

Figure 8: Distribution of conformal p-values $p(y^*)$ on the 2,000-point test set. Under a valid conformal predictor the p-values are super-uniform (density ≤ 1 everywhere); slight conservatism relative to the uniform baseline reflects the approximate exchangeability within bins of a heteroscedastic process. The p-values are discrete, taking values on $\{1/(m+1), 2/(m+1), \ldots, 1\}$ for the m training points in each bin; the histogram is smoothed by the varying bin sizes across test points.

Conditional coverage and calibration. Marginal coverage averages over all bins; a more demanding diagnostic is *conditional* coverage, evaluated separately within each bin. Figure 9 reports the empirical coverage at three levels. Bins 2–6 are close to the nominal level at all three thresholds. Bin 1 (the leftmost, $m = 86$) under-covers: test points near the right edge of the bin face a true conditional distribution shifted relative to the training ECDF, an inherent price of binning a continuously varying process.

Figure 10 shows the Probability Integral Transform (PIT) histograms per bin. Under perfect calibration the PIT values $\hat{F}_b(y^*)$ are uniform; the observed departures mirror the coverage pattern. Bin 1 shows a U-shape (boundary effect), while bins 5–6 exhibit a mild

rightward skew. The interior bins are close to uniform, confirming that the CRPS-optimal partition produces well-calibrated predictive distributions where the bin size is adequate.

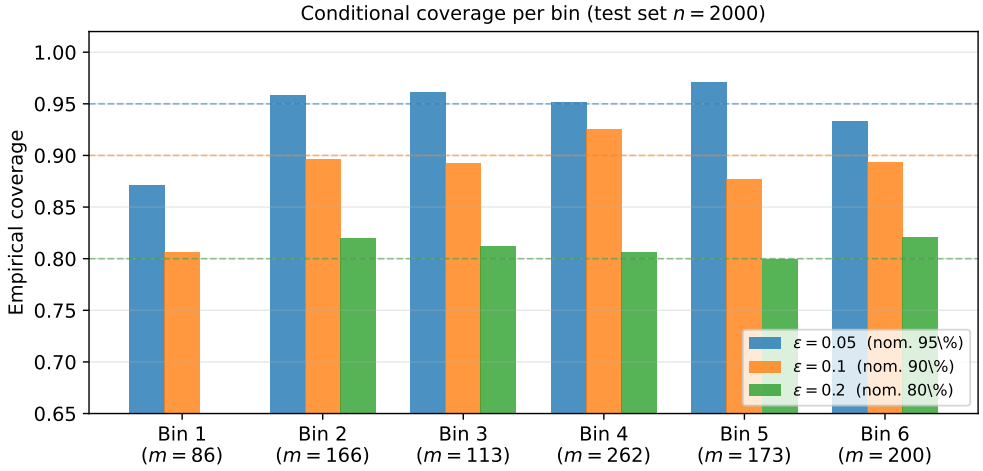

Figure 9: Conditional coverage per bin on the 2,000-point test set, at three levels $\varepsilon \in \{0.05, 0.10, 0.20\}$. Dashed lines mark the nominal level. Bin 1 ($m = 86$, leftmost) undercovers due to the within-bin shift of the true conditional distribution; the remaining bins are close to their nominal targets.

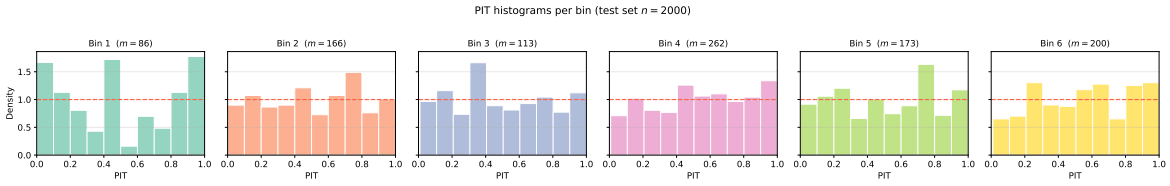

Figure 10: PIT histograms per bin. The PIT for a test point (x^*, y^*) assigned to bin b is $\hat{F}_b(y^*) = m_b^{-1} \sum_{i=1}^{m_b} \mathbf{1}[y_{b,i} \le y^*]$. Under perfect calibration the histogram is uniform (dashed line at density 1). Boundary bins (1 and 6) show departures due to the within-bin heterogeneity of the true conditional distribution.

10 Real-Data Experiments

We evaluate the method on two real datasets that stress different aspects of distributional non-stationarity: a bimodal conditional distribution and a strongly heteroscedastic one. All experiments use $\varepsilon = 0.10$ (nominal 90% coverage). Competitors are evaluated over $R = 200$ random 50/50 train/calibration splits, with standard errors reported throughout.

Competitors.

- **Gaussian split conformal.** OLS fit on the training half; absolute residuals on the calibration half determine the constant-width prediction interval.

- **CQR (cubic).** Conformalized Quantile Regression [14] with cubic polynomial base quantile regressors ($\tau = 0.05$ and 0.95), calibrated on the same held-out half.
- **CQR-QRF.** Quantile Regression Forest [16] with 500 trees fitted on the training half, conformalized with the $\max(q_{\alpha/2}(x) - y,\ y - q_{1-\alpha/2}(x))$ score (same as CQR) calibrated on the held-out half.
- **CQR-IDR.** Isotonic Distributional Regression [29] fitted on the training half, with quantile predictions at levels $\alpha/2$ and $1 - \alpha/2$ conformalized via the same CQR score on the calibration half.

Our method is transductive: all n observations are used for both partitioning and conformal calibration (full-data conformal with the within-bin ECDF as the nonconformity score), with no data held out. Split-conformal competitors must reserve a calibration set, effectively halving the sample available for model fitting. This data-efficiency advantage is inherent to full-data conformal methods [23]. To disentangle the effect of the method itself from this sample-size advantage, each table below reports two rows for our method: the *full-n* row reflects the method as deployed (transductive, no data splitting), while the $n/2$ row handicaps it to the same training set available to competitors, providing a controlled comparison.

10.1 Old Faithful: bimodal conditional distribution

The `faithful` dataset ($n = 272$) records eruption duration and waiting time between eruptions of the Old Faithful geyser. We set $x =$ waiting time (min) and $y =$ eruption duration (min). The marginal distribution of eruption duration is bimodal: short eruptions (≈ 2 min) and long eruptions (≈ 4.5 min), with the mixture proportion shifting as a function of waiting time. This is a natural stress-test for methods that assume a unimodal conditional distribution.

Cross-validation selects $K^* = 4$, with bin boundaries at waiting times 63.0, 67.5, and 71.5 minutes. Figure 11 (left) shows the partition; Figure 11 (right) shows the within-bin ECDFs. The two outer bins have clearly unimodal within-bin distributions (short eruptions for short waits, long eruptions for long waits), while the two inner bins capture the transition region with finer resolution.

Figure 12 compares 90% prediction intervals across methods. Gaussian split conformal and CQR produce intervals of roughly constant width across all waiting times, whereas our method adapts: narrower in both regimes where the conditional distribution is concentrated, and wide only in the transition region where the within-bin ECDF spans both modes.

Table 3 reports coverage and mean interval width averaged over $R = 200$ random 50/50 splits. In the matched-sample comparison (top block), our method ($n/2$, italic row) achieves mean width 1.27 min with slightly sub-nominal coverage ($88.5 \pm 0.3\%$), narrower than all competitors: Gaussian split conformal (1.68 min), CQR (1.49 min), CQR-IDR (1.29 min), and CQR-QRF (1.33 min). When all n observations are used (bold row, below the line), our method achieves $90.6 \pm 0.1\%$ coverage with mean width 1.22 min. All split-conformal methods exceed nominal coverage (91.2–91.4%).

10.2 Motorcycle accident: heteroscedastic benchmark

The `mcycle` dataset ($n = 133$) records head acceleration of a motorcycle dummy as a function of time after a simulated impact. The response variance changes dramatically: near-zero

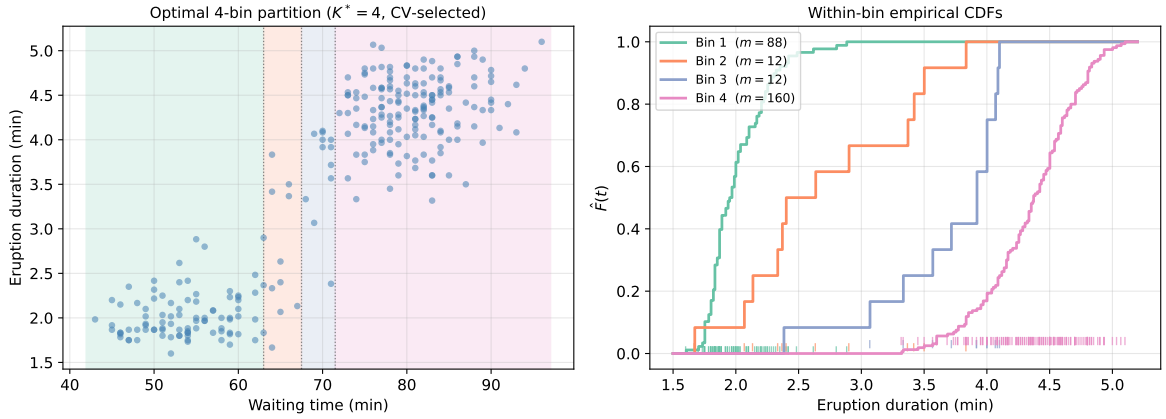

Figure 11: Old Faithful. *Left:* scatter of eruption duration vs. waiting time with the optimal 4-bin partition. Boundaries at 63.0, 67.5, and 71.5 minutes resolve the transition between the short-eruption and long-eruption regimes. *Right:* within-bin empirical CDFs; the outer bins are unimodal, confirming the partition captures the regime structure.

Method	Coverage (%)	Mean width (min)
Our method $(n/2)$	88.5 ± 0.3	1.270 ± 0.010
Gaussian split conformal	91.2 ± 0.0	1.683 ± 0.006
CQR (cubic)	91.2 ± 0.0	1.490 ± 0.006
CQR-QRF	91.4 ± 0.0	1.333 ± 0.006
CQR-IDR	91.4 ± 0.0	1.294 ± 0.007
Our method (full n**)**	90.6 ± 0.1	1.217 ± 0.002

Table 3: Old Faithful ($n = 272$): empirical coverage and mean width of prediction intervals at nominal level $1 - \varepsilon = 0.90$, averaged over $R = 200$ random 50/50 splits ($\pm$ one standard error). *Top block*: all methods use the same training half $(n/2)$; *Our method* $(n/2)$ is directly comparable to the competitors. *Below the line*: **Our method (full** n**)** uses all n observations, a design advantage inherent to full-data conformal methods.

before $\approx$ 15 ms, explosive in the 15–30 ms deformation phase, and moderating thereafter. This is the standard benchmark for heteroscedastic prediction intervals in the nonparametric regression literature.

Cross-validation selects $K^* = 6$, with boundaries at 15.1, 17.6, 24.4, 27.2, and 38.0 ms, clustered in the high-variance impact region. Figure 13 shows the K-selection curve and the resulting partition. With $n = 133$ observations, $K^* = 6$ implies bins of roughly 10–30 observations; the narrower bins in the chaotic 15–30 ms window improve within-bin exchangeability at the cost of fewer ECDF atoms. This is the small-sample manifestation of the bias–variance trade-off discussed in Section 8.5: the CV criterion selects the best available tradeoff, not a guaranteed minimum bin size, and the coarseness of the within-bin ECDF is absorbed into interval width rather than miscoverage.

Figure 14 and Table 4 compare prediction intervals averaged over $R = 200$ random 50/50 splits. In the matched-sample comparison (top block), Gaussian split conformal is drastically inefficient (172.4 g), CQR (cubic) and CQR-IDR are moderately better (134.1 g and 127.6 g

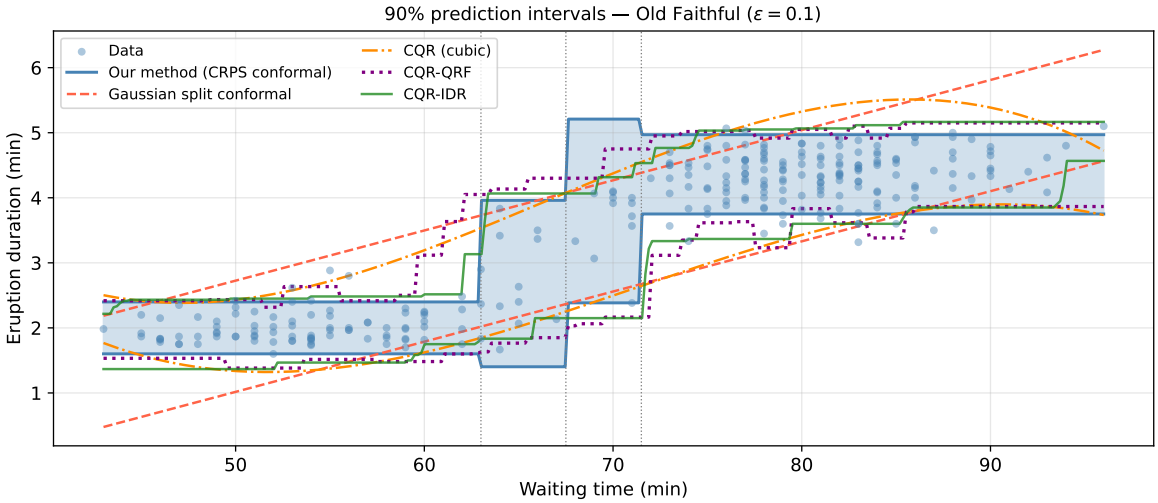

Figure 12: Old Faithful: 90% prediction intervals from five methods. Our CRPS-based conformal intervals adapt to the local conditional spread within each bin, producing narrower intervals than all competitors in both regimes. In the transition region around 67–70 minutes, the proposed method's interval is deliberately wide: the single bin spanning both eruption modes must cover the full bimodal spread, an honest consequence of the partition geometry.

respectively), and CQR-QRF adapts most flexibly (87.9 g). Our method on the same training half (italic row, $n_{\text{tr}} = 66$) yields mean width 100.6 g with sub-nominal coverage ($86.9 \pm 0.5\%$); at this small sample size, CQR-QRF achieves narrower intervals, reflecting the bias–variance trade-off discussed in Section 8.5: fewer observations per bin produce a coarser within-bin ECDF and wider intervals. When all n observations are used (bold row, below the line), our method achieves $90.3 \pm 0.2\%$ coverage with mean width 77.3 g, narrower than all competitors. All split-conformal methods exceed nominal coverage (92.5–93.1%).

Method	Coverage (%)	Mean width (g)
Our method $(n/2)$	86.9 ± 0.5	100.6 ± 1.3
Gaussian split conformal	92.5 ± 0.0	172.4 ± 1.0
CQR (cubic)	92.5 ± 0.0	134.1 ± 1.5
CQR-QRF	93.1 ± 0.1	87.9 ± 0.7
CQR-IDR	92.7 ± 0.0	127.6 ± 0.6
Our method (full n**)**	90.3 ± 0.2	77.3 ± 0.2

Table 4: Motorcycle accident ($n = 133$): empirical coverage and mean width of prediction intervals at nominal level $1 - \varepsilon = 0.90$, averaged over $R = 200$ random 50/50 splits ($\pm$ one standard error). *Top block*: all methods use the same training half $(n/2)$; *Our method* $(n/2)$ is directly comparable to the competitors. *Below the line*: **Our method (full** n**)** uses all n observations, a design advantage inherent to full-data conformal methods.

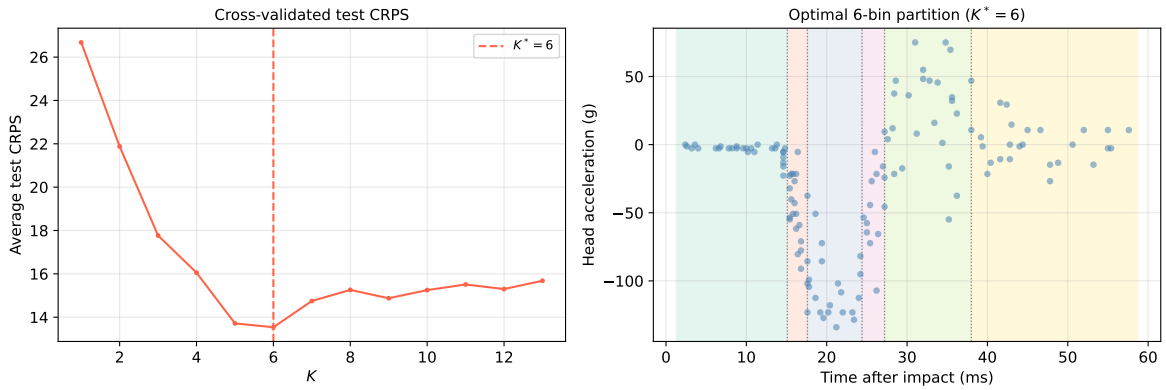

Figure 13: Motorcycle accident. *Left:* cross-validated test CRPS as a function of K with the selected $K^* = 6$. *Right:* scatter with the 6-bin partition; boundaries are concentrated in the high-variance impact phase.

11 Discussion

11.1 Other possible scoring rules

The DP requires only that the per-bin cost decompose additively; any leave-one-out proper scoring rule satisfies this. Among strictly proper rules for the full distribution, we are not aware of any other whose LOO sum reduces to a single scalar computable from pairwise differences: rules based on the log score or the Brier score evaluated on a fine grid would require $O(mn)$ or $O(m \log m)$ work per bin update rather than $O(\log n)$, compromising the tractability of the precomputation step. Moreover, CRPS serves double duty as the conformal nonconformity score (Section 8.2), so the same criterion governs both reference-class selection and the prediction sets formed from that reference class. Whether this coherence yields formal benefits, as in tighter coverage bounds or more efficient prediction sets relative to a mismatched pair of binning criterion and nonconformity score, is an open question we regard as a promising direction for further analysis.

11.2 Scoring against a held-out empirical distribution: the Cramér distance

The cross-validated criterion of Section 7.1 accumulates individual CRPS values, one per held-out observation. A natural variant is to form the empirical CDF $\hat{G}_n$ of all held-out observations in a bin and evaluate the predictive CDF F against it as a distributional object, using the *Cramér distance* [22, 21]

$$d_C(F, \hat{G}_n) = \int_{-\infty}^{\infty} \big(F(t) - \hat{G}_n(t)\big)^2 \, \mathrm{d}t.$$

When $\hat{G}_n = \delta_y$ (a single observation) this reduces to $\mathrm{CRPS}(F, y)$ [10], so the Cramér distance is the natural generalisation of CRPS to the case where the "observation" is itself a distribution.

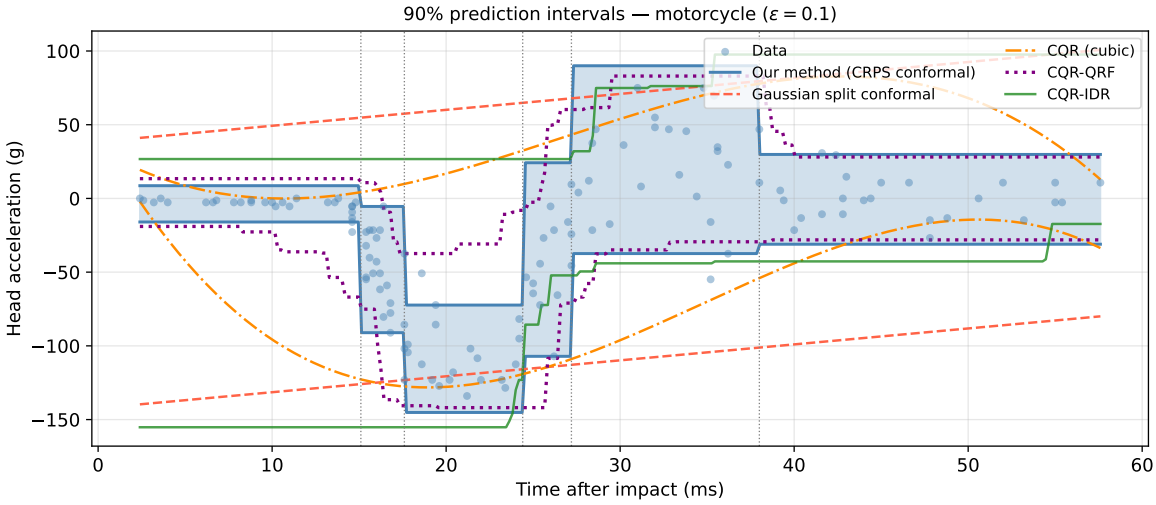

Figure 14: Motorcycle accident: 90% prediction intervals. Gaussian split conformal is constrained to constant width; our method, CQR-QRF, and CQR-IDR adapt to the non-stationary variance structure. A small number of data points in the high-variance impact phase (20–30 ms) fall outside the proposed method's intervals, consistent with the approximate within-bin exchangeability in this small and strongly heteroscedastic dataset.

Equivalence to average CRPS. Expanding the square and comparing term by term with the average CRPS over the held-out set $\{y_1, \ldots, y_n\}$ yields the identity

$$d_C(F, \hat{G}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{CRPS}(F, y_i) + \int_{-\infty}^{\infty} \hat{G}_n(t)\big[\hat{G}_n(t) - 1\big]\,\mathrm{d}t. \tag{3}$$

The second term on the right depends only on $\hat{G}_n$, not on F. Consequently, for any two candidate predictive distributions F and F',

$$d_C(F, \hat{G}_n) - d_C(F', \hat{G}_n) = \frac{1}{n}\sum_{i=1}^{n}\big[\mathrm{CRPS}(F, y_i) - \mathrm{CRPS}(F', y_i)\big].$$

The correction term cancels exactly in every pairwise comparison.

Implications. The two criteria are therefore identical for every purpose that matters in our setting: they share the same minimiser over F, the same ranking of candidate partitions, and the same variance for the model-selection statistic. The Cramér distance framing is nevertheless conceptually appealing: it makes explicit that the predictive distribution F is being judged against the full empirical distribution of held-out responses, rather than against individual outcomes summed post hoc. It also connects the cross-validation criterion to the classical Cramér two-sample test [21], which uses $d_C(\hat{F}_m, \hat{G}_n)$ as a statistic for testing $F = G$ and has power against all distributional alternatives.

12 Conclusion

We have presented a method for non-parametric conditional distribution estimation that combines three independently motivated ideas into a coherent pipeline: optimal bin-boundary placement by LOO-CRPS minimisation, cross-validated bin-count selection, and conformal prediction using the within-bin ECDF as both the predictive distribution and the nonconformity score.

The central technical contribution is the closed-form cost $\text{cost}(S) = mW/(m-1)^2$, which reduces the LOO-CRPS of any bin to a single pairwise-dispersion scalar and enables exact globally optimal partitioning in $O(n^2K)$ time via dynamic programming. The closed-form cost also explains why within-sample LOO-CRPS fails as a model-selection criterion: the $m/(m-1)^2$ prefactor makes size-2 bins cheap relative to their actual predictive contribution, so the DP exploits accidental y-homogeneity and the objective nearly monotonically decreases. The cross-validated TestCRPS criterion eliminates this optimism and yields a U-shaped model-selection curve that balances exchangeability against statistical efficiency.

On the predictive side, convexity of the CRPS nonconformity score guarantees contiguous prediction intervals in the split-conformal case; in our transductive setting, single-interval structure is observed in all experiments. The k-NN score provides a bandwidth-free alternative that yields non-contiguous highest-density regions for multimodal within-bin distributions. The two scores are therefore complementary: CRPS for efficiency in unimodal regimes, k-NN for distributional fidelity when the within-bin response is multimodal.

On the real datasets, the full-n method is competitive or superior to all conformalized competitors (Gaussian split conformal, CQR, CQR-QRF) in interval efficiency: 11–40% narrower intervals on Old Faithful (bimodal conditional) and 2.2× narrower than Gaussian split conformal on the motorcycle benchmark (strongly heteroscedastic), while maintaining near-nominal coverage in both cases. A matched-sample comparison (restricting our method to the same training half as the competitors) shows the cost of halving the data: on Old Faithful the method remains narrower than all competitors; on the motorcycle dataset ($n/2 = 66$, coverage $87.0 \pm 0.5\%$) CQR-QRF achieves modestly narrower intervals (87.9 vs 100.5 g), reflecting the small-sample bias-variance regime.

Limitations. The method requires a one-dimensional covariate and a contiguous binning structure. In small samples the CV criterion can select a large K, resulting in bins with few observations and hence coarse prediction intervals; this is a feature of the bias-variance tradeoff under limited data, not a failure of validity.

Multi-dimensional extension. The LOO-CRPS cost function is dimension-free: for any subset S of training points assigned to a bin, the cost is $W(S) \cdot m/(m-1)^2$ regardless of the dimension of x. The DP tractability, by contrast, is entirely one-dimensional: it relies on the total order induced by sorting x, which makes contiguous bins form a chain and gives the optimal-substructure property. Two natural extensions exist for $d > 1$, each with a distinct drawback.

The first is projection onto a learned one-dimensional index: find a function $g : \mathbb{R}^d \to \mathbb{R}$, sort by $g(x)$, and run the existing DP. The partition is globally optimal *given* g, and the DP cost is unchanged. The difficulty is choosing g: alternating optimisation (fix g, optimise

partition; fix partition, improve g) is tractable, but global optimality in g is not guaranteed. An interesting open question is whether the data-adaptive g that minimises LOO-CRPS recovers a form of sufficient dimension reduction for the conditional *distribution* $P(Y \mid X = x)$ rather than merely the conditional mean.

The second is CART-style greedy binary splitting: at each step choose the axis and threshold that most reduces total LOO-CRPS, then recurse. This requires $O(dn)$ evaluations per split and $K - 1$ splits total, giving $O(Kdn)$ complexity with no projection required and interpretable rectangular regions. As in the previous case, global optimality is not guaranteed: the greedy splits are locally optimal but may not yield the best overall partition.

Open problems. At least three questions raised by this work remain unresolved. First, whether the coherence between the LOO-CRPS binning criterion and the CRPS nonconformity score can be proved formally to yield benefits, namely tighter finite-sample bounds or more efficient prediction sets, relative to a mismatched pair. Second, the monotone-coverage question: as K increases (bins narrow), conditional coverage at a fixed x^* should improve; making this monotonicity precise would require bounding the exchangeability violation as a function of bin width and the local smoothness of the Data Generating Process (DGP). Third, the consistency of K^*: does $K^* = \arg\min_K \text{TestCRPS}(K)$ converge to a meaningful oracle value as $n \to \infty$, and under what conditions on the DGP does the population $\text{TestCRPS}(K)$ have a unique minimum? A population-level analysis would require characterising the limiting test CRPS as a functional of the DGP and the partition geometry — an open question even for simpler scoring rules.

Acknowledgements

The author acknowledges Prof. Alexander Gammerman for his helpful guidance and Matteo Fontana for providing pointers to relevant prior work. The author used Claude (Anthropic) as a writing and coding assistant during the preparation of this manuscript. Claude assisted with drafting and editing prose, implementing experiment scripts, and making LaTeX edits. The method, proofs, experimental design, and all intellectual content are the author's own.

Software and Reproducibility

The method is implemented in the `crpsconfreg` Python package, available at https://pypi.org/project/crpsconfreg. All figures and numerical results in this paper can be reproduced using the scripts in the accompanying repository at https://github.com/ptocca/crps-conformal-regression. An interactive browser demo (no installation required) is hosted at https://ptocca.github.io/crps-conformal-regression/.

References

[1] Brown, T. A. (1974). Admissible scoring systems for continuous distributions. *RAND Corporation Memorandum*, P-5235.

[2] Matheson, J. E. & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, **22**(10), 1087–1096.

[3] Unger, D. A. (1985). A method to estimate the continuous ranked probability score. *Preprints, 9th Conference on Probability and Statistics in Atmospheric Sciences*, Virginia Beach, VA, American Meteorological Society, 206–213.

[4] Bouttier, F. (1994). Sur la prévision probabiliste et sa vérification. *Note de Centre CNRM*, No. 21, Météo France, Toulouse.

[5] Fenwick, P. M. (1994). A new data structure for cumulative frequency tables. *Software: Practice and Experience*, **24**(3), 327–336.

[6] Knuth, D. E. (1971). Optimum binary search trees. *Acta Informatica*, **1**(1), 14–25.

[7] Yao, F. F. (1980). Efficient dynamic programming using quadrangle inequalities. In *Proceedings of the 12th Annual ACM Symposium on Theory of Computing (STOC)*, 429–435.

[8] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, **81**(394), 461–478.

[9] Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed., Section 7.4. Springer.

[10] Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**(477), 359–378.

[11] Devroye, L. & Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley. (Chapter 5 covers the exponential limit of nearest-neighbour distances.)

[12] Vovk, V., Gammerman, A. & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

[13] Papadopoulos, H., Proedrou, K., Vovk, V. & Gammerman, A. (2002). Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning (ECML)*, pp. 345–356. Springer.

[14] Romano, Y., Patterson, E. & Candès, E. J. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, vol. 32.

[15] Vovk, V. & Petej, I. (2014). Venn–Abers predictors. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 829–838.

[16] Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, **7**, 983–999.

[17] Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A. Y. & Schuler, A. (2020). NGBoost: Natural gradient boosting for probabilistic prediction. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 2690–2700.

[18] Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**(1), 33–50.

[19] Auger, I. E. & Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, **51**(1), 39–54.

[20] Killick, R., Fearnhead, P. & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107**(500), 1590–1598.

[21] Baringhaus, L. & Franz, C. (2004). On a new multivariate two-sample test. *Journal of Multivariate Analysis*, **88**(1), 190–206.

[22] Székely, G. J. & Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, **143**(8), 1249–1272.

[23] Barber, R. F., Candès, E. J., Ramdas, A. & Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*, **49**(1), 486–507.

[24] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, **113**(523), 1094–1111.

[25] Sesia, M. & Romano, Y. (2021). Conformal prediction using conditional histograms. In *Advances in Neural Information Processing Systems*, vol. 34.

[26] Allen, S., Gavrilopoulos, G., Henzi, A., Kleger, G.-R. & Ziegel, J. F. (2025). In-sample calibration yields conformal calibration guarantees. *arXiv preprint*, 2503.03841.

[27] Chernozhukov, V., Wüthrich, K. & Zhu, Y. (2021). Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, **118**(48), e2107794118.

[28] Randahl, D., Williams, J. P. & Hegre, H. (2026). Bin-conditional conformal prediction of fatalities from armed conflict. *Political Analysis*, **34**(1), 96–108.

[29] Henzi, A., Ziegel, J. F. & Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B*, **83**(5), 963–993.