CRPS-Optimal Conformal Regression

Non-parametric Conditional Distribution Estimation via Optimal Binning and Conformal Prediction

Paolo Toccaceli

Centre for Reliable Machine Learning
Royal Holloway, University of London

Outline

The uncertainty problem

- Standard supervised learning: given $(x_1, y_1), \ldots, (x_n, y_n)$, predict y for new x.
- Most algorithms output a **bare prediction** $\hat{y}$ — no quantification of uncertainty.
- Bare predictions are often insufficient for real decisions.

Two complementary goals

- **Coverage**: the true label should fall inside a prediction region with chosen probability.
- **Efficiency**: prediction regions should be as small (narrow) as possible.

- Statistical learning theory bounds are too loose for practical training set sizes.
- Hold-out estimates work but require a separate calibration set and give no formal guarantee.

Conformal Prediction: the framework

- **Conformal Predictors** (Vovk, Gammerman, Shafer 2005) are a framework, not a single algorithm.
- They wrap *any* underlying ML method and output a **prediction set** Γ^ε at a chosen significance level ε.

Validity guarantee (i.i.d. assumption only)

$$\mathbb{P}(y^* \notin \Gamma^\varepsilon) \leq \varepsilon$$

Valid at every chosen ε; does not depend on the algorithm.

- The only assumption: training and test data are **i.i.d.**
- **Efficiency** (small prediction sets) depends on the underlying algorithm; validity does not.
- CP frees us to focus entirely on improving efficiency.

How Conformal Prediction works

Key idea: measure how *strange* a hypothetical label is, compared to the training set.

- The **nonconformity measure** $\alpha_i = \mathcal{A}(\{z_1, \ldots, z_{n+1}\} \setminus z_i,\, z_i)$ quantifies how anomalous example i is.
- For a test object x^* and candidate label y_h, form the hypothetical example (x^*, y_h) and compute:

$$p(y_h) = \frac{\#\{i : \alpha_i \geq \alpha_{n+1}(y_h)\}}{n+1}$$

- This is a **conformal p-value**: not a posterior probability, but a rank-based exchangeability test.
- The prediction set at level ε is:

$$\Gamma^{\varepsilon} = \{y_h \in \mathcal{Y} : p(y_h) > \varepsilon\}$$

Inductive CP (ICP): split the data into a *proper training set* and a *calibration set*; train once, compute α_i on the calibration set. Scalable; same validity guarantee.

From prediction sets to predictive distributions

- A **predictive distribution** $F(y, x)$ estimates $\mathbb{P}[Y \leq y \mid X = x]$.
- Unlike a prediction set, it encodes the full conditional distribution, not just one coverage level.
- We seek F that is:
  1. **Valid** (calibrated): $F(y_i, x_i)$ are approximately $U(0, 1)$.
  2. **Specific** (sharp): prediction intervals are as narrow as possible.
- Conformal Predictive Distributions (CPDs, Vovk et al. 2017) achieve validity under i.i.d. assumption alone — *no prior, no parametric assumption*.

Evaluating predictive distributions

PIT (Probability Integral Transform): $F_i(y_i)$ should be $\sim U(0, 1)$ — checks validity.

CRPS (Continuous Ranked Probability Score): measures sharpness given validity — lower is better:

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} \big(F(t) - \mathbf{1}[t \geq y]\big)^2 dt$$

The Reference Class Problem

- John Venn (1866): *"Every individual thing has an indefinite number of properties observable in it, and might be considered as belonging to an indefinite number of different classes of things."*
- **Which class** do we use when estimating probabilities for a test object?
- The more specific the class, the better the statistics — but the fewer examples it contains.

Venn Predictors: the key idea

- Partition the training examples into **categories** using an ML method.
- For a test object x^*, use the empirical label distribution of its category as the predicted distribution.
- The partition is found automatically from the data — the reference class problem is solved by optimisation.

Venn Predictors: formal definition

- For each candidate label y_h, form the hypothetical example (x^*, y_h).
- Assign (x^*, y_h) to a category T (using a taxonomy function, e.g. nearest-neighbour label).
- Output the **empirical distribution of y over category T**, including (x^*, y_h) itself.
- Repeat for each y_h — the output is a *multi-probabilistic* prediction (a set of distributions).

Calibration guarantee

For an appropriate choice P from the multi-set of outputs,

$$\mathbb{P}[Y = y \mid P_y = p] = p \quad \text{(long-run calibration)}$$

Validity holds under the i.i.d. assumption alone.

Venn-ABERS (Vovk & Petej 2014): the isotonic-regression implementation for binary classification — the categories are the piecewise-constant regions of the isotonic calibrator.

Setting and goal

- Training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^2$; single covariate, continuous response.
- Sort by x: $x_{(1)} \leq \cdots \leq x_{(n)}$, write y_i for the paired response.
- A K-**partition** is a sequence of cut points $0 = b_0 < b_1 < \cdots < b_K = n$.
- For a test point $x^* \in [x_{(b_{k-1}+1)}, x_{(b_k)}]$, bin B_k provides the predictive distribution via its within-bin empirical CDF.

Core question

How should bin boundaries be placed, and how many bins K should be used, to minimise prediction error — without any parametric assumption on $P(Y|X = x)$?

Answer: minimise total **leave-one-out CRPS** over the partition, via dynamic programming.

The CRPS as bin cost

- For predictive CDF F and scalar outcome y:

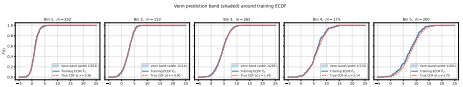
$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} \big(F(t) - \mathbf{1}[t \geq y]\big)^2 dt$$

  Geometrically: integrated squared gap between F and the step at y.

- Equivalent energy-score form (Gneiting & Raftery 2007):

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

- CRPS is **strictly proper**: uniquely minimised when $F = P$ (the true distribution).
- For an empirical CDF $\hat{F}_m$ with m atoms, CRPS is large when the forecast is mis-centred or



LOO-CRPS cost of a bin: closed form

- For bin S with responses $y_1, \ldots, y_m$, define:

$$W = \sum_{\ell<r} |y_\ell - y_r|, \quad D = 2W.$$

- The leave-one-out predictive distribution for observation k is the ECDF of $\{y_\ell : \ell \neq k\}$.

Proposition 1 (closed-form cost)

The total leave-one-out CRPS of bin S is:

$$\mathrm{cost}(S) = \sum_{k \in S} \mathrm{CRPS}(\hat{F}_{S\setminus\{k\}}, y_k) = \frac{mW}{(m-1)^2}$$

where $W = \sum_{\ell<r} |y_\ell - y_r|$ is the pairwise dispersion.

- For a bin spanning sorted indices i through j (size $m = j - i + 1$):

$$c(i, j) = \frac{j - i + 1}{} W(i, j)$$

Dynamic Programme

Define $\mathrm{dp}[k][j]$ = minimum total LOO-CRPS for partitioning observations $1, \ldots, j$ into exactly k bins.

Base case:

$$\mathrm{dp}[1][j] = c(1, j), \quad j = 2, \ldots, n$$

Recurrence (for $k \geq 2$, $j \geq k$):

$$\mathrm{dp}[k][j] = \min_{k-1 \leq i < j} \left\{ \mathrm{dp}[k-1][i] + c(i+1, j) \right\}$$

The last bin covers $i+1$ to j; the first i observations are split optimally into $k-1$ bins.

Solution: $\mathrm{dp}[K][n]$; boundaries recovered by backtracking.

Optimal substructure (Prop. 3)

Any prefix of a globally optimal K-partition is itself optimal for its size. This guarantees the DP finds the *exact* global optimum — unlike greedy splitting.

Complexity:

- Precompute cost matrix: $O(n^2 \log n)$ using Fenwick trees.
- Fill DP table: $O(n^2 K)$ time, $O(n^2)$ space.

Efficient precomputation with Fenwick trees

Goal: precompute $c(i, j)$ for all $1 \le i \le j \le n$ efficiently.

- Fix i; scan $j = i, i+1, \ldots, n$. Adding y_{j+1} updates $W(i, j)$ via:

$$\sum_{\ell=i}^{j} |y_\ell - y_{j+1}| = y_{j+1} \cdot r - S_{\le} + S_{>} - y_{j+1} \cdot (m - r)$$

  where $r = \#\{y_\ell \le y_{j+1}\}$, $S_{\le} = \sum_{\ell: y_\ell \le y_{j+1}} y_\ell$, $S_{>} = \sum_{\ell: y_\ell > y_{j+1}} y_\ell$.
- These are **prefix-count** and **prefix-sum** queries over a dynamically growing set.
- A **Fenwick tree** (Binary Indexed Tree, Fenwick 1994) answers each query in $O(\log n)$.
- Two trees suffice: one for counts, one for values.
- Total: $O(n^2 \log n)$ time, $O(n^2)$ space.

Potential speedup via Knuth–Yao theorem

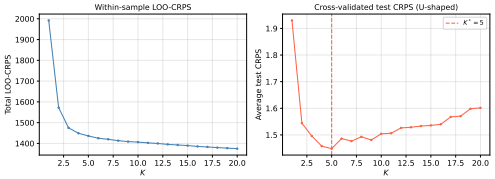
If $c(i, j)$ satisfies the *quadrangle inequality*, the optimal split point $\mathrm{opt}_k(j)$ is non-decreasing in j, enabling divide-and-conquer: $O(nK)$ DP. Whether $c(i, j)$ satisfies QI is an open problem.

Selecting K: avoiding in-sample optimism

- The within-sample LOO-CRPS $\mathrm{dp}[K][n]$ is **biased**: the partition and its LOO evaluation are optimised jointly on the same data.
- Result: the in-sample criterion decreases monotonically in K (no useful minimum).
- The $m/(m-1)^2$ prefactor makes size-2 bins artificially cheap, driving over-partitioning.

Cross-validated K selection (Sec. 7.1):

1. Split sorted data into alternating halves $\mathcal{T}$ and $\mathcal{V}$ (preserving x-order).
2. For each K: fit optimal partition on $\mathcal{T}$, evaluate CRPS on $\mathcal{V}$.
3. Select $K^* = \arg\min_K \mathrm{TestCRPS}(K)$.



Left: in-sample LOO-CRPS (nearly monotone). Right: cross-validated test CRPS (U-shaped, clear minimum at $K^* = 5$).

Predictive output I: the Venn prediction band

Once K^* is selected and the partition is re-fitted on all data, each test point x^* falls in bin B_k with m training responses $y_1, \ldots, y_m$.

- **Venn prediction** (Vovk et al.): the family of augmented ECDFs indexed by hypothetical label $y_h \in \mathbb{R}$:

$$F_{y_h}(t) = \frac{\#\{i : y_i \leq t\} + \mathbf{1}[y_h \leq t]}{m+1}$$

- This is a *band* of width $1/(m+1)$ around the training ECDF $\hat{F}_m$:

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

- By exchangeability, the true F_{y^*} lies inside the band — a valid multi-probabilistic prediction.
- **Limitation**: band width $1/(m+1)$ carries no information about local density or position of y^* within the bin.

Predictive output II: CRPS conformal prediction set

Nonconformity score: for candidate y_h,

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

where W is fixed (independent of y_h). Scores for training points use LOO-ECDFs.

Proposition 2 (conformal coverage)

Under exchangeability of $(y_1, \ldots, y_m, y^*)$:

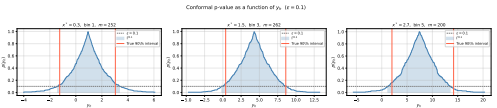
$$\mathbb{P}(y^* \in \Gamma^\varepsilon) \geq 1 - \varepsilon, \quad \text{where } \Gamma^\varepsilon = \{y_h : p(y_h) > \varepsilon\}$$

- **Why CRPS?** It is *strictly proper*, so $\alpha(y_h)$ is large precisely when y_h is surprising under the within-bin distribution — not just far from the mean.
- The LOO structure ensures the training and test scores are computed symmetrically.
- **Coherence**: the DP selects bins by minimising LOO-CRPS; the conformal step evaluates using the same criterion.

Structure of Γ^{ε}: always a single interval

- $\alpha(y_h) = \frac{1}{m}\sum_i |y_i - y_h| - \frac{W}{m^2}$ is **convex** and piecewise linear in y_h, with unique minimum at the empirical median.
- Consequence: $\{y_h : \alpha(y_h) \leq c\}$ is always a **connected interval**.
- Γ^{ε} is a connected interval centred near the median, with width shrinking as $m \to \infty$.

Bimodal caveat: for bins with bimodal within-bin distribution, Γ^{ε} spans both modes and the inter-modal gap — valid but informationally wasteful.



Conformal p-value $p(y_h)$ as a function of y_h at three test points. The shaded region is $\Gamma^{0.1}$.

Non-convex extension: the k-NN score

For bimodal within-bin distributions: the CRPS score spans the inter-modal gap; the 1-NN score resolves this.

- Define $\alpha^{(1)}(y_h) = d_{(1)}(y_h, \{y_1, \ldots, y_m\})$ (distance to nearest training point).
- Local minima near each mode; large in the inter-modal gap.
- $\Gamma^{\varepsilon,(1)}$ decomposes into **two disjoint intervals**, one around each mode.
- The effective resolution is data-driven; no bandwidth needed.
- Result: a non-parametric highest-density region (HDR) of the within-bin distribution.

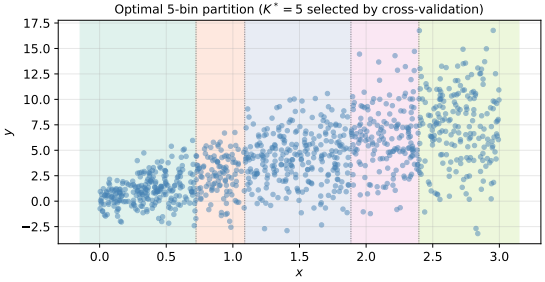
Score	Coverage	Mean set size
CRPS	$92.6 \pm 0.2\%$	7.51 ± 0.01
1-NN	$91.0 \pm 0.3\%$	4.09 ± 0.03

Bimodal bin ($m = 50$, $\varepsilon = 0.10$, $R = 500$ seeds).
Both achieve super-uniform coverage; 1-NN set is $1.84\times$ smaller.

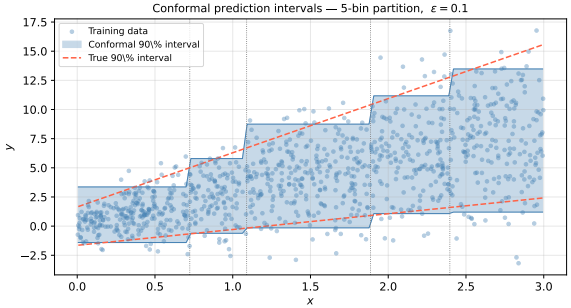
The DP uses LOO-CRPS regardless: CRPS measures predictive accuracy of the ECDF, not local density.

Numerical illustration: heteroscedastic synthetic data

Data: $n = 1000$, $X_i \sim \mathrm{Unif}(0, 3)$, $Y_i | X_i = x \sim \mathcal{N}(3x, (1+x)^2)$. Conditional variance grows $16\times$ over $[0, 3]$.

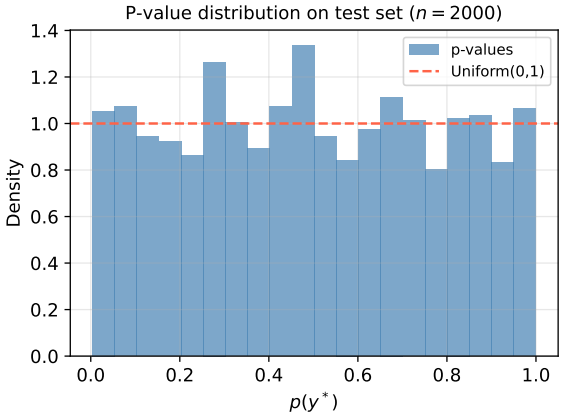


Cross-validation selects $K^* = 5$ bins. The narrow bin B_2 ($\Delta x = 0.37$) isolates the high-slope transition region.



Conformal 90% intervals (blue) vs. true 90% intervals (dashed red). Widths grow monotonically with x, tracking the increasing variance.

Numerical illustration: results summary



P-value distribution on 2000-point test set.
Near-uniform, consistent with the super-uniform

- Empirical coverage at $\varepsilon = 0.10$: **89.8%** (within ± 0.7 pp of nominal).
- Bin boundaries adapt to the jointly growing mean and variance.
- Conformal intervals track the oracle widths closely; slight conservatism at $x = 0.3$ and $x = 1.5$ reflects residual within-bin heterogeneity.

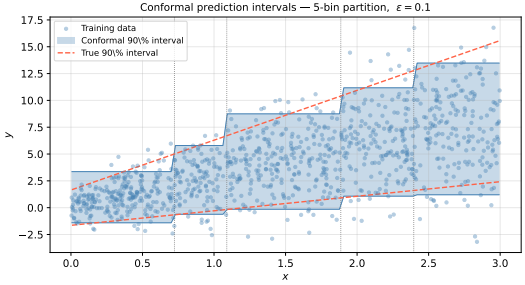
Key takeaway

CV selects the best bias–variance trade-off automatically. The within-sample criterion would have driven $K \to K_{\max}$.

Real-data experiments: Old Faithful (bimodal)

Dataset: $n = 272$; $x =$ waiting time (min), $y =$ eruption duration (min).
Marginal distribution of y is **bimodal**; mixture proportion shifts with x.

- CV selects $K^* = 2$; boundary at $x = 67.5$ min separates short- and long-eruption regimes.
- Each bin is unimodal — the partition captures the regime transition.
- 90% conformal intervals adapt: narrower inside each regime, wider only in the transition region.



Method	Cov. (%)	Width (min)
Ours (full n)	**90.3±0.1**	**1.200±0.001**
Ours ($n/2$)	*88.3±0.3*	*1.262±0.010*
Gaussian split conf.	91.2±0.0	1.683±0.006

Full-n method is **11–40% narrower** than all split-conformal competitors.

Real-data experiments: motorcycle accident (heteroscedastic)

Dataset: $n = 133$; x = time after impact (ms), y = head acceleration (g).
Variance near-zero before 15 ms, explosive in 15–30 ms, moderating thereafter.

- CV selects $K^* = 10$; boundaries concentrated in the high-variance impact phase.
- Narrow bins in 15–30 ms window improve within-bin exchangeability at the cost of fewer ECDF atoms.

Method	Cov. (%)	Width (g)
Ours (full n)	**91.0±0.2**	**78.9±0.3**
Ours ($n/2$)	*87.0±0.5*	*100.5±1.2*
Gaussian split conf.	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

Averaged over $B = 200$ random 50/50 splits.

- Full-n method is **2.2× narrower** than Gaussian split conformal.
- At matched sample size ($n/2 = 66$): CQR-QRF slightly narrower, reflecting bias–variance cost of halving the data.
- All methods achieve near-nominal coverage.

Key takeaway

Full-data conformal with the within-bin ECDF as both fitting and calibration object eliminates the fitting/calibration split cost.

Conclusion

Three independently motivated ideas form a coherent pipeline

1. **Optimal bin placement**: minimise total LOO-CRPS via DP — globally optimal, $O(n^2K)$.
2. **Cross-validated K selection**: alternating-split TestCRPS gives a genuine U-shaped minimum.
3. **Conformal prediction**: CRPS nonconformity score yields finite-sample coverage; coherent with the binning objective.

Key properties

- Distribution-free validity under i.i.d.
- No parametric model for $P(Y|X = x)$
- Γ^ε is always a connected interval (CRPS score) or a HDR (1-NN score)
- Full-n design: all data used for both

Open problems

- Does $c(i, j)$ satisfy the quadrangle inequality? ($O(nK)$ DP speedup)
- Does coherence between LOO-CRPS binning and CRPS conformal scoring yield formally tighter bounds?