On training locally adaptive CP

Nicolo Colombo   NICOLO.COLOMBO@RHUL.AC.UK
Royal Holloway, University of London, Egham, Surrey, UK

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

We address the problem of making Conformal Prediction (CP) intervals locally adaptive. Most existing methods focus on approximating the object-conditional validity of the intervals by partitioning or re-weighting the calibration set. Our strategy is new and conceptually different. Instead of re-weighting the calibration data, we redefine the conformity measure through a trainable change of variables, $A \to \phi_X(A)$, that depends explicitly on the object attributes, X. Under certain conditions and if ϕ_X is monotonic in A for any X, the transformations produce prediction intervals that are guaranteed to be marginally valid and have X-dependent sizes. We describe how to parameterize and train ϕ_X to maximize the interval efficiency. Contrary to other CP-aware training methods, the objective function is smooth and can be minimized through standard gradient methods without approximations.

Keywords: Conformal prediction, local adaptivity, conditional validity, conformity score, regression.

1. Introduction

Two features of Conformal Prediction (CP) contribute to their increasing popularity, finite-sample validity and straightforward applicability. Unlike the Bayesian approach, CP algorithms are computationally cheap, provide practical non-asymptotic guarantees, and can augment any pre-existing or pre-trained point-prediction model. The very reasons for their success may hide a few structural limitations. On the one hand, finite-sample validity is hard to achieve on heteroskedastic data, as the prediction intervals should have attribute-dependent sizes. On the other hand, defining the same (fixed) CP algorithm for any point-prediction model may be a sub-optimal strategy for certain types of tasks or data. Lastly, it is unlikely that machine learning algorithms trained to generate good point-like predictions are automatically CP-efficient without being retrained.

We introduce a new CP-aware fine-tuning scheme to make CP algorithms more efficient and locally adaptive. Compared to analogous strategies, our approach does not break data exchangeability. This implies the obtained locally-adaptive prediction intervals are marginally valid, as the non-adaptive ones, but have attribute-dependent sizes. Moreover, since the scheme does not modify the underlying point-prediction model, it can be merged with other localization methods. The idea is to boost the local adaptivity of the prediction intervals by optimizing a set of learnable transformations of the conformity scores. The trained transformations are guaranteed to produce marginally valid prediction sets by construction. The prediction sets become locally adaptive when we map them back to the label space by inverting the trained transformations.

The scheme generalizes the Error Re-weighted Conformal approach (ERC) of Papadopoulos et al. (2008, 2011), in which the conformity scores are re-weighted by a pre-fitted model of the residuals. More precisely, let f be a pre-trained point-prediction algorithm, $f(X) \approx \mathrm{E}_{Y|X}(Y)$, A the standard conformity score, $A = (f(X) - Y)^2$, and $(1 - \alpha) \in [0, 1]$ a user-defined confidence level. ERC can be viewed as a coordinate transformation $A \to \phi_X(A) = \frac{A}{g(X)^2 + \gamma}$, where g is a model of the conditional prediction error, i.e. $g(X) \approx \mathrm{E}_{Y|X}(f(X) - Y)$, and $\gamma > 0$. The boundaries of the prediction intervals are the solutions of $\phi_X(A) = \frac{(Y - f(X))^2}{g(X)^2 + \gamma} = \hat{q}$, where $\hat{q}$ is the $(1 - \alpha)$-quantile of the conformity scores distribution, estimated on the calibration set. By definition of empirical quantile, the prediction interval at X, $C = \{y : \phi_X(A(X, y)) \leq \hat{q}\}$, is marginally valid. The size of C is $|C| = \sqrt{\hat{q}(g(X)^2 + \gamma)}$ and explicitly depends on X, which means the intervals grow and shrink over the attribute space.

We extend the ERC idea in two ways, [1]

- we define a general parametric class of coordinate transformations, Φ_θ, that are guaranteed to produce marginally valid and locally adaptive prediction intervals (Sections 2.3, 2.4 and 2.6) and

- we show how to find a Φ_θ that maximizes the efficiency of the intervals (Sections 2.7 and 2.8).

To guarantee the validity of the transformed prediction sets and their interpretability as prediction intervals in the label space, we require the attribute-dependent transformations, $\phi_X \in \Phi_\theta$, to be monotonic in A and have the same codomain for all X (Assumption 1). Validity and local adaptability are proven in Theorem 1 by adapting a standard proof of CP marginal validity and using the assumed invertibility of ϕ_X. To train Φ_θ, we introduce a confidence-specific cost function that measures the non-efficiency of the intervals at a given confidence level (ℓ_α in (21)). The confidence-specific cost function contains non-differentiable terms but becomes smooth and tractable when it is averaged over all possible confidence levels and estimated empirically on a finite-size data set (ℓ_{all} in (22)). This makes the learning task a smooth and unconstrained optimization problem, which can be solved with gradient approaches even if the inverse map, ϕ_X^{-1}, is not available analytically. In the experiments, we compare four classes of attribute-dependent transformations with the standard non-adaptive CP algorithm and ERC.

1.1. Related work

Redefining the conformity functions to include an attribute-dependent factor is not new. The first example of this technique is the ERC method of Papadopoulos et al. (2008, 2011) (see also Section 5 of Romano et al. (2019)). Our transformation of the conformity score is more general and is not pre-trained by fitting the residuals.

A growing stream of works addresses the problem of producing prediction intervals that are approximately object-conditional valid (whit finite-size data). Approximate conditional validity is obtained through an attribute-dependent re-weighting of the calibration samples.

---

1. For simplicity, we focus on regression and make standard i.i.d. assumptions, but the approach generalizes with minor changes to classification problems and exchangeable data.

Data exchangeability is temporarily broken and needs to be restored at the end. The idea of Lei and Wasserman (2014) and Vovk (2012) has been refined in the last few years by establishing links to Kernel Density Estimation and the covariate shift problem (Barber et al. (2022); Han et al. (2022); Guan (2023)). Here, we follow a conceptually different path, where data exchangeability is preserved at all stages.

In the Conformalized Quatile Regression model of Romano et al. (2019), the standard conformity function is replaced with a proxy of the pinball loss. Given $(1 - \alpha) \in [0, 1]$, the optimum of the pinball loss is an estimate of the conditional $(1 - \alpha)$-quantile function. This implies the method can not quantify the uncertainty of a pre-trained point-prediction model and needs to be retrained for each $\alpha \in [0, 1]$. See Sesia and Candès (2020) for a review of conformal quantile regression methods.

Evaluation scores similar to ℓ_α defined in (21) are popular in the CP literature. Recent works exploit objective functions similar to (21) to train the underlying prediction models in a CP-aware way. Here, we fix the underlying prediction model and learn a new conformity function, $\phi_X(A)$. We also average ℓ_α over $\alpha \in [0, 1]$, which makes the optimization problem smooth. This is a significant improvement compared to most CP-aware objectives used in the past, which often require smooth relaxations that may be hard to optimize (Bellotti (2020); Stutz et al. (2021); Einbinder et al. (2022)).

A recent work, Einbinder et al. (2022), is particularly close to ours. The authors propose to train a point-prediction model by forcing the distribution of conformity scores to be uniform on $[0, 1]$ for all X. The definition of the conformity scores does not change. It would be interesting to investigate whether tuning the underlying model or the conformity score is equivalent, theoretically and practically. Also, while our approach extends straightforwardly to classification tasks (provided the distribution of the conformity scores is continuous), it may be hard to adapt Einbinder et al. (2022) to the regression setup. As for Conformalized Quantile Regression, the CP algorithm of Einbinder et al. (2022) requires fine-tuning the underlying model and cannot quantify the uncertainty of a pre-trained non-conformal classifier.

Using the Inverse Function Theorem to compute the gradient of implicit functions is a common technique for minimizing objective functions without analytical form (Krantz and Parks (2002)). The idea has become popular in the domain of implicit layers optimization (David Duvenaud and Johnson (2020)).

1.2. Contribution

To the best of our knowledge, this is the first time locally adaptive and efficient conformal prediction intervals are obtained from a data-driven definition of the conformity scores. Previous works on training CP, e.g. Bellotti (2020), Stutz et al. (2021), or Einbinder et al. (2022), also aim to optimize the efficiency of the intervals but focus on tuning the underlying point-predictor instead of the CP construction on top of it. See Section 1.1 for a brief discussion of those works.

Proposing a differentiable CP-aware objective function is also relevant. The averaged objective function presented here may help define a differentiable version of Bellotti (2020), Stutz et al. (2021), or Einbinder et al. (2022).

As Papadopoulos et al. (2008), our approach is orthogonal to methods that fine-tune the underlying point-prediction model, e.g. Romano et al. (2019), and is compatible with localization approaches based on the selection/re-weighting of the calibration objects (Lei and Wasserman (2014); Izbicki et al. (2019); Romano et al. (2020); Foygel Barber et al. (2021); Sesia and Romano (2021); Guan (2023)). Merging different localization techniques, as in Guan (2023), may boost the usability of CP on specific tasks.

Finally, we believe this work is one of the first to exploit, at the same time, the invariance of CP intervals under monotone transformations of the conformity scores and recent ideas on conformity-aware training.

2. Methods

2.1. A regression task

Let $\{D, (X_{test}, Y_{test})\} = \{(X_n, Y_n) \in \mathcal{X} \times \mathbb{R}\}_{n=1}^{N} \cup (X_{test}, Y_{test})$ be a collection of i.i.d. random variables following an unknown joint distribution, P_{XY}, $f : \mathcal{X} \to \mathbb{R}$ a pre-trained point-prediction model that approximates the object-conditional mean of Y, i.e. $f(X) \approx \mathrm{E}_{Y|X}(Y)$, and

$$\mathcal{D} = \{(x_n, y_n) \in \mathcal{X} \times \mathbb{R}\}_{n=1}^{N+1} \tag{1}$$

a sample of $\{D, (X_{test}, Y_{test})\}$. In what follows, we use $\mathcal{D}$ to train a set of conformity score transformations, Φ_θ, through a Leave-One-Out strategy where N samples are interpreted as a calibration set and the remaining sample as a test object. The resulting CP algorithm will be tested on unseen data, i.e. a new sample of $\{D, (X_{test}, Y_{test})\}$.

2.2. Prediction intervals

Given f and $\{D, (X_{test}, Y_{test})\}$, CP algorithms use a conformity function, $a = a(f(X), Y)$, to evaluate the conformity of $f(X_{test})$ with the predictions made on the calibration set, $\{f(X_n)\}_{n=1}^{N}$. Let $(1 - \alpha) \in [0, 1]$ be a user-defined confidence level. The corresponding prediction set

$$C = \{y : a(f(X_{test}, y)) \le \hat{q}\}, \quad \hat{q} \text{ s.t. } \left|\{a(f(X_n), Y_n) \le \hat{q}\}_{n=1}^{N}\right| = \lceil (N+1)(1-\alpha) \rceil \tag{2}$$

where $|\mathcal{S}|$ is the cardinality of $\mathcal{S}$, is said to be marginally valid with confidence $(1 - \alpha)$ because

$$\mathrm{Prob}\,(Y_{test} \in C) \ge 1 - \alpha \tag{3}$$

Lower bounds for $\mathrm{Prob}\,(Y_{test} \in C)$ can also be obtained as a function of N. When a is monotonic in $|f(X) - Y|$, C is a symmetric interval centred on $f(X_{test})$. The size of C can be associated with the reliability of $f(X_{test})$. For example, if $a(f(X), Y) = (f(X) - Y)^2$, the prediction interval can be written explicitly as $C = [f(X_{test}) - \Delta, f(X_{test}) + \Delta]$, where $\Delta = 2a^{-1}(\hat{q}) = 2\sqrt{\hat{q}}$ and a^{-1} is defined by $S = a^{-1}(a(S))$. Other possible choices for a include $a_{abs} = |f(X) - Y|$ or $a_{log} = \log((f(X) - Y)^2)$. Here, we always let $a = (f(X) - Y)^2$. Some advantages of using a_{log} are outlined in Section 2.5.

4

2.3. Coordinate transformations

CP intervals are invariant under composition with a monotone function, e.g. if $a \to a_{log} = \log \sqrt{a}$ or $a \to a_{abs} = \sqrt{a}$. This happens because the intervals only depend on the ranking of the conformity scores, which is preserved if the conformity scores are transformed monotonically. This work is a generalization of this idea. Let $\Phi_\theta = \{\phi_X : \mathbb{R}_+ \to \mathcal{B}_X, X \in \mathcal{X}\}$ be a set of attribute-dependent parametric functions of a base conformity scores, $A = a(f(X), Y) \in \mathbb{R}_+$. The task is to find a Φ_θ that maximizes the efficiency of a CP algorithm based on the conformity scores transformed by $\phi_X \in \Phi_\theta$. To avoid overfitting, all $\phi_X \in \Phi_\theta$ should share a smooth functional dependence on X, e.g. we may require $\phi_X(A) \sim \phi_{X'}(A)$ if $X \sim X'$ for all $A \in \mathbb{R}_+$. Some assumptions are needed to guarantee that $B = \phi_X(A)$ is a well-defined conformity score. If so, the transformed prediction sets are defined like in 2 for any A and X. More explicitly, we let

$$C = \{y : \phi_{X_{test}}(A), y)) \leq \hat{q}\}, \quad \hat{q} \text{ s.t. } \left|\{\phi_{X_n}(A_n) \leq \hat{q}\}_{n=1}^N\right| = \lceil (N+1)(1-\alpha) \rceil \quad (4)$$

where $(X_n, Y_n) \in D$, $\phi_{X_n} \in \Phi_\theta$, $A_n = a(f(X_n), Y_n)$, $n = 1, \ldots, N$, and $\alpha \in [0, 1]$. If $a(f(X), Y) = (f(X) - Y)^2$ and under certain assumptions on Φ_θ, e.g. if Φ_θ is defined as in Assumption 1 (see below), C can be rewritten as

$$C = [f(X_{test}) - \Delta, f(X_{test}) + \Delta], \quad \Delta = \sqrt{\phi_{X_{test}}^{-1}(\hat{q})} \quad (5)$$

Differently from the case of a rigid, i.e. object-independent, monotone transformation, the prediction intervals produced by different Φ_θ may not be equivalent. See Section 2.5 for an example.

2.4. Model classes

In the examples of Section 2.5, the coordinate transformations belong to simple model classes, e.g. $\Phi_\theta = \{\phi_X(A) = \sqrt{A} - \theta X, \theta \in \mathbb{R}\}$. Varying the value of the parameter with a class produces non-equivalent intervals. Choosing different classes may also change the intervals. For concreteness, we restrict ourselves to model classes that satisfy the following assumption.

Assumption 1 *Let $\mathcal{X} = \mathbb{R}^d$, $\mathcal{B}_X \subseteq \mathbb{R}$, and*

$$\Phi_\theta = \{\phi_X : \mathbb{R}_+ \to \mathcal{B}_X, X \in \mathcal{X}\} \quad (6)$$

Φ_θ *is such that*

$$\phi'_X(A) > 0, \quad \text{for all } X \in \mathcal{X} \text{ and all } A \in \mathbb{R}_+ \quad (7)$$
$$\mathcal{B}_X = \mathcal{B}_{X'}, \text{ for all } X, X' \in \mathcal{X} \quad (8)$$

where $A = a(f(X), Y)$, $a = (f(X) - Y)^2$, and $h'(s) = \frac{d}{ds}h(s) = \frac{d}{ds'}h(s')|_{s'=s}$.

Requiring $\phi'_X > 0$ guarantees that $B = \phi_X(A)$ is a well-defined conformity score and the existence of the inverse map, $\phi_X^{-1} : \mathcal{B}_X \to \mathbb{R}_+$, which is defined implicitly by $\phi_X^{-1}(\phi_X(A)) = A$. Requiring the codomain of the transformations to be the same for all X guarantees that

5

we can convert the transformed prediction sets in (4) into lable-space symmetric intervals centred in $f(X_{test})$. Imagine that Φ_θ does not fulfil the second requirement of Assumption 1 and, for example, $\mathcal{B}_{test} \subset \mathcal{B}_{X_n}$ for any $n = 1, \ldots, N$. Then the inequalities in (4) may not have a real solution because $\hat{q} = \phi_{X_{n*}}(A_{n*})$, $n* \in \{1, \ldots, N\}$, may lay outside the codomain of $\phi_{X_{test}}$. Both assumptions are satisfied if, for example, $\phi_X(A) \to \phi_X(\log A) + \epsilon \log A$, $\phi'_X(A) > 0$ for all X and A, and $\epsilon > 0$.

2.5. Examples

2.5.1. INEQUIVALENT INTERVALS

Let $\mathcal{X} = \mathbb{R}$, $B = \phi_X(A) = \sqrt{A} - \theta X$, and $\theta_\pm = \pm 1$. Let the calibration set be $\{(x_n, y_n)\}_{n=1}^3$, f such that $(a_1, x_1) = (1, 1)$, $(a_2, x_2) = (2, 2)$, and $(a_3, x_3) = (3, 3)$, $x_{test} = 0$, and $\alpha = \frac{1}{2}$. If $\theta = 1$, $\{b_n\}_{n=1}^3 = \{2, \sqrt{2}+2, \sqrt{3}+3\}$ and $\hat{q} = \hat{q}_+ = \sqrt{2}+2$. If $\theta = -1$, $\{b_n\}_{n=1}^3 = \{0, \sqrt{2}-2, \sqrt{3}-3\}$ and $\hat{q} = \hat{q}_- = \sqrt{2}-2$. The prediction intervals, $C_\pm = [f(x_{test})-\Delta_\pm, f(x_{test})+\Delta_\pm]$ with $\Delta_\pm = \sqrt{\phi_{x_{test}}^{-1}(\hat{q}_\pm)} = \sqrt{(\hat{q}_\pm + \theta x_{test})^2} = |\hat{q}_\pm|$, have sizes $|C_+| = 2(2 + \sqrt{2}) > |C_-| = 2(2 - \sqrt{2})$, i.e. C_+ and C_- have different sizes.

2.5.2. ERC

Let $a = (f(X)-Y)^2$ and ϕ_X be the ERC coordinate transformations of Papadopoulos et al. (2008), i.e.

$$\phi_X(A) = \frac{A}{g(X)^2 + \gamma}, \quad A = (f(X) - Y)^2, \quad g : \mathbb{R}^d \to \mathbb{R}, \quad \gamma > 0 \tag{9}$$

$\Phi_\theta = \{\phi_X : \mathbb{R}_+ \to \mathcal{B}_X, X \in \mathcal{X}\}$ fulfils the requirements of Assumption 1 because $\phi'_X(A) = \frac{1}{g(X)^2+\gamma} > 0$ and $\mathcal{B}_X = \mathbb{R}_+$ for all X. The inverse transformation is $\phi_X^{-1}(B) = B(g(X)^2+\gamma)$ and $C = [f(X_{test})-\Delta, f(X_{test})+\Delta]$, where $\Delta = \sqrt{\hat{q}(g(X_{test})^2 + \gamma)}$ and $\hat{q}$ is the transformed conformity score of a calibration object. If $\hat{q}$ is the $(\lceil(N + 1)(1 - \alpha)\rceil)$th smallest object of $\{(\phi_{X_n}(A_n))\}_{n=1}^N$, the interval is marginally valid with confidence $(1 - \alpha)$ and has X_{test}-dependent size $|C| = 2\Delta$. In Papadopoulos et al. (2008), g is chosen to be a model of the conditional residuals, i.e. $g(X) \approx E_{Y|X}(f(X) - Y)$, but other choices are also possible.

2.5.3. NON-ADAPTIVE TRANSFORMATIONS

If the transformation does not depend on X, the second requirement in Assumption 1 is always satisfied. For example, consider

$$\phi_0(A) = \log A + \gamma, \quad A = (f(X) - Y)^2, \quad \gamma \in \mathbb{R} \tag{10}$$

then $\phi'_0(A) = \frac{1}{A} > 0$, $\mathcal{B}_X = (-\infty, \infty)$ for all X, and $\phi_0^{-1}(B) = \exp(B - \gamma)$. The corresponding prediction intervals have size $|C| = 2\sqrt{\exp(\hat{q} - \gamma)}$, which does not depend on X_{test}.

2.5.4. Different codomains

A set of transformations that satisfy $\phi'_X(A) > 0$ for all $X \in \mathcal{X}$ but may have $\mathcal{B}_X \neq \mathcal{B}_{X'}$ is

$$\phi_X(A) = A + g(X)^2, \quad A = (f(X) - Y)^2, \quad g : \mathbb{R}^d \to \mathbb{R} \tag{11}$$

as $\phi'_X(A) = 1 > 0$ and $\mathcal{B}_X = [g(X)^2, \infty) \subset \mathbb{R}_+$ for all X and A. The inverse transformation, $\phi_X^{-1}(B) = B - g(X)^2$ is well defined for all B and all X. The prediction intervals, however, may not have a straightforward interpretation in the label space. If $\hat{q} = \phi_{X_{n*}}(A_{n*}) = A_{n*} + g(X_{n*})$ is the ($\lceil (N+1)(1-\alpha) \rceil$)=th smallest object of $\{B_n = A_n + g(X_n)^2\}_{n=1}^N$, the interval at X_{test} is defined by $A_{test} \leq A_{n*} + g(X_{n*})^2 - g(X_{test})^2$, which is meaningless if $A_{n*} + g(X_{n*})^2 < g(X_{test})^2$ because $A_{test} = (f(X_{test}) - Y_{test})^2 > 0$. The problem does not arise if $\phi(A) \to \phi(\log A)$.

2.6. Validity of the prediction intervals

Global monotone transformations of the conformity scores do not affect the resulting prediction intervals. The first example of Section 2.5 shows this is not the case for locally-defined transformations. Under Assumption 1, however, the inverse map produces label-space prediction intervals that are marginally valid and have attribute-dependent sizes.

Theorem 1 *Let $(X_1, Y_1), \ldots, (X_N, Y_N), (X_{test}, Y_{test}) \sim P_{XY}$ be a collection of i.i.d. random variables, $f \sim \mathrm{E}_{Y|X}(Y)$ a point-prediction model, and $A_n = (f(X_n) - Y_n)^2$, $n = 1, \ldots, N$. Let Φ_θ be a family of strictly increasing scalar functions satisfying the requirements of Assumption 1 and $B_n = \phi_{X_n}(A_n)$, $n = 1, \ldots, N$. Assume $B_n \neq B_{n'}$ if $n \neq n'$. Then there exists a permutation of $\{1, \ldots, N\}$, $\{m_n = \pi(n)\}_{n=1}^N$ such that $B_{m_1} < \cdots < B_{m_n} < \cdots < B_{m_N}$ and, for any $\alpha \in [\frac{1}{N+1}, 1]$, the interval*

$$C = [f(X_{test}) - \Delta, f(X_{test}) + \Delta], \quad \Delta = \sqrt{\phi_{X_{test}}^{-1}(B_{m*})}, \quad m* = \lceil (N+1)(1-\alpha) \rceil \tag{12}$$

is marginally valid with confidence $1 - \alpha$, i.e.

$$\mathrm{Prob}\,(Y_{test} \in C)) \geq 1 - \alpha \tag{13}$$

Proof of Theorem 1 If $B_n \neq B_{n'}$ if $n \neq n'$, the exchangeability of (X_n, Y_n), $n - 1, \ldots, N$, and (X_{test}, Y_{test}) implies

$$\mathrm{Prob}(B_{test} \leq B_{m_n}) = \frac{m_n}{N+1}, \quad n = 1, \ldots, N \tag{14}$$

where $B_{test} = \phi_{X_{test}}(A_{test}) = \phi_{X_{test}}((f(X_{test}) - Y_{test})^2)$. In particular,

$$\mathrm{Prob}(B_{test} \leq B_{m*}) = \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1} \geq 1 - \alpha \tag{15}$$

or, equivalently, $\mathrm{Prob}(\phi_{X_{test}}(A_{test}) \leq B_{m*}) \geq 1 - \alpha$. The strict monotonicity of $\phi_X(A)$ for all $X \in \mathbb{R}^d$ implies that $\phi_{X_{test}}$ is invertible with inverse $\phi_{X_{test}}^{-1} : \mathcal{B}_{X_{test}} \to \mathbb{R}_+$ defined by

$A = \phi_{X_{test}}^{-1}(\phi_{X_{test}}(A))$. $\phi_{X_{test}}^{-1}$ and is also strictly increasing because $\phi_{X_{test}}^{-1'} = \frac{1}{\phi'_{X_{test}}}$ and $\phi'_{X_{test}} > 0$. [2] The validity of (12) follows from

$$1 - \alpha \leq \text{Prob}\left(B_{test} \leq B_{m*}\right) \tag{16}$$

$$= \text{Prob}\left(\phi_{X_{test}}^{-1}(B_{test}) \leq \phi_{X_{test}}^{-1}(B_{m*})\right) \tag{17}$$

$$= \text{Prob}\left(A_{test} \leq \phi_{X_{test}}^{-1}(B_{m*})\right) \tag{18}$$

$$= \text{Prob}\left((f(X_{test}) - Y_{test})^2 \leq \phi_{X_{test}}^{-1}(B_{m*})\right) \tag{19}$$

$$= \text{Prob}\left(Y_{test} \in [f(X_{test}) - \Delta, f(X_{test}) + \Delta]\right) \tag{20}$$

where $\Delta = \sqrt{\phi_{X_{test}}^{-1}(B_{m*})}$ and the first equality holds because $\phi_{X_{test}}^{-1}$ is strictly increasing. $\square$

2.7. Efficiency of the prediction intervals

If the conformity function does not depend on X, the size of the prediction intervals is minimal. [3] If the conformity scores are locally re-defined, i.e. $A \to \phi_X(A)$, the intervals can grow and shrink over the attribute space. Under Assumption 1, Theorem 1 ensures that the obtained locally-adaptive intervals remain valid. Their sizes, $|C| = 2\sqrt{\phi_{X_{test}}^{-1}(\hat{q})}$, which depend on X_{test}, Φ_θ, and the data, are not guaranteed to be optimal. Exact (optimal) conditionally-valid prediction intervals for all X can be obtained only if a finite-size data set is available for any $X \in \mathcal{X}$, which is impossible for real-valued attributes (Vovk (2012); Lei and Wasserman (2014)). Like other locally-adaptive CP algorithms, our scheme aims to approximate such unachievable conditional validity with finite data. To evaluate the efficiency of Φ_θ, we compute the average size of the prediction intervals at a user-defined confidence level, $(1 - \alpha) \in [0, 1]$, i.e.

$$\ell_\alpha = \text{E}\left(2\sqrt{\phi_{X_{test}}^{-1}(\hat{q})}\right), \quad \hat{q} \text{ s.t. } |\{\phi_{X_n}(A_n) \leq \hat{q}\}| = \lceil(N + 1)(1 - \alpha)\rceil \tag{21}$$

where $A_n = a(f(X_n), Y_n)$, $a(f(X), Y) = (f(X), Y)^2$, and $(X_n, Y_n) \in D$. Training Φ_θ by minimizing (21) over θ has two disadvantages. Firstly, ℓ_α depends on the empirical quantile of the transformed conformity scores, i.e. on the ranking of $\{B_n = \phi_{X_n}(A_n)\}_{n=1}^N$, which needs to be estimated numerically. Moreover, the dependence of $\hat{q}$ on Φ_θ may be hard to estimate without ad-hoc smooth relaxations. A second possible issue of (21) is it requires retraining the model if the target confidence level changes.

A possible way out is to average ℓ_α over all possible confidence levels, $(1 - \alpha) \in [0, 1]$. Taking the expectation of (21) over α produces a model that works well for any $\alpha \in [0, 1]$. More importantly, it bypasses the non-differentiability of (21) without introducing arbitrary relaxations or approximations. Intuitively, this happens because sorting the transformed conformity scores becomes redundant if we assume a flat prior over $\alpha \in [0, 1]$, i.e. if we let

$$\ell_{all} = \text{E}_{\alpha \sim \mathcal{U}_{[0,1]}}(\ell_\alpha) = \int_{[0,1]} d\alpha \, \ell_\alpha \tag{22}$$

---

2. $\phi_{X_{test}}^{-1'} = \frac{1}{\phi_{X_{test}'}}$ follows from $1 = \frac{d}{dA}\phi_{X_{test}}^{-1}(\phi_{X_{test}}(A)) = \phi'_X(\phi_{X_{test}}^{-1}(A))\phi_{X_{test}}^{-1'}(A)$.
3. From the definition of empirical quantile.

where ℓ_α is defined in (21). The number of non-equivalent confidence levels in an empirical version of ℓ_{all} coincides with the size of the calibration set, $|D| = N$. Each non-equivalent confidence level is associated with one of the transformed conformity scores, $\{B_n = \phi_{X_n}(A_n)\}_{n=1}^N$. Since the integral becomes a sum, the commutativity of the addends makes the reordering redundant. The size-N approximation of (22),

$$\ell_{all} = \mathrm{E}\left(\sum_{n=1}^n \sqrt{\phi_{X_{test}}^{-1}\left(\phi_{X_n}(A_n)\right)}\right) \tag{23}$$

is then smooth in θ if Φ_θ is smooth in θ for all X and A. In practice, Φ_θ is a fixed set of parametric transformations. The free parameter, θ, controls the shared functional dependence of all $\phi_X \in \Phi_\theta$. Specific functional dependencies may be chosen to ensure that Φ_θ is differentiable in θ and satisfies Assumption 1 (see Section 3 for a few examples). Given such Φ_θ, we can safely estimate (23) and its gradient (see Section 2.8) from the training data set, $\mathcal{D}$ defined in (1), by computing $B_n = \phi_{x_n}((f(x_n) - y_n)^2)$, $n = 1, \ldots, N$, their empirical quantile, $\hat{q}$, and $\phi^{-1}{}_{x_{test}}(\hat{q})$. A data-efficient strategy is to average over $N + 1$ splits of $\mathcal{D}$ deformed in (1). In the n-th split, we use (x_n, y_n) as a test object and the remaining objects as a calibration set.

2.8. Gradient optimization

The smoothness of (23) allows using any gradient-based optimization method. As $\phi_X^{-1}(\phi_{X'}(A))$ depends on θ directly and through its argument, the parameter updates are $\theta \to \theta + \eta d_\theta \ell_{all}$, where $\eta > 0$ and $d_\theta \ell_{all}$ is the total derivative of (23) with respect to θ, i.e.

$$d_\theta \ell_{all} = \mathrm{E}\left(\sum_{n=1}^n \frac{\nabla_\theta \phi_{X_{test}}^{-1}\left(\phi_{X_n}(A_n)\right) + \phi_{X_{test}}^{-1\,'}\left(\phi_{X_n}(A_n)\right)\nabla_\theta \phi_X(A_n)}{2\sqrt{\phi_{X_{test}}^{-1}\left(\phi_{X_n}(A_n)\right)}}\right) \tag{24}$$

Under Assumption 1, the monotone transformations $\phi_X \in \Phi_\theta$ are guaranteed to be invertible. Assumption 1, however, does not guarantee a closed-form expression of ϕ_X^{-1}. In that case, $\phi_X^{-1}(B)$ should be obtained numerically through standard root-finding methods, e.g. the Bisection method. As for the backward pass, $\nabla_\theta \phi_X^{-1}$ and $\phi_X^{-1\,'}$ can be obtained from $\nabla_\theta \phi_X$ and ϕ_X' using

$$0 = d_\theta \phi_X\left(\phi_X^{-1}(B)\right) = \nabla_\theta \phi_X\left(\phi_X^{-1}(B)\right) + \phi_X'\left(\phi_X^{-1}(B)\right)\nabla_\theta \phi_X^{-1}(B) \tag{25}$$

$$1 = \frac{d}{dB}\phi_X\left(\phi_X^{-1}(B)\right) = \phi_X'\left(\phi_X^{-1}(B)\right)\phi_X^{-1\,'}(B) \tag{26}$$

Similar techniques are increasingly popular in machine learning. Especially because they allow differentiating prediction models that include implicitly-defined outputs, e.g. the numerical solutions of an ODE (David Duvenaud and Johnson (2020)).

3. Experiments

We test the feasibility and performance of the proposed scheme by training and comparing the prediction intervals produced by four different CP algorithms. Each algorithm is based

on one of the following classes of transformations,

$$\Phi_{\texttt{ERC}} = \{\phi_X(A) = \frac{A}{g(X)^2 + \gamma}, \gamma > 0\} \tag{27}$$

$$\Phi_{\texttt{linear}} = \{\phi_X(A) = \log A + g(X)\} \tag{28}$$

$$\Phi_{\texttt{exp}} = \{\phi_X(A) = Ae^{g(X)}\} \tag{29}$$

$$\Phi_{\texttt{sigma}} = \{\phi_X(A) = \sigma\left(\log A + g(X)\right)\} \tag{30}$$

where g is an unconstrained trainable function of X. In all experiments, we let g be a fully connected ReLu network with five hidden layers of size 100. All model classes satisfy Assumption 1 and are trained by minimizing ℓ_{all} defined in (23). The ERC model, $\Phi_{\texttt{ERC}}$ is also trained as suggested in Papadopoulos et al. (2008), i.e. by minimizing $\text{E}((g(X) - (f(X) - Y)^2)^2)$. In the plots and tables, we refer to that setup as ERC (error fit). Our baseline is the non-adaptive CP algorithm, $\Phi_0 = \{\phi_X(A) = A\}$. The underlying point-prediction model is the same for all CP algorithms We use a KNN algorithm with the Euclidean distance and cross-validated K and pre-train it on a separate proper-train data set. The model efficiencies are measured by computing the interval sizes and empirical validities.

We test all models on a synthetic data set (Section 3.1) and five real-world benchmark data sets (Section 3.2). To optimize the parameterized localization network, $g = g(X, \theta)$, over θ, we use a PyTorch implementation of ADAM (Kingma and Ba (2014)), with a fixed learning rate, default parameters, and mini-batches of size 16. To avoid overfitting, we evaluate the efficiency of the intervals on a small validation set at each epoch and retrain the model up to the one associated with the most efficient intervals. All data sets are normalized and split into three parts, one to define the KNN point-prediction algorithm, one for training Φ_θ, and one for testing.

3.1. Synthetic data

In Figure 1, we show the locally adaptive intervals obtained by the models on four randomly generated synthetic data sets. In this case, the confidence level is set to $\alpha = 0.05$. Each data set consists of 1000 samples generated by perturbing an order-2 polynomial regression model, $f_{true}(X) = w_0 + w_1 X + w_2 X^2$, $w \in \mathcal{N}(0,1)^3$, with one of the four following X-depedent error functions,

$$\varepsilon_{\texttt{cos}} = \left(\rho + 2\cos\left(\frac{\pi}{2}|X|\right) \mathbf{1}(|X| < 0.5)\right)\xi \tag{31}$$

$$\varepsilon_{\texttt{squared}} = \left(\rho + 2X^2 \mathbf{1}(|X| > 0.5)\right)\xi \tag{32}$$

$$\varepsilon_{\texttt{inverse}} = \left(\rho + \frac{2}{\rho + |X|} \mathbf{1}(|X| > 0.5)\right)\xi \tag{33}$$

$$\varepsilon_{\texttt{linear}} = \left(\rho + (2 - |X|) \mathbf{1}(|X| < 0.5)\right)\xi \tag{34}$$

$\rho = 0.1$, $\xi \sim \mathcal{N}(0,1)$. In all cases, we sample X uniformly in $[-1,1]$ and then normalize the input vector $(1, X, X^2)^T$ across all samples. [4]

---

4. Normalizing the input causes the x-coordinate of some dots to exceeds $[-1,1]$ in Figure 1.
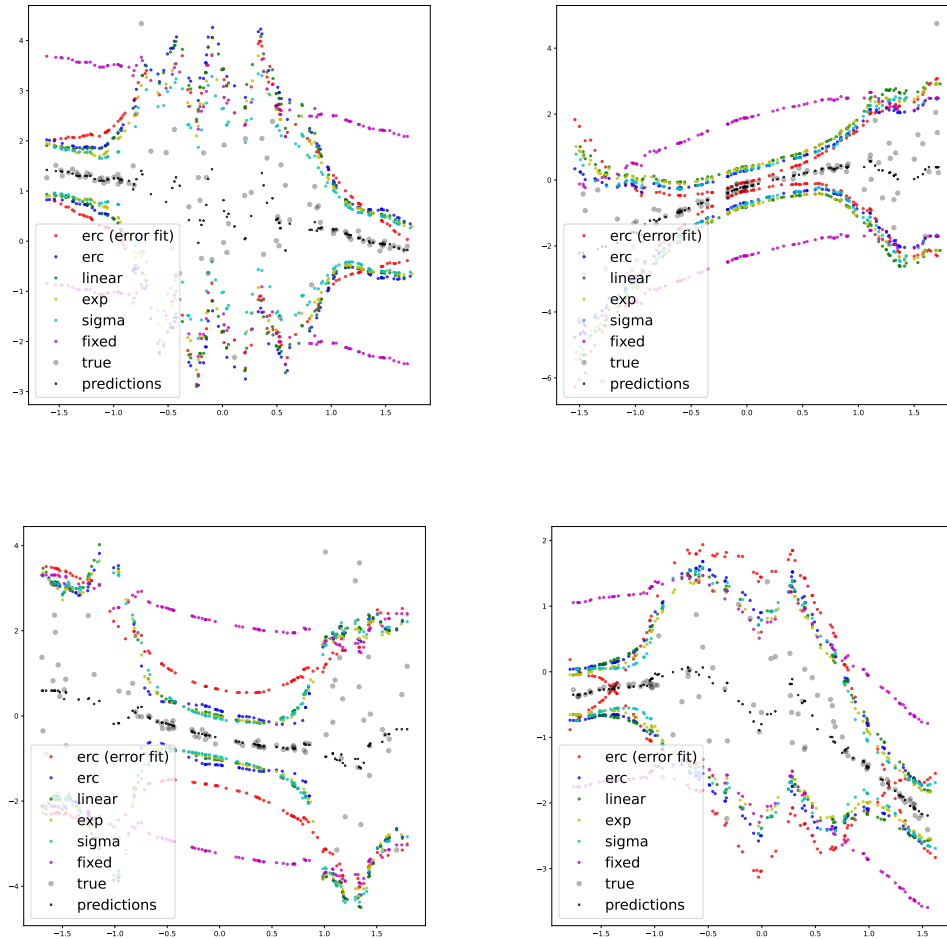
Figure 1: Examples of the locally adaptive intervals obtained with the proposed approach.

Figure 1 shows that the trained models can produce intervals of locally adaptive size. As noted in previous works, e.g. Romano et al. (2019), ERC may become unstable when the localization network, g, is trained by fitting the conditional squared errors. For the fairness of the comparison, we regularize ERC (error fit) through the cross-validated early-stopping technique we use for the other models instead of tuning an additional regularization hyperparameter, γ.

3.2. Real data

To check the scalability and efficiency of the algorithms on real-world data comparison, we have used the following regression data sets,

- energy, the Energy Efficiency Data Set, 768 observations with eight attributes, Tsanas and Xifara (2012),

- concrete, the Concrete Compressive Strength Data Set, 1030 observations with nine attributes, Yeh (2007),

- homes, the King County Home Price Data Set, 21613 observations with 14 attributes, King County (2015),

- CASP, the Physicochemical Properties of Protein Tertiary Structure Data Set, 45730 observations with nine attributes, Rana (2013),

- facebook_1, the Facebook Comment Volume Data Set, 40949 observations with 54 attributes, Singh (2016).

These are popular public benchmarks for evaluating the local adaptability of conformal prediction. Visit Tsanas and Xifara (2012); Yeh (2007); King County (2015); Rana (2013); Singh (2016) or see Romano et al. (2019); Sesia and Candès (2020); Guan (2023) for more details about the data sets. Table 1 shows the interval sizes and empirical validities on all datasets and for different confidence levels. Each value is obtained by averaging the performance over five runs of the training-testing procedure.

The results confirm that ERC may be unstable when the localization function is fitted to the conditional residual. All models, however, are consistently more efficient than the non-adaptive baseline, ϕ_0. The choice of the model class does not look to be crucial, with minor and data set-dependent performance differences.

4. Limitations and future work

We recognize that our work should have included a more extensive real-world validation of the methods. Firstly, we only consider a simple point-prediction model and do not run a careful ablation study involving more or less advanced regression models. While, in many cases, a localization network with five hidden layers of size 100 seems to make $\Phi_\theta \in \{\Phi_{\texttt{ERC}}, \Phi_{\texttt{linear}}, \Phi_{\texttt{exp}}, \Phi_{\texttt{sigma}}\}$ flexible enough, the efficiency of the intervals may be improved further by considering more general model classes. For example, it may be interesting to train a linear combination of $\Phi_{\texttt{ERC}}$, $\Phi_{\texttt{linear}}$, $\Phi_{\texttt{exp}}$, and $\Phi_{\texttt{sigma}}$ with nonnegative weights. While the model automatically satisfies Assumption 1, its inverse is not available in analytic form. The parameters of the localization network should then be trained through the implicit differentiation strategy described in Section 2.8.

As mentioned in Section 1.1, the proposed approach is not incompatible with existing methods that approximate object-conditional validity by re-weighting the empirical distribution of the conformity scores, e.g. Vovk (2012); Han et al. (2022); Barber et al. (2022); Guan (2023), Combining a global redefinition of the conformity measure with a sample-specific re-weighting of the calibration distribution may help localize the interval on specific tasks. Including the proposed strategy in the quantile regression approaches of Sesia and Candès (2020) would also be possible.

In a follow-up of this work, we will also explore how to interpret Φ_θ as a Normalizing Flow between the distribution of the original conformity scores and their transformed versions.

energy	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.32$	
	size	val	size	val	size	val
erc (error fit)	0.86 ± 0.582	0.968 ± 0.016	0.431 ± 0.289	0.924 ± 0.015	0.103 ± 0.025	0.684 ± 0.066
erc	0.259 ± 0.195	0.944 ± 0.023	0.184 ± 0.121	0.888 ± 0.02	0.083 ± 0.017	0.684 ± 0.023
linear	0.172 ± 0.018	0.948 ± 0.01	0.134 ± 0.011	0.9 ± 0.028	0.078 ± 0.009	0.688 ± 0.041
exp	0.161 ± 0.014	0.928 ± 0.02	0.131 ± 0.007	0.868 ± 0.027	0.075 ± 0.006	0.66 ± 0.046
sigma	0.169 ± 0.027	0.948 ± 0.016	0.135 ± 0.009	0.884 ± 0.008	0.078 ± 0.01	0.672 ± 0.016
fixed	0.192 ± 0.015	0.952 ± 0.01	0.143 ± 0.008	0.884 ± 0.023	0.082 ± 0.008	0.66 ± 0.038

concrete	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.32$	
	size	val	size	val	size	val
erc (error fit)	0.736 ± 0.17	0.936 ± 0.015	0.484 ± 0.072	0.888 ± 0.032	0.27 ± 0.024	0.656 ± 0.015
erc	0.903 ± 0.652	0.968 ± 0.016	0.637 ± 0.359	0.912 ± 0.016	0.29 ± 0.049	0.672 ± 0.041
linear	0.569 ± 0.088	0.96 ± 0.013	0.457 ± 0.049	0.9 ± 0.018	0.267 ± 0.033	0.668 ± 0.035
exp	0.567 ± 0.078	0.952 ± 0.016	0.45 ± 0.032	0.92 ± 0.022	0.274 ± 0.034	0.72 ± 0.025
sigma	0.557 ± 0.073	0.952 ± 0.032	0.453 ± 0.039	0.904 ± 0.015	0.268 ± 0.024	0.688 ± 0.016
fixed	0.61 ± 0.08	0.964 ± 0.02	0.472 ± 0.049	0.904 ± 0.046	0.271 ± 0.025	0.688 ± 0.052

homes	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.32$	
	size	val	size	val	size	val
erc (error fit)	1.664 ± 1.165	0.936 ± 0.008	0.864 ± 0.562	0.892 ± 0.01	0.234 ± 0.062	0.68 ± 0.031
erc	0.526 ± 0.038	0.936 ± 0.023	0.409 ± 0.047	0.876 ± 0.023	0.195 ± 0.017	0.656 ± 0.027
linear	0.522 ± 0.065	0.944 ± 0.015	0.425 ± 0.05	0.904 ± 0.015	0.206 ± 0.024	0.704 ± 0.06
exp	0.521 ± 0.054	0.94 ± 0.022	0.392 ± 0.053	0.892 ± 0.01	0.197 ± 0.018	0.68 ± 0.025
sigma	0.533 ± 0.077	0.936 ± 0.02	0.421 ± 0.058	0.892 ± 0.027	0.203 ± 0.027	0.632 ± 0.045
fixed	0.721 ± 0.087	0.94 ± 0.025	0.484 ± 0.093	0.896 ± 0.027	0.192 ± 0.024	0.672 ± 0.037

CASP	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.32$	
	size	val	size	val	size	val
erc (error fit)	1.375 ± 0.093	0.948 ± 0.02	1.073 ± 0.058	0.904 ± 0.023	0.643 ± 0.058	0.664 ± 0.056
erc	1.412 ± 0.191	0.936 ± 0.027	1.045 ± 0.048	0.888 ± 0.037	0.637 ± 0.056	0.648 ± 0.07
linear	1.392 ± 0.203	0.96 ± 0.018	1.018 ± 0.053	0.916 ± 0.015	0.631 ± 0.07	0.684 ± 0.032
exp	1.363 ± 0.187	0.944 ± 0.023	1.014 ± 0.07	0.892 ± 0.037	0.635 ± 0.042	0.68 ± 0.022
sigma	1.321 ± 0.207	0.956 ± 0.023	1.03 ± 0.108	0.904 ± 0.034	0.629 ± 0.075	0.688 ± 0.03
fixed	1.492 ± 0.207	0.952 ± 0.02	1.093 ± 0.086	0.9 ± 0.031	0.659 ± 0.063	0.668 ± 0.037

facebook_1	$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.32$	
	size	val	size	val	size	val
erc (error fit)	3.404 ± 1.063	0.956 ± 0.015	1.473 ± 0.293	0.916 ± 0.029	0.281 ± 0.062	0.692 ± 0.032
erc	3.332 ± 1.36	0.948 ± 0.024	1.389 ± 0.434	0.908 ± 0.027	0.282 ± 0.061	0.684 ± 0.029
linear	3.091 ± 0.694	0.944 ± 0.02	1.523 ± 0.353	0.904 ± 0.015	0.289 ± 0.068	0.672 ± 0.03
exp	3.268 ± 1.024	0.96 ± 0.022	1.534 ± 0.305	0.916 ± 0.041	0.281 ± 0.049	0.708 ± 0.02
sigma	3.088 ± 1.211	0.944 ± 0.015	1.376 ± 0.439	0.9 ± 0.025	0.292 ± 0.068	0.664 ± 0.05
fixed	3.612 ± 1.243	0.948 ± 0.035	1.519 ± 0.273	0.908 ± 0.035	0.276 ± 0.063	0.656 ± 0.054

Table 1: Size (left) and empirical validity (right) of the prediction intervals produced by the models on different datasets and for different confidence levels.

The goal is then to force the joint distribution of the conformity scores and the attributes to factorize, i.e. Φ_θ should be trained so that $P_{\phi(A)X} = P_{\phi(A)}P_X$, as this would make conditional and marginal validity equivalent.

The code to reproduce the numerical experiments described in Section 3 and possibly updated versions of this manuscript can be found in this Github directory.

Acknowledgements

References

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*, 2022.

Anthony Bellotti. Constructing normalized nonconformity measures based on maximizing predictive efficiency. In *Conformal and Probabilistic Prediction and Applications*, pages 41–54. PMLR, 2020.

Zico Kolter David Duvenaud and Matt Johnson. Deep implicit layers - neural odes, deep equilibirum models, and beyond. *NIPS 2020 Tutorial*, 2020.

Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. *arXiv preprint arXiv:2205.05878*, 2022.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

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

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

Rafael Izbicki, Gilson T Shimizu, and Rafael B Stern. Flexible distribution-free conditional predictive bands using density estimators. *arXiv preprint arXiv:1910.05575*, 2019.

WA King County. House Sales in King County, USA. Kaggle, 2015. DOI: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Steven George Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96, 2014.

Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69, 2008.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

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

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

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

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 34:6304–6315, 2021.

Kamaljot Singh. Facebook Comment Volume Dataset. UCI Machine Learning Repository, 2016. DOI: 10.24432/C5Q886.

David Stutz, Ali Taylan Cemgil, Arnaud Doucet, et al. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*, 2021.

Athanasios Tsanas and Angeliki Xifara. Energy efficiency. UCI Machine Learning Repository, 2012. DOI: 10.24432/C51307.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490. PMLR, 2012.

I-Cheng Yeh. Concrete Compressive Strength. UCI Machine Learning Repository, 2007. DOI: 10.24432/C5PK67.