Normalizing Flows for Conformal Regression

Abstract

Conformal Prediction algorithms estimate the uncertainty of a prediction model by calibrating its outputs on labelled data. The same calibration scheme usually applies to any model and data without modifications. The obtained prediction intervals are valid by construction but could be inefficient, i.e. unnecessarily big, if the errors are not uniformly distributed over the input space.

We present a general scheme to localize the intervals by training the calibration process, where the distance metrics used for calibration are learned from the data and depend explicitly on the object attributes. This is equivalent to training a Normalizing Flow that acts on the joint distribution of the prediction errors and the inputs. We apply the method to the Error-Reweighting model of Lei and Wasserman [2012]. The new framework allows estimating the gap between nominal and empirical conditional validity.

The approach is compatible with recent locally-adaptive CP strategies based on reweighting the calibration samples and applies to any point-prediction model without retraining.

1 INTRODUCTION

In natural sciences, calibration often refers to comparing measurements of the same quantity made by new and reference devices.[1] In data science, calibrating a model means quantifying the uncertainty of its predictions. Parametric and non-parametric approaches for calibrating prediction models have been proposed for decades. Examples of trainable post hoc calibration approaches are Platt scaling [Platt et al., 1999], Isotonic regression [Zadrozny and Elkan, 2002], and Bayesian Binning [Naeini et al., 2015]. In the regression setup, $Y \in \mathbb{R}$, the task is to define an algorithm that transforms a model prediction, $f(X) \approx \mathrm{E}_{Y|X} Y$, where $X \in \mathcal{X}$ are the attributes of the test object, into a valid Prediction Interval (PI) with target coverage, $C \subseteq \mathbb{R}$ such that $\mathrm{Prob}(Y \in C) \geq 1-\alpha, (1-\alpha) \in (0,1)$. Conformal Prediction (CP) algorithms output PIs that are non-asymptotically valid by construction [Vovk et al., 2005, Shafer and Vovk, 2008]. The properties hold without assumptions on f or the data generating distribution, P_{YX}. Different CP algorithms, however, may produce non-equivalent PIs for the same f. Several efficiency criteria have been proposed to assess their efficiency [Vovk et al., 2016]. For regression problems, a straightforward criterion is the average size of the PIs, $|C|$. Given f, a typical Conformal Calibration task is to reduce $|C|$ on specific regions of $\mathcal{X}$ where the Prediction Error (PE) is small while maintaining the overall coverage, i.e. to make the PIs locally adaptive [Vovk et al., 2020].

For simplicity, assume we have a set of calibration samples, $(X_1, Y_1), \ldots, (X_N, Y_N)$, and a test object, (X, Y), drawn independently from the same distribution. We use f, and a conformity function, ψ, to compute a series of calibration scores, $A_n = \psi(f(X_n), Y_n)$, $n = 1, \ldots, N$. Intuitively, ψ quantifies the quality of model predictions by comparing them with the corresponding labels. Conformal PIs are obtained from the sample quantile of the calibration scores, $\{A_n\}_{n=1}^{N}$. They are marginal PIs because coverage is defined in terms of the data joint distribution $\mathrm{Prob}(Y \in C_X) = P_{XYX_1Y_1 \ldots X_N Y_N}(Y \in C)$, without conditioning on either the object label or X [Vovk, 2012]. When the PE distribution varies across $\mathcal{X}$, e.g. data are heteroskedastic, marginal PIs may be inefficient. For example, they can be unnecessarily large if X is in a region

[1] The International Bureau of Weights and Measurements defines calibration as the *"operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties (of the calibrated instrument or secondary standard) and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication."*

where $f(X) - Y$ is small. An established research stream aims to improve the adaptability of the PIs [Lei and Wasserman, 2014, Vovk, 2012]. Unlike most existing methods, we address the task by training an X-dependent conformity function, ψ, without reweighting the calibration scores and hence breaking the data exchangeability, e.g. as in Lin et al. [2021], Tibshirani et al. [2019], Guan [2023].

1.1 OUTLINE

Technically, our goal is to approximate object-conditional PIs. Distribution-free and object-conditionally valid PIs can not be obtained from finite-size data [Lei and Wasserman, 2012, Vovk, 2012, Foygel Barber et al., 2021]. The strategy of Lin et al. [2021], Tibshirani et al. [2019], Guan [2023] is to replace the sample quantile of $\{A_n\}_{n=1}^N$, with the quantile of a reweighted empirical distribution, $\sum_{n=1}^N w_n \delta_{A_n}$, where $w = (w_1, \ldots, w_N)$ depends on the test attribute through a given localization function. Finding a localization function may be as challenging as estimating the underlying conditional distribution, $P_{A|X}$. Our approach is conceptually different. We learn a set of X-dependent conformity functions, $\phi_X(A)$, and use them without reweighting in the computation of the PIs. Data exchangeability is not broken because the transformation does not depend on the test object. In the transformed space, the PIs are marginally valid by construction. Local adaptability arises when we map them back to the label space (by inverting ϕ_X). Our starting point is to interpret ϕ_X as a Normalizing Flow (NF), i.e. a coordinate transformation that maps a target distribution, P, into a target distribution, P' [Papamakarios et al., 2021]. The target distribution is the joint distribution of the conformity scores and the object attributes, P_{AX}. The target distribution is a factorized distribution, $P'_{\phi_X X} = U_{\phi_X} P_X$, where U_{ϕ_X} is the composition between ϕ_X and an arbitrary univariate distribution, U_B. In the transformed space, the PIs have constant size but are conditionally valid because the joint distribution factorizes. To enforce the factorization, we train ϕ_X by maximizing the likelihood of the transformed sample under $= U_B$. We then invert ϕ_X to compute the PIs in the label space. The obtained intervals are marginally valid by construction but are as efficient as the ideal conditional PIs. While, in practice, we can only obtain a non-exact factorization, the scheme provides explicit error bounds on the validity of the obtained approximately conditional PIs.

1.2 AN EXAMPLE

Let $P_X = \text{uniform}(\mathcal{X})$ be the uniform distribution over $\mathcal{X} = [0, 1]$ and $(X_1, Y_1), \ldots, (X_N, Y_N), (X_{test}, Y_{test})$ a collection of i.i.d. random variables from $P_{XY} = P_{Y|X} P_X$ where

$$P_{Y|X} = \mathbf{1}(X < 0.5)\mathcal{N}(0, 1) + \mathbf{1}(X > 0.5)\mathcal{N}(0, 5) \quad (1)$$

We assume we have an exact prediction model, i.e. $f(X) = 0$ for all $X \in \mathcal{X}$, and choose the usual regression conformity measure, $\psi(f(X), Y) = |f(X) - Y| = |Y|$. We use f and ψ to form $D_A = \{(A_n, X_n) = (|Y_n|, X_n)\}_{n=1}^N$. By definition, D_A is also a collection of i.i.d. random variables. We often refer to the elements of D_A as the conformity scores. Let $1 - \alpha$ be the target confidence level. The corresponding marginal PIs are $C_{marginal} = [-\hat{Q}_A, \hat{Q}_A]$, where $\hat{Q}_A$ is the $(1 - \alpha)$-th sample quantile of D_A, i.e. the m_*-th smallest element of D_A, with $m_* = \lceil (1 - \alpha)(N + 1) \rceil$. If $N = 100$ and $\alpha = 0.05$, $m_* = 96$. The exchangeability of D_A implies $\text{Prob}(Y_{test} \in C_{marginal}) = \frac{m*}{N+1}$. Since f is constant over $\mathcal{X}$, $C_{marginal}$ is also constant over $\mathcal{X}$, i.e. the PIs have the same size for any test object X_{test}.

According to (1), the model PE depends deterministically on X. The data heteroscedasticity makes the marginal PIs inefficient (see Figure 1). In particular, $C_{marginal}$ are too large when $X_{test} < 0.5$ and too small when $X_{test} > 0.5$. Our conformal calibration scheme aims to increase their efficiency by learning a set of locally adaptive conformity functions, $\Phi(A) = \{\phi_X(A), X \in \mathcal{X}\}$. In particular, Φ will reduce the PIs for $X < 0.5$ and increase them for $X > 0.5$. The coverage guarantees of the marginal PIs remain the same because the same Φ applies to D_A and X_{test}, i.e. Φ does not break the data exchangeability.

Let C_Φ denote the marginal PIs obtained through Φ. Assuming (1), we can compute the exact conditional PIs and compare them with C_Φ and $C_{marginal}$. Let $D_{X<0.5} = \{(A, X) \in D_A, X < 0.5\}$, $D_{X>0.5} = \{(A, X) \in D_A, X > 0.5\}$, and $C_{conditional} = [-\hat{Q}_X, \hat{Q}_X]$, where $\hat{Q}_X = \mathbf{1}(X < 0.5)\hat{Q}_{A|X<0.5} + \mathbf{1}(X > 0.5)\hat{Q}_{A|X>0.5}$ and $\hat{Q}_{A|X<0.5}$ is the $(1 - \alpha)$-th sample quantile of $D_{X<0.5}$, i.e. its $m_{X<0.5}$-th smallest element, and $m_{X<0.5} = \lceil (1 - \alpha)(|D_{X<0.5}| + 1) \rceil$ (idem for $X > 0.5$). In words, the conditional PIs for $X < 0.5$ and $X > 0.5$ are the marginal PIs of the regions $[0, 0.5] \subset \mathcal{X}$ and $[0.5, 1] \subset \mathcal{X}$. Computing $C_{marginal}$ without knowing the ground truth distribution is unfeasible. Let

$$\Phi_\theta = \{\phi_X(A) = \frac{A}{\theta_1 + \theta_2 \sigma(MX))}, X \in [0, 1]\} \quad (2)$$

where $\theta_1, \theta_2 > 0$ are free parameters, $\sigma(t) = (1 + e^{-t})^{-1}$, and $M = 30$. For any X and θ, ϕ_X is a monotonic (and hence invertible) function of A. Φ_θ is the conformity function of Papadopoulos et al. [2008] with $\gamma = \theta_1$ and $g^2(X) = \theta_2 \sigma(MX)$. Let $D_{B_\Phi} = \{B_n = \phi_{X_n}(A_n), \phi_X \in \Phi, (A_n, X_n) \in D_A\}_{n=1}^N$. We refer to D_{B_Φ} as the transformed calibration-training set. In Figure 2, we compare a sample of D_A and a sample of $D_{B_{\Phi_\theta}}$ for two different choices of θ. Let $\hat{Q}_{B_\theta}$ be the $(1 - \alpha)$-th (marginal) sample quantile of $D_{B_{\Phi_\theta}}$, i.e. $\hat{Q}_{B_\theta} = A_{n_*}(\theta_1 + \theta_2 \sigma(MX_{n_*}))^{-1}$, with n_* such that there are m_* elements of $D_{B_{\Phi_\theta}}$ smaller than or equal to $\phi_{X_{n_*}}(A_{n_*})$.

The exchangeability of $B_1, \ldots, B_n$ and $B_{test} =$

$\phi_{X_{test}}(A_{test})$ implies $\text{Prob}(B_{test} \leq \hat{Q}_{B_\theta}) = \frac{m_*}{N+1}$. Equivalently, $\hat{Q}_{B_\theta}$ defines the marginally valid PIs in the transformed space, $[0, \hat{Q}_{B_\theta}] \subseteq \mathbb{R}$. The monotonicity of $\phi^{-1}_{X_{test}}(B)$ allows us to convert the transformed PIs back to the original space through $\text{Prob}(B_{test} \leq \hat{Q}_{B_\theta}) = \text{Prob}(A_{test} \leq \phi^{-1}_{X_{test}}(\hat{Q}_{B_\theta})) = \text{Prob}(|Y_{test}| \leq \phi^{-1}_{X_{test}}(\hat{Q}_{B_\theta})) = \text{Prob}(|Y_{test}| \leq \hat{Q}_{B_\theta}(\theta_1 + \theta_2\sigma(MX_{test}))) = \text{Prob}(Y_{test} \in C_\Phi)$. The chain of equality shows how the monotonicity of $\phi^{-1}_{X_{test}}(B)$ guarantees that validity of $C_\Phi = [-\Delta_\Phi, \Delta_\Phi]$, $\Delta_\Phi = \hat{Q}_{B_\theta}(\theta_1 + \theta_2\sigma(MX_{test}))$. If $M \to \infty$, $\theta = (1, 5)$, and $\hat{Q}_{B_\theta}$ is the $(1-\alpha)$-th sample quantile of a collection of $|D_{X<0.5}| \sim |D_{X>0.5}|$ Gaussian random variables with unit variance, Δ_Φ becomes $\hat{Q}_{Z\sim\mathcal{N}(0,1)}(1 + 5\mathbf{1}(X_{test} > 0.5))$, i.e. C_θ is equivalent to $C_{marginal}$.

If $P_{Y|X}$ is unknown, we need an optimization strategy to find θ. In the Locally Reweighted (ER) approach of Papadopoulos et al. [2008], θ_1 is a hyper-parameter and $\theta_2\sigma(MX)$ is a model of the conditional residuals, i.e. $\theta_{ER2} = \arg\min_t \sum_{(X,Y)\in D} |Y^2 - t^2\sigma^2(MX)|^2$. The results in Figures 1 and 2 are for $\theta_1 = 0.5$. In the proposed approach, we interpret Φ as an NF acting on $(A, X) \sim P_{AX}$ and train it by maximizing the likelihood of the transformed scores under a target input-independent distribution U_B. Choosing $U = \mathcal{N}(0, 1)$, we obtain

$$u_\Phi(A) = \frac{\exp\left(-\frac{1}{2}A^2(\theta_1 + \theta_2\sigma(MX))^{-2}\right)}{\sqrt{2\pi}(\theta_1 + \theta_2\sigma(MX))} \quad (3)$$

where the Jacobian of Φ is added because we evaluate the density at $B = \Phi(A)$ but use samples from A. In this case, we need to minimize $-\sum_{(A,X)\in D_A} \log(u_{\Phi(A)}(A, X)) = -\sum_{(A,X)\in D_A} \log(u_\Phi(A, X) - \text{const}$ over θ. Figure 1 shows the PIs obtained through the above procedure, C_{flow}, in red, and the ER approach, C_{ER}, in blue.

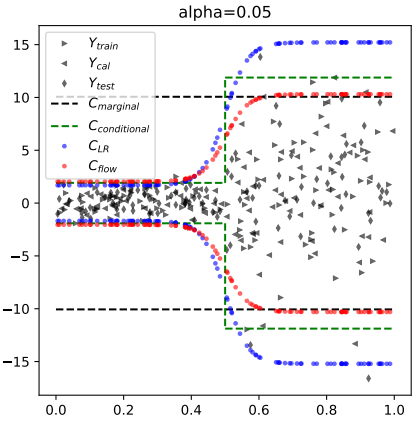

Figure 1: Samples and marginal and conditional PIs for the example of Section 1. C_{ER} and C_{flow} are the (marginal) PIs obtained through the ER approach of Papadopoulos et al. [2008] and the proposed method.

2 THEORY

In this section, $\mathcal{X}$ is an arbitrary attribute space and $\{(X_n, Y_n) \in \mathcal{X} \times \mathbb{R}\}_{n=1}^{N+1}$ is a collection of i.i.d. random variables from an unknow joint distribution, $P_{XY} = P_{Y|X}P_X$. We often refer to (X_{N+1}, Y_{N+1}) as (X_{test}, Y_{test}). $f(X_n) \approx \text{E}_{Y_n|X_n} Y_n$, $n = 1, \ldots, N+1$ is a pre-trained point-prediction model.

2.1 QUANTILES

Given a distribution, P_Z, let $F_Z(z) = P_Z(Z \leq z)$ be the Cumulative Distribution Function of P_Z. The $(1-\alpha)$-th quantile of $Z \sim P_Z$ is

$$Q_Z = \inf_q \{q : F_Z(q) \geq (1-\alpha)\} \quad (4)$$

When P_Z is continuous, F_Z is strictly increasing and $Q_Z = F_Z^{-1}(1-\alpha)$. The $(1-\alpha)$-th sample quantile of a collection of i.i.d. random variables, $\{Z_n \sim P_Z\}_{n=1}^N$, is the quantile of the empirical distribution $\hat{P}_Z = \frac{1}{N}\sum_{n=1}^N \delta_{Z_n}$, i.e.

$$\hat{Q}_Z = \inf_q \{q, |\{Z_n \leq q\}_{n=1}^N| \geq \lceil (N+1)(1-\alpha) \rceil\} \quad (5)$$

where $|S|$ is the cardinality of S and $\lceil s \rceil$ the smallest integer greater than or equal to s. Assuming ties occur with probability 0, i.e. $\text{Prob}(Z_n = Z_{n'}) = 0$ for any $n \neq n'$, $\hat{Q}_Z$ is the $\lceil (N+0)(1-\alpha) \rceil$-th smallest element of $\{Z_n \sim P_Z\}_{n=1}^N$. CP validity is a direct consequence of

Lemma 2.1 (Quantile Lemma Tibshirani et al. [2019]) *Let $Z_1, \ldots, Z_N, Z_{test} \in \mathbb{R}$ be a collection of i.i.d. random variables and $\hat{Q}_Z$ be the $(1-\alpha)$-th sample quantile of $\{Z_n\}_{n=1}^N$ defined in* (5)*. If $\text{Prob}(Z_n = Z_{n'}) = 0$ for any $n \neq n'$,*

$$\text{Prob}\left(Z_{test} \leq \hat{Q}_Z\right) = \frac{\lceil (1-\alpha)(N+1) \rceil}{N+1} \quad (6)$$

The standard bound, $1 - \alpha \leq \text{Prob}(Z_{test} \leq \hat{Q}_Z) \leq 1 - \alpha + \frac{1}{N+1}$, follows from $\lceil s \rceil - s \geq 0$ and $(1-\alpha)(N+1) \leq \lceil (1-\alpha)(N+1) \rceil \leq (1-\alpha)(N+1) + 1$. Asymptotically, $\hat{Q}_Z$ is normally distributed around Q_Z with variance $\sigma^2 = \frac{(1-\alpha)\alpha}{N p_Z(Q_Z)}$, where $p_Z(Q_Z)$ is the density of P_Z evaluated at $z = Q_Z$, with Q_Z defined in (4).

2.2 CONFORMITY SCORES

A conformity score is a random variable, $A = \psi(f(X), Y)$, that describes the conformity between a prediction, $f(X)$, and the corresponding label, Y. A standard choice is $\psi(s) = |f(X) - Y|$. Let P_{AX} be the joint distribution of the i.i.d.

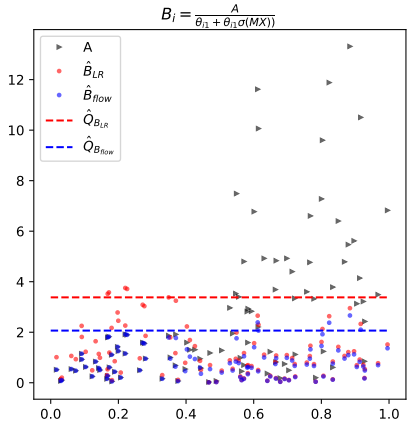

Figure 2: Calibration samples of the original and transformed random variables, $A = |Y|$, $B_{ER} = A(0.5 + \theta_{ER}\sigma(30X))^{-1}$, and $B_{flow} = A(\theta_{flow1} + \theta_{flow2}\sigma(30X))^{-1}$, where $\theta_{ER2} = \arg\min_t \mathrm{E}_{XY}|Y^2 - t^2\sigma^2(MX)|^2$, $\theta_{flow} = \arg\max_\theta \mathrm{E}_{XY} \log(u_\Phi(|Y|, X))$, and $u_\Phi(A, X)$ is defined in (3).

random variables $\{(A_n = |f(X_n) - Y_n|, X_n)\}_{n=1}^{N+1} =$. In this case, Lemma 2.1 directly guarantees the validity of the symmetric PI

$$C = [f(X_{test}) - \Delta_A, f(X_{test}) + \Delta_A] \tag{7}$$

when $X_{test} = X_{N+1}$, $\Delta_A = \hat{Q}_A$ and $\hat{Q}_A$ is the $(1-\alpha)$-th sample quantile of $\{A_n = |f(X_n) - Y_n|\}_{n=1}^N$. We may also choose $A = \phi(A)$, where ϕ is a global monotonic function of its argument, e.g. $\phi(s) = -s^{-1}$ or $\phi(s) = \log s$. In that case, we obtain the PIs by inverting ϕ, i.e. $\Delta_{\phi(A)} = \phi^{-1}(\hat{Q}_{\phi(A)})$, where $\hat{Q}_{\phi(A)}$ is the $(1-\alpha)$-th sample quantile of $\{\phi(A_n)\}_{n=1}^N$. For example, $\Delta_{-A^{-1}} = -\frac{1}{\hat{Q}_{-A^{-1}}}$ or $\Delta_{\log A} = \exp(\hat{Q}_{\log A})$. Assuming ties occur with probability 0, $\hat{Q}_A$ is the $\lceil (1-\alpha)(N+1) \rceil$-th smallest element of $\{A_n\}_{n=1}^N$. Let A_{n_*} be that element. The $(1-\alpha)$-th sample quantile of the transformed scores, $\hat{Q}_{\phi(A)}$, is the $\lceil (1-\alpha)(N+1) \rceil$-th smallest element of $\{\phi(A_n)\}_{n=1}^N$. If ϕ is monotonic and applies globally to all samples, $\phi(A_n) < \phi(A_{n'})$ if and only if $A_n < A_{n'}$, for any $n \neq n'$. Then $\hat{Q}_{\phi(A)} = \phi(A_{n_*})$ and $\Delta_{\phi(A)} = \phi^{-1}(\hat{Q}_{\phi(A)}) = \hat{Q}_A = \Delta_A$, i.e. the size of the PIs does not depend on ϕ. If ϕ is not applied globally on $\{A_n\}_{n=1}^N$, e.g. if it depends on the input, the above is untrue.

2.3 NORMALIZING FLOWS

This work is about finding a non-global transformation $\Phi = \{\phi_X(A), X \in \mathcal{X}\}$ that changes the PIs to make them locally adaptive. In what follows, we assume Φ always satisfies

Assumption 2.2 *Let $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}$. $\Phi = \{\phi_X : \mathcal{A} \to \mathcal{B}, X \in \mathcal{X}\}$ is such that*

1. *the domain and codomain of ϕ_X, $\mathcal{A}$ and $\mathcal{B}$, are the same for all X,*
2. *ϕ_X is strictly increasing on $\mathcal{A}$, i.e. $J_\Phi(A, X) = \frac{d}{dA}\phi_X(A) > 0$ for all $(A, X) \in \mathcal{A} \times \mathcal{X}$.*

The assumption on the domain and codomain of ϕ_X guarantees $\phi_X^{-1}(\phi_{X'}(A))$ is well defined for any $X \neq X'$. We avoid overfitting by imposing a smooth functional dependence of Φ on X and A. Since Φ acts on random variables, we may interpret it as an NF. Let P_Z and U_Z be two given distributions. An NF is an invertible coordinate transformation, $\phi : \mathcal{Z} \to \mathcal{U}$, such that

$$Z' = \phi(Z) \sim U_{Z'}, \quad Z = \psi^{-1}(Z') \sim P_Z \tag{8}$$

In our case, $Z = (A, X)$ and $Z' = (B, X)$, i.e. we have $\psi(A, X) = (\phi_X(A), X)$. In this case, the Jacobian of ψ is a $(d+1) \times (d+1)$ diagonal matrix, J_ψ, with all but the first element on the diagonal equal to one. The invertibility of ψ is guaranteed by the strict monotonicity of Φ because $\det(J_\psi(A, X)) = \prod_{i=1}^{d+1} J_{\psi\ ii}(A, X) = J_\Phi(A, X)$ and $J_\Phi(A, X) > 0$ for all A and X if Φ satisfies Assumption 2.2. See Papamakarios et al. [2021] for a review of using NFs in inference tasks.

2.4 VALIDITY

Given an NF, Φ, the marginal PI at X_{test} is

$$C_\Phi = [f(X_{test}) - \Delta_\Phi, f(X_{test}) + \Delta_\Phi] \tag{9}$$

$$\Delta_\Phi = \phi_{X_{test}}^{-1}(\hat{Q}_B) \tag{10}$$

where $\hat{Q}_B$ is the $(1-\alpha)$-th sample quantile of $\{B_n = \phi_{X_n}(A_n)\}_{n=1}^N$. In this case, we do not apply the same monotonic transformation to all samples because, in general, $\phi_{X_n} \neq \phi_{X_{n'}}$ for $n \neq n'$. Then $A_1 < A_2 < \cdots < A_N$ does not imply $B_1 < B_2 < \cdots < B_N$, i.e. we may have $A_n < A_{n'}$ and $\phi_{X_n}(A_n) > \phi_{X_{n'}}(A_{n'})$. If ties occur with probability 0, the validity of C_Φ is guaranteed by

Lemma 2.3 *Let Φ satisfy Assumption 2.2 and C_Φ be the PI defined in* (9)*. Then*

$$\mathrm{Prob}(Y_{test} \in C_\Phi) = \frac{\lceil (1-\alpha)(N+1) \rceil}{N+1} \tag{11}$$

While validity is not affected, Φ may still change the ranking of the calibration samples. This happens because, unlike for $\phi(A) = -A^{-1}$ or $\phi(A) = \log A$, we apply a different transformation to each sample. This observation is the core motivation of this work. Lemma 2.4 shows there exists a test object for which $|C_\Phi| \neq |C|$, under the further mild assumption that Φ is not constant.

Lemma 2.4 *LetΦ satisfy Assumption 2.2 and assume there exists X, X' such that $\phi_X(A) \neq \phi_{X'}(A)$ for any $A \in \mathcal{A}$. Then there exists X_{test} for which*

$$|C_\Phi| \neq |C| \tag{12}$$

where C_Φ and C are the marginal PIs defined in (9) *and* (7)*.*

2.5 EXACT NORMALIZING FLOWS

In some cases, marginal PIs are conditionally valid for any $X_{test} \in \mathcal{X}$, i.e. are such that

$$\text{Prob}(Y_{test} \in C | X_{test}) \tag{13}$$

with C defined in (7). In general, this is not true and may happen when P_{AX} has a specific form. For example, when the data is not heteroscedastic, i.e. when $P_{AX} = P_A P_X$ and hence $P_{A|X} = P_A$. The equivalence of marginal and conditional PIs is proven in

Theorem 2.5 *Let $P_{AX} = P_A P_X$ for any $X \in \mathcal{X}$. For any $X_{test} \in \mathcal{X}$,*

$$\text{Prob}(Y_{test} \leq C | X_{test}) = \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1} \tag{14}$$

where C is defined in (7)*.*

Theorem 2.5 is a straightforward consequence of the Bayesian theorem and Lemma 2.1. We include it here because it suggests we may make C_Φ approximately conditionally valid if Φ is such that $(\phi_X(A), X) = (B, X) \sim P_{BX} \approx P_B P_X$. Interpreting Φ as an NF, we train it by maximizing the likelihood of the transformed scores under an arbitrary target distribution, U_B, that does not depend on the input. As we only have samples from A, we need the composition between the target distribution and Φ, which we call U_Φ. The optimization problem is

$$\min_\Phi \ell, \quad \ell = -\text{E}_{AX} \log \left(|J_\Phi(A, X)| u_B(\phi_X(A)) \right) \tag{15}$$

where u_B is the (known) density of the target distribution U_B. The Jacobian of the transformation is added because we evaluate the density at $B = \Phi(A)$. Assume there exists a target distribution, U_B, and an NF satisfying Assumption 2.2, Φ, such that $P_{BX} = U_B P_X$ if $B = \Phi(A)$ for any (A, X). Under these assumptions, C_Φ defined in (9) is conditionally valid at X_{test}, as we show in

Corollary 2.6 *Let U_B be an arbitrary univariate distribution and $\Phi = \Phi(A)$ an NF satisfying Assumption 2.2. If $(B, X) = (\Phi(A), X) \sim\sim P_{BX} = U_B P_X$ for any $(A, X) \in \mathcal{A} \times \mathcal{X}$,*

$$\text{Prob}(Y_{test} \in C_\Phi | X_{test}) = \frac{\lceil (1-\alpha)(N+1) \rceil}{N+1} \tag{16}$$

with C_Φ defined in (9)*.*

Corollary 2.6 follows from Lemma 2.3 and the monotonicity of Φ. There is no contradiction with the negative result of Lei and Wasserman [2012], Vovk [2012] because exact factorization can not be achieved with finite data.

2.6 NON-EXACT NORMALIZING FLOWS

Let $\hat{\Phi}$ be an NF trained by minimizing a finite-sample empirical estimation of the likelihood defined in (15). We should not expect $\hat{\Phi}$ factorizes P_{BX} exactly. In what follows, we assume $\hat{\Phi}$ approximates Φ defined in Corollary 2.6. The assumption is technical and used for proving the error bounds below. More precisely, let $\epsilon > 0$ quantify the discrepancy between the target distribution, $P_{BX} = U_B P_X$, and the joint distribution of the optimized scores and the data, $P_{\hat{B}X}$, in the Huber sense, i.e.

$$P_{\hat{B}X} = (1-\epsilon) U_{\hat{B}} P_X + \epsilon S_{\hat{B}X}, \tag{17}$$

where U_B does not depend on X and $S_{\hat{B}X}$ is an unknown joint distribution representing the non-factorized part of the joint distribution of the transformed scores and the attributes, $P_{\hat{B}X}$. In terms of densities, we have $p_{\hat{B}X} = |J_{\hat{\Phi}}(A, X)| u_{\hat{\Phi}(A)} = (1-\epsilon) u_B + \epsilon s_{\hat{B}X}(A)$, where u_B and $s_{\hat{B}X}$ are the densities of U_B, and the error distribution, $S_{\hat{B}X}$. Theorem 2.7 characterizes the validity and size of $C_{\hat{\Phi}}$, i.e. the PIs obtained through the inexact NF, $\hat{\Phi}$, as in (9). In the theorem, we assume Φ and $\hat{\Phi}$ fulfil the requirements of Assumption 2.2, Φ satisfies the assumption of Corollary 2.6 and $\hat{\Phi}$ is the minimizer of (15) for a given target distribution U_B. We bound the size and validity of $C_{\hat{\Phi}}$ in terms of the variation distance between $B = \Phi(A)$ and $\hat{B} = \hat{\Phi}(A)$,

$$d_{\text{TV}}(P_B, P_{\hat{B}}) = \sup_b \| p_B(b) - p_{\hat{B}}(b) \| = 2\text{Prob}(B \neq \hat{B}) \tag{18}$$

where $p_Z(z)$ is the density of P_Z and the second equality follows from the Maximal Coupling Theorem.[2] Assume $\hat{\Phi}$ approximates Φ, which obeys the requirements of Corollary 2.6. Let $\epsilon > 0$ be the discrepancy parameter between the target and the modelled distribution defined in (17).

An NF which achieves an exact factorization would imply $P_{BX} = U_B P_X$, i.e. it would fulfil the assumptions of Corollary 2.6.

Theorem 2.7 *Let $\Phi(A)$, $\hat{\Phi}(A)$ be the functional defined in Assumption 2.2. Let U_B be an arbitrary target distribution. Assume $B = \Phi(A)$ and $\hat{B} = \hat{\Phi}(A)$ obey the Huber expansion* (17)*. Then*

$$\text{Prob}\left(Y_{test} \in C_{\hat{\Phi}} | X_{test} \right) \geq (1-\alpha)(1 + \frac{\epsilon}{2})^N \tag{19}$$

$$\geq 1 - \alpha - \delta \tag{20}$$

[2] See Lindvall [2002] or Ross and Peköz [2023] for an overview of coupling methods.

where $C_{\hat{\Phi}}$ is defined as in (9)*. An less tight additive boun* $\delta = 1 - e^{N \log(1-\frac{\epsilon}{2})}$.

Theorem 2.7 connects our work with the non-exchangeability gaps obtained in Barber et al. [2022] in a different framework. Using the Bretagnolle-Huber inequality

$$d_{\text{TV}}(P, P') \leq \sqrt{1 - e^{-KL(P,P')}} \tag{21}$$

Theorem 2.7 may allow an *a posteriori* estimation of the validity gap through an empirical estimate of the KL divergence between the NF and the target.

3 IMPLEMENTATION

We compare the ER model and a similar conformity function trained as described in Lei and Wasserman [2012] and through the proposed scheme.

3.1 DATA

We generate 4 synthetic data sets by perturbing the output of a polynomial regression model of order 2 with four types of heteroskedastic noise, e.g. $\epsilon \sim \sigma_{\texttt{synth-cos}}(X)\mathcal{N}(0,1)$, $\sigma_{\texttt{synth-cos}}(X) = 0.1 + 2\cos(\frac{\pi}{2}X)\mathbf{1}(X < 0.5)$. For the real-data experiments, we use 6 benchmark regression data sets from the UCI database: the Bike Sharing Data Set Fanaee-T [2013], `bike`, the Blog Feedback Data Set Buza [2014], `blog`, the Physicochemical Properties of Protein Tertiary Structure Data Set Rana [2013], `CASP`, the Concrete Compressive Strength Data Set Yeh [2007], `concrete`, the Communities and Crime Data Set Redmond [2009], `community`, the Energy Efficiency Data Set Tsanas and Xifara [2012], `energy`, and the Facebook Comment Volume Data Set of Singh [2016], `facebook_1`. More details are available in Appendixes B.1 and B.2

3.2 MODELS

The base conformity score is $A = |f(X) - Y|$ in all cases. The obtained PIs are not affected by this choice because they are invariant under global monotonic transformations. $f(X) \approx \text{E}_{Y|X}Y$ is a Random Forest model pre-trained using the `scikit-learn` optimizer on a separate proper-training set. $\Phi_{\texttt{ER}} = \{\phi_X(A) = \frac{e^A}{\gamma + g^2(X)}\}$ is the ER model of Papadopoulos et al. [2008], trained by minimizing $\ell_{ER} = \text{E}_{XY}|g^2(X) - (f(X) - Y)^2|^2$. $\Phi_{\texttt{ER-flow}}$ has the same functional form as $\Phi_{\texttt{ER}}$ but it is trained as described in Section 2 with $U_B = \text{Uniform}[0,1]$ as target distribution, i.e. $u_B(B) = \mathbf{1}(0 \leq B \leq 1)$. $\Phi_{\texttt{ERExp}} = \{\phi_X(A) = Ae{-}(\gamma + g(X)^2)\}$ is also trained by minimizing either the squared distances from the conditional residuals (`ERExp`) or (15) (`ERExp-flow`). The target distribution is the same as for the other models. g is a fully connected ReLu neural network with 5 hidden layers of 100 neurons. The data sets are split into a training and a test set of the same size. We use the training and test set for training $f(X)$ and Φ and evaluate the PIs. We use the ADAM gradient descent algorithm of Adam and Lorraine [2019] to solve all optimization problems. We set γ to $10e^{-4}$ in $\Phi_{\texttt{ER}}$ and $\Phi_{\texttt{ERExp}}$.

3.3 RESULTS

To evaluate the PIs, we look at their size, empirical validity, and approximated input-conditional coverage. Table 1 summarizes our numerical results across the 4 synthetic and 6 real data sets. We report the averages and standard deviations over 5 random train-test splits. `ER` produces the PIs with smaller average sizes and `ER-flow` the PIs with the best conditional coverage. In the synthetic experiments, fitting g directly may be the best strategy because we generate the data using $Y \sim f(X) + g(X)\epsilon$, $\epsilon \sim \mathcal{N}(0,1)$, the ER assumptions are 'exact'. The likelihood approach produces better conditional coverage as it does not penalize the PI average size. On synthetic data, `ERExp` underperforms `ER`, likely because the associated optimization is unstable. `ERExp-flow` achieves the best trade-off between efficiency and conditional coverage on real data. Such a trade-off between average size and conditional coverage was expected. The average size of the PIs is often incompatible with extreme local adaptivity. Figure 3 shows the correlation between the PI sizes and the approximate conditional validity for the real data sets. More details about the implementation are in Appendix B. The code for reproducing all numerical simulations is available in this `https://github.com/nicoloRHUL/ConformalCalibrationTraining`.

4 RELATED WORK

Calibration training The literature contains many examples of calibration optimization in data science [Platt et al., 1999, Zadrozny and Elkan, 2002, Naeini et al., 2015]. See Guo et al. [2017] for an introduction and empirical comparison of different calibration methods for neural networks. **Object-dependent conformity measures.** In ER Papadopoulos et al. [2008, 2011], the conformity measure is the ratio between the PE and a pre-trained model of the conditional residuals. Section 5 of Romano et al. [2019] contains a detailed discussion on the limitations of ER.

ER is intuitive and empirically effective but has been poorly investigated theoretically. Our work justifies it as approximating the unachievable conditional validity [Lei and Wasserman, 2012, Foygel Barber et al., 2021]. Recent work about ER includes Vovk et al. [2020], which is a theoretical study of the validity of oracle conformity measures

	synthetic data ($\alpha = 0.05$)		
	cover	size	WSC
basline	0.951(0.026)	4.571(0.967)	0.789(0.199)
ER	0.961(0.016)	2.591(0.493)	0.917(0.142)
ER-flow	0.961(0.031)	5.194(1.406)	0.930(0.094)
ERexp	0.964(0.021)	2.953(0.553)	0.928(0.08)
ERexp-flow	0.953(0.032)	4.764(1.168)	0.937(0.104)
	real data ($\alpha = 0.05$)		
	val	cover	WSC
basline	0.949(0.019)	2.253(0.846)	0.857(0.132)
ER	0.944(0.005)	3.072(2.269)	0.970(0.043)
ER-flow	0.966(0.014)	6.603(1.484)	0.976(0.041)
ERexp	0.964(0.015)	NA	0.935(0.091)
ERexp-flow	0.961(0.013)	2.467(0.749)	0.968(0.044)

Table 1: Averages and standard deviation of the coverage, size, and the Worst Slab Coverage (WSC) estimate of the input-conditional coverage Cauchois et al. [2020] of the PI obtained by the models over 4 synthetic and 6 real data sets. The reported averages and standard deviation are computed over 5 random training-test splits. More details about the experiments are in Appendix B. One value is NA because the optimized model produced PIs of unusually large size.

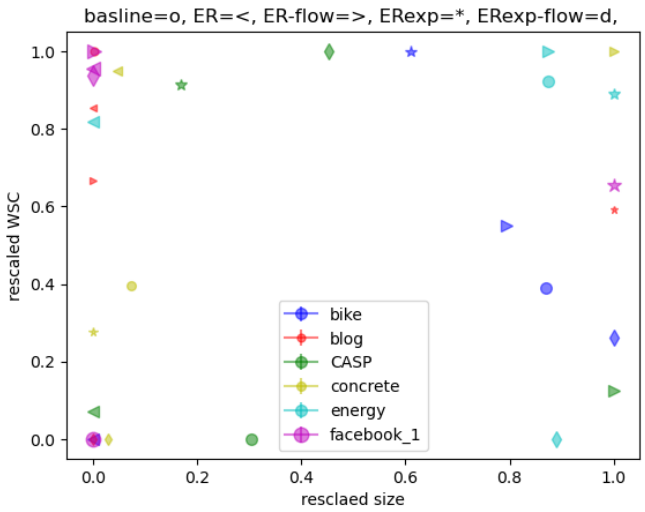

Figure 3: Trade-off between average size and average conditional coverage of the PIs. To compare different experiments, we use min-max rescaled quantities. The markers represent the averages over 5 runs for all models and real data sets. The size of the markers is proportional to the RSS of the underlying regressor. Table 1 shows aggregate averages and standard deviations over all experiments.

and Bellotti [2021], where the conformity score is iteratively updated to make the PI conditionally valid by minimizing an approximate empirical measure of the validity gap. Similarly to other work on CP localization, Bellotti [2021] requires estimating the empirical conditional probability, which is usually unreliable. Besides Papadopoulos et al. [2008], Bellotti [2020], conformity scores other than $A = |f(X) - Y|$ have been rarely used. In Romano et al. [2019], the conformity function is redesigned to mimic the pinball loss of quantile regression problems. We are unaware of works where the conformity measure is explicitly optimized. Henceforth, our scheme is orthogonal to all methods above, except for Papadopoulos et al. [2008]. Papadopoulos et al. [2008] is an exception because the conformity function is trained by minimizing $\mathrm{E}_{XY}|A^2 - g^2(X)|^2$. In Section 3, we show that training $g(X)$ as an NF may produce better PIs on real-world data. In Anonymous [2023] (to appear and by one of the authors), a series of trained conformity functions are tested empirically. Compared to this work, the learning scheme is not analyzed theoretically and uses a different learning loss. **Approximate conditional validity.** In Lei and Wasserman [2014], Vovk [2012], Lin et al. [2021], Guan [2023], Deutschmann et al. [2023], locally adaptive PI are constructed by reweighting the calibration samples and temporarily breaking data exchangeability. The weights transform the marginal distribution into an estimate of the object-conditional distribution. Often, computing the localizing wights requires a density estimation step based on one or more hyperparameters [Lei and Wasserman, 2014, Vovk, 2012, Guan, 2023, Deutschmann et al., 2023]. **Non-exchangeability.** Barber et al. [2022] is a study of CP under data non-exchangeability, with no explicit connections to the local adaptivity problem. In the context of online CP, Xu and Xie [2023] exploits the bounds of Barber et al. [2022] for proving the asymptotic convergence of the estimated PIs to the exact conditional PIs. Theorem 4 in Guan [2023] guarantees exact conditional coverage for a calibration-reweight method up to corrections to the estimated PI. The NF setup allows more explicit bounds on the validity of the actual algorithm outputs (Theorem 2.7 in Section 2). Similar to Barber et al. [2022] and Guan [2023], we exploit the similarity between approximate conditional CP and CP under non-exchangeability. In Einbinder et al. [2022], a point-prediction model is trained to guarantee $P_{AX} = UP_X$, where $U = \mathrm{Uniform}([0, 1])$. It is unclear whether tuning the point-prediction model or the conformity function produces equivalent PIs. This work is intuitively close to conformity-aware training, which aims to optimize the output of a standard CP algorithm by tuning the underlying model [Colombo and Vovk, 2020, Bellotti, 2020, Stutz et al., 2021, Einbinder et al., 2022]. The two ideas are compatible and could be implemented simultaneously. We leave this for future work.

5 DISCUSSION AND LIMITIATIONS

This is mainly a theoretical and methodological work. We recognize our numerical simulations are limited, especially regarding the model complexity. We also miss a full comparison with existing localization approaches. We focus on the ER conformity function to underline the efficiency of the proposed learning strategy without bias coming from the definition of more or less suitable model classes. Generalizing the approach to more complex NF is possible and straightforward, provided $\Phi(A)$ remains invertible, i.e. monotonic in A. We leave a more systematic classification and empirical validation to future work. A comparison with other localization methods goes beyond our scope because calibration training is orthogonal to many existing strategies, e.g. algorithms based on reweighting the calibration samples. The proposed scheme could be used on top of them and help provide theoretical guarantees. CP-aware retraining of the prediction model could also be combined with calibration training.

References

G Adam and J Lorraine. Understanding neural architecture search techniques. *arXiv preprint arXiv:1904.00438*, 2019.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Author Anonymous. On training locally adaptive cp. *arXiv preprint arXiv:2306.04648*, 2023.

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.

Anthony Bellotti. Approximation to object conditional validity with inductive conformal predictors. In *Conformal and Probabilistic Prediction and Applications*, pages 4–23. PMLR, 2021.

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

Maxime Cauchois, Suyash Gupta, and John Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *arXiv preprint arXiv:2004.10181*, 2020.

Nicolo Colombo and Vladimir Vovk. Training conformal predictors. In *Conformal and Probabilistic Prediction and Applications*, pages 55–64. PMLR, 2020.

Nicolas Deutschmann, Mattia Rigotti, and Maria Rodriguez Martinez. Adaptive conformal regression with jackknife+ rescaled scores. *arXiv preprint arXiv:2305.19901*, 2023.

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

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

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

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

Jing Lei and Larry Wasserman. Distribution free prediction bands. *arXiv preprint arXiv:1203.5422*, 2012.

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.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Locally valid and discriminative prediction intervals for deep learning models. *Advances in Neural Information Processing Systems*, 34:8378–8391, 2021.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence, 29-1*, 2015.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002.

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.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

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

Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C53W3X.

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

Sheldon M Ross and Erol A Peköz. *A second course in probability*. Cambridge University Press, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

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.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

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.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings 5*, pages 23–39. Springer, 2016.

Vladimir Vovk, Ivan Petej, Paolo Toccaceli, Alexander Gammerman, Ernst Ahlberg, and Lars Carlsson. Conformal calibrators. In *conformal and probabilistic prediction and applications*, pages 84–99. PMLR, 2020.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, pages 38707–38727. PMLR, 2023.

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

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

Conformal Calibration Training with Normalizing Flows
Supplementary Material

A PROOFS

$\mathcal{X}$ is an arbitrary attribute space and $\{(X_n, Y_n) \in \mathcal{X} \times \mathbb{R}\}_{n=1}^{N+1}$ is a collection of i.i.d. random variables from an unknow joint distribution, $P_{XY} = P_{Y|X}P_X$. We often refer to (X_{N+1}, Y_{N+1}) as (X_{test}, Y_{test}). $f(X_n) \approx \mathrm{E}_{Y_n|X_n}(Y_n)$, $n = 1, \ldots, N+1$ is a pre-trained point-prediction model. The $(1-\alpha)$-th sample quantile of a collection of i.i.d. random variables, $\{Z_n \sim P_Z\}_{n=1}^N$, is the quantile of the empirical distribution $\hat{P}_Z = \frac{1}{N}\sum_{n=1}^N \delta_{Z_n}$, i.e.

$$\hat{Q}_Z = \inf_q \{q, |\{Z_n \leq q\}_{n=1}^N| \geq \lceil (N+1)(1-\alpha) \rceil\} \tag{22}$$

where $|S|$ is the cardinality of S and $\lceil s \rceil$ the smallest integer greater than or equal to s.

Proof of Lemma 2.1 . Assume ties occur with probability 0. The definition of $\hat{Q}_Z$ in (5) implies that $\hat{Q}_Z$ is the $n* = \lceil (1-\alpha)(N+1) \rceil$-th smallest element of $\{Z_n\}_{n=1}^N$. Assume the samples have been labelled so that $Z_1 < Z_2 \cdots < Z_{N-1} < Z_N$. By assumption, $Z_1, \ldots, Z_n$, and Z_{test} are exchangeable. This implies Z_{test} falls with equal probability in any of the $N+1$ intervals

$$\begin{aligned}(-\infty, Z_1), &(Z_1, Z_2) \ldots, (Z_{n*-1}, \hat{Q}_Z), \\ &(\hat{Q}_Z, Z_{n*+1}) \ldots (Z_{N-1}, Z_N), (Z_N, \infty)\end{aligned} \tag{23}$$

This implies

$$\mathrm{Prob}(Z_{test} \leq \hat{Q}_Z) = \frac{n*}{N+1} = \frac{\lceil (1-\alpha)(N+1) \rceil}{N+1} \tag{24}$$

The lemma first appeared in Papadopoulos et al. [2002]. See also Lei and Wasserman [2014], Tibshirani et al. [2019], Angelopoulos and Bates [2021]. □

A conformity score is a random variable, $A = \psi(f(X), Y)$, that describes the conformity between a prediction, $f(X)$, and the corresponding label, Y. A standard choice is $\psi(s) = |f(X) - Y|$. Let P_{AX} be the joint distribution of the i.i.d. random variables $\{(A_n = |f(X_n) - Y_n|, X_n)\}_{n=1}^{N+1} =$. We focus on marginal PIs that are symmetric intervals of $\mathbb{R}$ centered in $f(X_{test})$, i.e.

$$C = [f(X_{test}) - \Delta_A, f(X_{test}) + \Delta_A] \tag{25}$$

where $\Delta_A = \hat{Q}_A$ and $\hat{Q}_A$ is the $(1-\alpha)$-th sample quantile of $\{A_n = |f(X_n) - Y_n|\}_{n=1}^N$. We consider non-global transformations $\Phi = \{\phi_X(A), X \in \mathcal{X}\}$ that satisfy (2.2). An NF is an invertible coordinates transformation, $\phi : \mathcal{Z} \to \mathcal{U}$, such that

$$Z' = \phi(Z) \sim U_{Z'}, \quad Z = \psi^{-1}(Z') \sim P_Z \tag{26}$$

Here, $Z = (A, X)$ and $Z' = (B, X)$, i.e. we have $\psi(A, X) = (\phi_X(A), X)$. Given an NF, Φ, the marginal PI at X_{test} is

$$C_\Phi = [f(X_{test}) - \Delta_\Phi, f(X_{test}) + \Delta_\Phi] \tag{27}$$

$$\Delta_\Phi = \phi_{X_{test}}^{-1}(\hat{Q}_B) \tag{28}$$

where $\hat{Q}_B$ is the $(1-\alpha)$-th sample quantile of $\{B_n = \phi_{X_n}(A_n)\}_{n=1}^N$. The validity of C_Φ is guaranteed by Lemma 2.3

Proof of Lemma 2.3 $B_1, \ldots, B_N, B_{test}$ are continuous i.i.d. random variable because Φ is deterministic and $A_1, \ldots, A_N, A_{test}$ are continuous i.i.d. random variables. When Φ satisfies Assumption 2.2, $\mathrm{Prob}(A_n = A_{n'}) = 0$ for any $n \neq n'$ implies $\mathrm{Prob}(B_n = B_{n'}) = \mathrm{Prob}(A_n = \phi_{X_n}^{-1}(\phi_{X_{n'}}(A_{n'})) = 0$ for any $n \neq n'$. Let $\hat{Q}_B$ be the $(1-\alpha)$-th sample quantile of $\{\phi_{X_n}(A_n)\}_{n=1}^N$. From Lemma 2.1, $\mathrm{Prob}(B_{test} \leq \hat{Q}_B) = \frac{n_*}{N+1}$, with $n_* = \lceil (1-\alpha)(N+1) \rceil$. Let

$\phi'_X(A) = J_\Phi(A, X)$. By Assumption 2.2, $\phi'_X(A) > 0$. From $\frac{d}{dB}\phi_X(\phi_X^{-1}(B)) = \phi'_X(\phi_X^{-1}(B))\frac{d}{dB}\phi_X^{-1}(B) = 1$, we obtain $\frac{d}{dB}\phi_X^{-1}(B) = \left(\phi'_X(\phi_X^{-1}(B))\right)^{-1} > 0$, i.e. $\phi_{X_{test}}^{-1}(B)$ is a monotonic function of B. Therefore,

$$\begin{aligned}
&\text{Prob}\left(B_{test} \leq \hat{Q}_B\right) && (29)\\
&= \text{Prob}\left(\phi_{X_{test}}^{-1}(B_{test}) \leq \phi_{X_{test}}^{-1}(\hat{Q}_B)\right) && (30)\\
&= \text{Prob}\left(\phi_{X_{test}}^{-1}(\phi_{X_{test}}(A_{test})) \leq \phi_{X_{test}}^{-1}(\hat{Q}_B)\right) && (31)\\
&= \text{Prob}\left(A_{test} \leq \phi_{X_{test}}^{-1}(\hat{Q}_B)\right) && (32)\\
&= \text{Prob}\left(|f(X_{test}) - Y_{test}| \leq \phi_{X_{test}}^{-1}(\hat{Q}_B)\right) && (33)\\
&= \text{Prob}\left(Y_{test} \in C_\Phi\right) && (34)
\end{aligned}$$

□ Lemma 2.4 shows there exists a test object for which $|C_\Phi| \neq |C|$, under the further mild assumption that Φ is not constant.

Proof of Lemma 2.4 Let $\{B_n = \phi_{X_n}(A_n)\}_{n=1}^{N+1}$ and $A_{test} = A_{N+1}$, and $B_{test} = B_{N+1}$. By definition of sample quantile, there exist m_* and n_* such that $\hat{Q}_A = A_{m*}$ and $\hat{Q}_{\hat{B}} = \phi_{X_{n_*}}(A_{n*})$. We may have $n_* = m_*$ or $n_* \neq m_*$. If $n_* = m_*$, let X_{test} be such that $\phi_{X_{test}}(A_{n_*}) \neq\neq \phi_{X_{n_*}}(A_{n_*})$. Then

$$\begin{aligned}
|C_{\hat{\Phi}}| &= \phi_{X_{test}}^{-1}(\phi_{X_{m_*}}(A_{m_*}) && (35)\\
&= \phi_{X_{test}}^{-1}(\phi_{X_{n_*}}(A_{n_*}) && (36)\\
&\neq A_{n_*} = |C| && (37)
\end{aligned}$$

If $n_* \neq m_*$, let $X_{test} = X_{m*}$. Then

$$\begin{aligned}
|C_{\hat{\Phi}}| &= \phi_{X_{test}}^{-1}(\phi_{X_{m_*}}(A_{m_*}) && (38)\\
&= \phi_{X_{m_*}}^{-1}(\phi_{X_{m_*}}(A_{m_*})) && (39)\\
&= A_{m_*} && (40)\\
&\neq A_{n_*} = |C| && (41)
\end{aligned}$$

where $A_{m_*} \neq A_{n_*}$ because we assume there are no ties. □

The equivalence of marginal and conditional PIs when $P_{AX} = P_A P_X$ is proven in Theorem 2.5.

Proof of Theorem2.5 Let $\{A_n^{(X_{N+1})} \sim P_{A|X_{N+1}}\}_{n=1}^N$ be a collection of i.i.d random variables at X_{N+1} and $\hat{Q}_{A|X_{test}}$ the sample quantile of $\{A_n^{(X_{N+1})}\}_{n=1}^N$. Assume ties occur with probability zero, i.e. $\text{Prob}(A_n^{(X_{N+1})} = A_{n'}^{(X_{N+1})}) = 0$ for any $n \neq n'$. Let $C^{(X_{N+1})}$ be defined as in (7) with $\Delta = \hat{Q}_{A|X_{N+1}}$. By the Bayesian theorem, the assumption on P_{AX} implies $P_{A|X} = P_A = \sum_X P_{AX}$. Then, for any X_{test}, $A_{test} \sim P_A$ and the claim follows from Lemma 2.1 and $\text{Prob}(A_{test} \leq \hat{Q}_A) = \text{Prob}(Y_{test} \in C)$. □

We train Φ by maximizing the likelihood of $B = \Phi(A)$ under the target distribution U_B, i.e. by minimizing (15). Given a target distribution, U_B, assume there exists an NF, Φ, that satisfies Assumption 2.2 and is such that $P_{AX} = U_\Phi P_X$ for any X. Under these assumptions, C_Φ defined in (9) is conditionally valid at X_{test}, as we show in Corollary 2.6.

Proof of Corollary 2.6 The conditional PIs, $C^{(X_{test})}$ such that $\text{Prob}(Y_{test} \in C^{(X_{test})}|X_{test}) = \frac{m_*}{N+1}$, $m_* = \lceil (1-\alpha)(N+1) \rceil$, can be defined as in (7) with Δ_A replaced by $\Delta_{A|X} = \hat{Q}_{A|X_{test}}$, where $\hat{Q}_{A|Xtest}$ is the $(1-\alpha)$-th sample quantile of $\{A_n^{(X_{test})} \sim P_{A|X_{test}}\}_{n=1}^N$. The assumption $P_{AX} = P_{A|X} P_X = U_\Phi P_X$ implies $(A, X) \sim \phi_X(B), X)$ for any X. Then

$$\begin{aligned}
\hat{Q}_{A|X_{test}} &= \hat{Q}_{\phi_{X_{test}}^{-1}(B)|X_{test}} && (42)\\
&= \phi_{X_{test}}^{-1}(\hat{Q}_{B|X_{test}}) && (43)\\
&= \phi_{X_{test}}^{-1}(\hat{Q}_B) && (44)
\end{aligned}$$

where the first and second equalities hold because $\phi^{-1}_{X_{test}}$ is monotonic and applied globally to all samples and $P_{BX} = P_{B|X}P_X = UP_X$, which implies $\hat{Q}_{B|X} = \hat{Q}_B$ for any X. The claim is obtained by rewriting $\mathrm{Prob}(Y_{test} \in \hat{Q}_{A|X_{test}})$ as $\mathrm{Prob}(Y_{test} \in C^{(X_{test})}|X_{test}) = \frac{m_*}{N+1}$. $\square$

In Theorem 2.7, we use the NF formalism to prove a quantitative bound on the gap between nominal and empirical conditional validity of the PI obtained from $\hat{B} = \hat{\Phi}(A)$.

Proof of Theorem 2.7. Let $\{A_n = |f(X_n) - Y_n|\}_{n=1}^{N+1}$ be the collection of conformity scores and $\{B_n = \phi_{X_n}(A_n), \phi_X \in \Phi\}_{n=1}^{N+1}$ and $\{\hat{B}_n = \hat{\phi}_{X_n}(A_n), \hat{\phi}_X \in \hat{\Phi}\}_{n=1}^{N+1}$ the collections of conformity scores transformed by Φ and $\hat{\Phi}$. Assumption 2.2 and the assumption of Corollary 2.6 imply

$$\begin{aligned}
&\mathrm{Prob}(Y_{test} \in C_{\hat{\Phi}}|X_{test}) && (45)\\
&= \mathrm{Prob}(A_{test} \leq \hat{\phi}^{-1}_{X_{test}}(\hat{Q}_{\hat{B}})|X_{test}) && (46)\\
&= \mathrm{Prob}(\phi^{-1}_{X_{test}}(B_{test}) \leq \hat{\phi}^{-1}_{X_{test}}(\hat{Q}_{\hat{B}}))|X_{test}) && (47)\\
&= \mathrm{Prob}(B_{test} \leq \phi_{X_{test}}(\hat{\phi}^{-1}_{X_{test}}(\hat{Q}_{\hat{B}}))) && (48)
\end{aligned}$$

where C_Φ and $C_{\hat{\Phi}}$ are defined as in (9). We can drop the conditioning in the last line because, by assumption, $B_n \sim P_{B|X} = P_B$ for all X. The test and calibration data are not exchangeable. The coverage gap is bounded in terms of the total variation distance between the distribution of $\tilde{B} = \phi_{X_{test}}(\hat{\phi}^{-1}_{X_{test}}(\hat{B})$ and P_B, i.e.

$$\mathrm{d_{TV}}(P_B, P_{\tilde{B}}) = \sup_Z |P_B(Z) - P_{\tilde{B}}(Z)| \tag{49}$$

In particular, we use

$$\begin{aligned}
\mathrm{Prob}(B_{test} \leq \hat{Q}_{\hat{B}}) &= \mathrm{Prob}(B_{test} \leq \hat{Q}_B, \{B_n\}_{n=1}^N = \{\hat{B}_n\}_{n=1}^N) && (50)\\
&= \mathrm{Prob}(B_{test} \leq \hat{Q}_B)\mathrm{Prob}(\{B_n\}_{n=1}^N = \{\hat{B}_n\}_{n=1}^N) && (51)\\
&= \mathrm{Prob}(B_{test} \leq \hat{Q}_B) \prod_{n=1}^N \Big(1 - \mathrm{Prob}(B_n \neq \hat{B}_n)\Big) && (52)\\
&= \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1} \Big(1 - \mathrm{Prob}(B_1 \neq \hat{B}_1)\Big)^N && (53)\\
&= \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1} \left(1 - \frac{1}{2}\mathrm{d_{TV}}(P_B, P_{\tilde{B}})\right)^N && (54)
\end{aligned}$$

where $\hat{Q}_{\tilde{B}}$ is the sample quantile of

$$\{\tilde{B}_n = \phi_{X_{test}}(\hat{\phi}^{-1}_{X_{test}}(\hat{B}_n), \hat{B}_n = \hat{\phi}_{X_n}(A_n)\}_{n=1}^N \tag{55}$$

The joint probabilities factorize because the events $B_{test} \leq \hat{Q}_B$ and $\{B_n\}_{n=1}^N = \{\hat{B}_n\}_{n=1}^N$ are independent. The last equalities follow from the maximal coupling theorem (for example, see Lindvall [2002], Ross and Peköz [2023]), i.e.

$$\mathrm{Prob}(B \neq \hat{B}) = \frac{1}{2}\mathrm{d_{TV}}(P_B, P_{\tilde{B}}) \tag{56}$$

Under the assumption $P_{\hat{B}X} = (1-\epsilon)U_{\hat{B}}P_X + \epsilon S_{\hat{B}X}$, we have

$$\begin{aligned}
\mathrm{d_{TV}}(P_B, P_{\tilde{B}}) &= \sup_{(b,x)} \|u_B(b)p_X(x) - p_{\hat{B}}(b,x)\| && (57)\\
&= \sup_{(b,x)} \|u_B(b)p_X(x) - (1-\epsilon)p_B(b)p_X(x) - \epsilon s_{\hat{B}X}(b,x)\| && (58)\\
&= \epsilon \sup_{(b,x)} \|p_B(b)p_X(x) - s_{\hat{B}X}(b,x)\| && (59)\\
&\leq \epsilon && (60)
\end{aligned}$$

implying

$$\text{Prob}(B_{test} \leq \hat{Q}_{\hat{B}}) \geq \frac{\lceil (N+1)(1-\alpha) \rceil}{N+1} \left(1 - \frac{\epsilon}{2}\right)^N \tag{61}$$

$$\geq (1-\alpha)\left(1 - \frac{\epsilon}{2}\right)^N \tag{62}$$

The additive bound is obtained by defining $\hat{\delta} = 1 - \left(1 - \frac{\epsilon}{2}\right)^N$ and using

$$(1-\alpha)(1-\hat{\delta}) \geq 1 - \alpha - \hat{\delta} \tag{63}$$

□

B MORE ON THE EXPERIMENTS

We run all experiments 5 times with 5 random splits of each data set. The data is split into a training set, D_{train}, a calibration-training set, D_{cal}, and a test set, D_{test}, with ratio $[0.5, 0.25, 0.25]$. Let $N = |\mathcal{D}_{train}|$. The point prediction model, $f(X)$ is the Random Forest model of `sklearn.sklearn.ensemble` with default settings. We train f by running

```
Xtrain, ytrain = train
forest = sklearn.ensemble.RandomForestRegressor()
f = forest.fit(Xtrain, np.ravel(ytrain))
```

where `train` is the training data set. We use the same data set to train the calibration models. The pre-trained f is used to compute the base conformity scores and form $D_A = \{A_n = |f(X_n) - Y_n|, (X_n, Y_n) \in D_{cal}\}_{n=1}^N$. We then train all calibration models on D_A Φ_i, $i \in \{\texttt{ER}, \texttt{ER} - \texttt{flow}, \texttt{EREXp}, \texttt{EREXp} - \texttt{flow}\}$, where

$$\Phi_{\texttt{ER}} = \{\phi_X(A) = \frac{A}{\gamma + g^2(X)}, X \in \mathcal{X}\}, \quad \gamma = 1e^{-4} \tag{64}$$

$$\Phi_{\texttt{EREXp}} = \{\phi_X(A) = Ae{-}(\gamma + g^2(X)), X \in \mathcal{X}\}, \quad \gamma = 1e^{-4} \tag{65}$$

where $g(X)$ is a fully connected neural network defined in `pyTorch` as

```
g = [nn.Linear(in_dim, hidden_dim), nn.ReLU()]
for n in range(num_layers):
  g.append(nn.Linear(hidden_dim, hidden_dim))
  g.append(nn.ReLU())
  g.append(nn.Linear(hidden_dim, 1))
  g = OrderedDict([(str(i), v[i]) for i in range(len(v))])
g = nn.Sequential(v)
```

with `hidden_dim` and `num_layers` set to 100 and 5 in all experiments. We append `-flow` when g is trained by maximizing the likelihood

$$\ell = \sum_{(A,X) \in D_A} -\log\left((u_B(\phi_X(A))J_\Phi(A,X)\right), \quad i \neq \texttt{ER} \tag{66}$$

where $U_B = \text{Uniform}([0,1])$. To solve the optimization problem, we use the ADAM gradient descent algorithm Kingma and Ba [2014] with default parameters and learning rate $1e^{-4}$ for all models and all data sets. We avoid data and model-specific tuning to avoid bias in the comparison. We use the same optimization setup on all synthetic and real-data experiments.

B.1 SYNTHETIC DATA

The synthetic data sets consist of 1000 samples from the following generative model

$$X_1 \sim \text{Uniform}([-1, 1]), \tag{67}$$
$$X = [1, X_1, X_1^2], \tag{68}$$
$$Y = X^T w + \epsilon_i, \tag{69}$$
$$\epsilon_i = 0.1 + \sigma_{\texttt{synth-i}}(X)E \tag{70}$$
$$E \sim \mathcal{N}(0, 1) \tag{71}$$

where $w \in \mathbb{R}^3$ is a randomly generated fixed parameter, $\texttt{i} \in \{\texttt{cos}, \texttt{squared}, \texttt{inverse}, \texttt{linear}\}$, and

$$\sigma_{\texttt{synth-cos}}(X) = 2\cos\left(\frac{\pi}{2}X_1\right)\mathbf{1}(X_1 < 0.5) \tag{72}$$
$$\sigma_{\texttt{synth-squared}}(X) = 2X_1^2\mathbf{1}(X_1 > 0.5) \tag{73}$$
$$\sigma_{\texttt{synth-inverse}}(X) = 2\frac{1}{0.1 + |X_1|}\mathbf{1}(X_1 < 0.5) \tag{74}$$
$$\sigma_{\texttt{synth-linear}}(X) = 2|X_1|\mathbf{1}(X_1 > 0.5) \tag{75}$$

B.2 REAL DATA

We selected the following 6 public benchmark data sets from the UCI database:

- `bike`, the Bike Sharing Data Set of Fanaee-T [2013], 10886 observations with 18 attributes,
- `blog`, the Blog Feedback Data Set of Buza [2014], 52397 observations with 280 attributes,
- `CASP`, the Physicochemical Properties of Protein Tertiary Structure Data Set of Rana [2013], 45730 observations with 9 attributes,
- `concrete`, the Concrete Compressive Strength Data Set of Yeh [2007], 1030 observations with 8 attributes,
- `energy`, the Energy Efficiency Data Set of Tsanas and Xifara [2012], 768 observations with 8 attributes, and
- `facebook_1`, the Facebook Comment Volume Data Set of Singh [2016], 40948 observations with 54 attributes.

B.3 EVALUATION

To evaluate the PIs, we used their average size empirical validity

$$\text{size} = N^{-1}\sum_{n=1}^{N} |C_{\phi_{X_{test\ n}}}| \tag{76}$$
$$\text{cover} = N^{-1}\sum_{n=1}^{N} \mathbf{1}(Y_{test\ n} \in C_{\phi_{X_{test\ n}}}), \tag{77}$$
$$\tag{78}$$

where $\hat{Q}_B$ is the sample quantile of the transformed calibration set, $\{\phi_{X_n}(A_n)\}_{n=1}^N$, and the Worst Slab Coverage (WSC) estimate of the input-conditional coverage Cauchois et al. [2020]. At test time, we randomly split the test data set into two subsets, $\mathcal{D}_{test-cal}$ and $\mathcal{D}_{test-test}$ such that $|\mathcal{D}_{test-cal}| = |\mathcal{D}_{test-test}|$. We use $\mathcal{D}_{test-cal}$ to calibrate the test models and $\mathcal{D}_{test-test}$ to evaluate the PIs. Averages and standard deviations are over 5 random training-test splits.