Distribution-free risk assessment of regression-based machine learning algorithms

Sukrita Singh SSINGH5@MUNICHRE.COM
University of Oxford, Wellington Square, Oxford OX1 2JD, United Kingdom
Neeraj Sarna NSARNA@MUNICHRE.COM
Munich Re, Koeniginstr. 107, 80802, Munich, Germany
Yuanyuan Li YLI@MUNICHRE.COM
Munich Re, Koeniginstr. 107, 80802, Munich, Germany
Yang Lin YANG_LIN@HSB.COM
Hartford Steam Boiler, Connecticut, USA
Agni Orfanoudaki AGNI.ORFANOUDAKI@SBS.OX.AC.UK
University of Oxford, Wellington Square, Oxford OX1 2JD, United Kingdom
Michael Berger MBERGER@MUNICHRE.COM
Munich Re, Koeniginstr. 107, 80802, Munich, Germany

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

In safety-critical applications, such as medicine and healthcare, decision makers are hesitant to deploy machine learning models unless the expected algorithmic errors are guaranteed to remain within pre-defined tolerances. However, since ML algorithms are statistical in nature, a bounded error cannot be ensured for all possible data inputs. To the contrary, practitioners could be provided with an estimate of the probability the error exceeds the pre-defined tolerance interval. Thus, they will be able to better anticipate high magnitude ML errors and thus manage them more effectively. We refer to this as the risk-assessment problem and propose a novel solution for it. We propose a conformal prediction approach that translates the risk-assessment task into a prediction interval generation problem. The conformal prediction approach results in prediction intervals that are guaranteed to contain the true target variable with a given probability. Using this coverage property, we prove that our risk-assessment approach is conservative i.e., the risk we compute, under weak assumptions, is not lower than the true risk resulting from the ML algorithm. We focus on regression tasks and computationally study, and compare with other related methods, the performance of the proposed method both with and without covariate shift. We find that our method offers superior accuracy while being conservative.

Keywords: risk assessment, conformal prediction, machine learning safety

1. Introduction

Certain safety-critical applications demand tight error tolerances from ML algorithms. Consider healthcare, for settings in which ML algorithms perform dose optimization for chemotherapy and radiotherapy Feng et al. (2018); Huynh et al. (2020); Prinster et al. (2022). An under- or over-dose could severely harm the health of a patient or could even be lethal. Nevertheless, a 5 to 10 percentage deviation between the model's prediction and the true dosage—the absolute error in this case—is considered safe Gurney (2002); Cohen et al.

(1996). Since ML algorithms are statistical in nature, these error tolerances cannot be always guaranteed and a tolerance–violation is possible even for expectionally accurate ML models Hüllermeier and Waegeman (2021). To better prepare for such events, ML model users would like to reserve resources (financial, for instance), making it crucial to estimate the probability with which the error overshoots the tolerated error margin Bertsimas and Orfanoudaki (2021). To this end, we note that a ML model developer does not often have the flexibility to set the tolerance threshold based upon the model's performance. In practice, the model user and the field of application (healthcare in this case) dictates the magnitude of the error tolerance.

We refer to the above problem as *risk-assessment*. Any method that aims to solve the risk-assessment problem needs to satisfy two important properties. Firstly, it should be accurate i.e., the estimated probability of the error overshooting the threshold should be close to its true value. Secondly, it should be conservative i.e., the true probablity of the error overshooting the tolerance shouldn't be larger than its estimated value. Otherwise, the model user would be over-optimistic about the risk it is undertaking—thereby under-allocating resources in case the error tolerances are not met. Note that conservative risk-assessment does not necessarily provide an accurate risk-assessment or vice-versa—assuming that the error overshoots the tolerance with probability one is conservative but (usually) not accurate. Conservative risk-assessment has the caveat of an over-allocation of resources for when a violation of the error tolerances occurs. However, given the financial and potential health harm involved, we suppose that it is better to be over-prepared than under.

1.1. Risk-assessment: problem formulation

We formalize the risk-assessment problem mathematically. Let (X, Y) represent the input-output pair, where $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$. We denote by $\mu(x)$ the machine learning model and by $\mathbb{R}^+ \ni e(x, y) := |\mu(x) - y|$ the prediction error. Let $\tau(x) \in \mathbb{R}^+$ be the pre-defined error tolerance. Our goal is to estimate the probability with which the error overshoots this tolerance. Equivalently, *Risk assessment task:*

$$\text{Given } \tau(X) \text{ find } \alpha : \mathbb{P}(e(X, Y) \geq \tau(X)) \leq \alpha. \tag{1}$$

Let $\mathcal{I}(X)$ denotes a band where the prediction errors are below the pre-defined error tolerance, i.e., $\mathcal{I}(X) = [\mu(X) - \tau(X), \mu(X) + \tau(X)]$. Then the above risk-assessment task can also be expressed as

$$\text{Given } \mathcal{I}(X) \text{ find } \alpha : \mathbb{P}(Y \notin \mathcal{I}(X)) \leq \alpha. \tag{2}$$

Risk assessment shows similar characteristics with the prediction interval (PI) generation task. Owing to this similarity, we will tackle the risk assessment problem by an inverse PI generation method—Figure 1 provides an illustration. Consider a PI generation approach that, given a miss-coverage level α, outputs a prediction interval $\mathcal{T}(X; \alpha)$ with the coverage property $\mathbb{P}(Y \notin \mathcal{T}(X; \alpha)) \leq \alpha$. We observe that an α could be a solution to the risk-assessment problem if $\mathcal{T}(X; \alpha)$ satisfies two properties: a) it is contained inside the interval $\mathcal{I}(X)$; and b) it is the largest possible. The former property ensures a conservative risk-assessment. It follows from the inverse relation between the size of a prediction interval and the miss-coverage level α. Furthermore, the latter ensures accurate risk-assessment, avoiding an overly-conservative estimate.

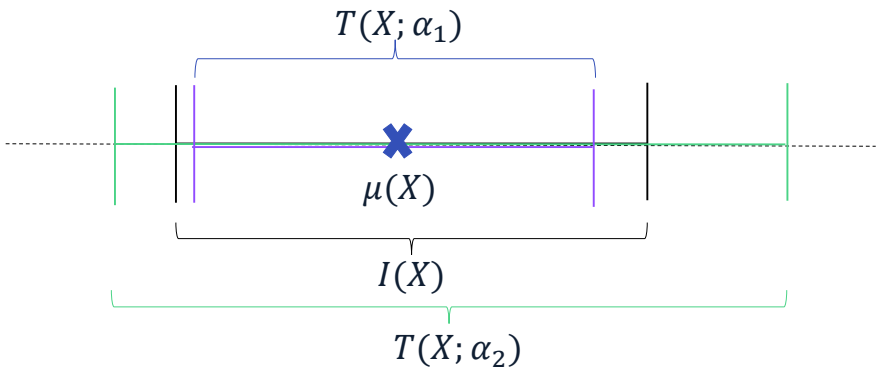


Figure 1: *Intervals visualized at some input $X = x$. α_1 and α_2 are two possible choices for the approximation of the failure probability. $\mathcal{T}(x;\alpha_1)$ is the largest PI contained inside $\mathcal{I}(X)$ whereas, $\mathcal{T}(x;\alpha_2)$ is the smallest PI that contains $\mathcal{I}(X)$. These PIs—depending upon the PI generation technique—might or might not be equal to $\mathcal{I}(X)$. Since $\mathcal{I}(x) \subset \mathcal{T}(x;\alpha_2)$, α_2 is smaller than the failure probability and thus, provides a non-conservative solution to risk-assessment. However, since $\mathcal{I}(x) \supset \mathcal{T}(x;\alpha_1)$, α_1 provides a conservative solution to risk-assessment.*

As for the prediction interval generation technique, we choose conformal prediction (CP) Papadopoulos et al. (2002); Vovk et al. (2005). Compared to other approaches outlined in Subsection 1.2, CP provides the following benefits: a) it is a distribution-free approach and thus, unlike some standard approaches Nix and Weigend (1994); Khosravi et al. (2011b), does not make any assumptions on the functional form of the data distribution Vovk et al. (2005); Shafer and Vovk (2008); b) as we verify later numerically, it provides accurate solutions to the risk-assessment problem; c) owing to its coverage property on the prediction interval Lei et al. (2018); Angelopoulos and Bates (2021), our risk-assessment is conservative and accurate—Section 3 provides further elaboration; d) it extends to problems with covariate shifts Tibshirani et al. (2019); Barber et al. (2022); Prinster et al. (2022); and e) it is model agnostic Papadopoulos et al. (2002); Lei et al. (2015) and therefore, does not require one to change the underlying model architecture. Owing to the aforementioned properties, in the context of regression CP has already been extended to various real-life uses cases—see Auer et al. (2023); Nolte et al. (2024); Bastos (2024) and references therein.

1.2. Previous works

Recall that we seek an α such that the corresponding prediction interval $\mathcal{T}(X;\alpha)$ is the largest possible and is contained inside $\mathcal{I}(X)$. One can use any prediction interval generation technique that efficiently solves this task. Broadly speaking, these methods have two categories. The first category models the distribution for $Y|X = x$. The CDF of the distribution provides the miss-coverage level α for an interval $\mathcal{I}(X)$. One possibility is to model the distribution via a Bayesian neural network or Gaussian Process Regression

David (1992); Williams and Rasmussen (1995). Another possibility is to assume a Gaussian distribution and compute its mean and variance either via samples of the error (collected over a hold-out set) or via additional outputs to a deep neural network Nix and Weigend (1994); Khosravi et al. (2011b). The second category trains a neural network by minimizing a loss function that enforces a small width and a desired coverage on the PIs Pearce et al. (2018); Khosravi et al. (2011a). Note that the loss function depends on α, and thus for our use case, this means minimizing multiple different loss functions until a desirable α is found, which we expect to be prohibitively expensive.

The above approaches either make assumptions on the underlying data distribution or require a substantial change to the predictive model's architecture. A CP-approach is able to address these limitations since it is both distribution-free and model agnostic. To the best of our knowledge, only the authors in Prinster et al. (2022) have considered such an approach to solve the risk-assessment problem. Our article builts upon this previous work but offers three key differences. Firstly, we capture the randomness in the solution to the risk-assessment problem via a hold-out set thereby, providing a lower variance solution. Secondly, while accounting for this randomness, we establish that our method is conservative. Thirdly, for independent and identically distributed (i.i.d) data, we prove the accuracy of our split-CP based method. Lastly, we perform extensive numerical experimentation to provide computational evidence that validate the desired properties of the proposed algorithm.

1.3. Contributions

Following is a summary of our contributions. Firstly, we formalize the risk-assessment problem and propose a framework to solve it for regression problems using CP techniques. Using the coverage property of the CP technique, we prove that our risk-assessment is conservative. Secondly, we capture the randomness in the solution to the risk-assessment task via a hold-out set. Our hold-out set does not require any information on the target variable and thus, could also be generated using generative-ML techniques Goodfellow et al. (2016). Lastly, to assess the accuracy of our risk-assessment algorithm, we conduct a comprehensive set of computational experiments on problems with and without covariate shifts. On a variety of real-life datasets, we compare our method to those outlined above and provide empirical evidence that our method is both the most accurate and conservative.

2. Conformal Prediction (CP)

We briefly summarize CP in the presence of covariate shift. The i.i.d setting is a special case of covariate shift. We consider the latter to be more practically relevant.

2.1. Covariate shift

The covariate shift problem has gained attention within the field of uncertainty quantification in recent studies Tibshirani et al. (2019); Barber et al. (2022); Shimodaira (2000); Sugiyama et al. (2007). This topic has also been highlighted in various ML applications, specifically in the context of healthcare Quinonero-Candela et al. (2008); Ovadia et al. (2019); Ulmer et al. (2020). Under covariate shift, $Y|X = x$ has the same distribution under training and testing.

However, the distribution for X changes under testing. Precisely,

$$\begin{aligned} \text{training:} \quad & (X,Y) \overset{i.i.d}{\sim} P_X \times P_{Y|X}, \\ \text{testing:} \quad & (X,Y) \overset{i.i.d}{\sim} \tilde{P}_X \times P_{Y|X}, \end{aligned} \tag{3}$$

where P_X and $\tilde{P}_X$ are distributions for X under training and testing, respectively. Co-variate shift violates the data exchangeability assumption in standard CP Shafer and Vovk (2008). Nevertheless, under *weight-exchangeability* (see Tibshirani et al. (2019)), CP techniques can be applied again and PIs with coverage guarantees can be recovered even under covariate shift Tibshirani et al. (2019).

Recall that under the exchageability assumption, the random variables $V_1, V_2, \dots V_n$ have a joint distribution $f(V_1, V_2, \dots V_n)$ that is invariant under permutations of these random variables. Furthermore, weight-exchageability means that the joint probability distribution could be factorized as $f(V_1, \dots, V_n) = \Pi_{i=1}^{n} w_i(V_i) g(V_1, \dots, V_n)$, where w_i is a weight function and g is invariant under permutations of the random variables.

If $\tilde{P}_X$ is absolutely continuous with respect to P_X, the data under the covariate shift are weighted exchangeble with the weight functions being the likelihood ratio given as $w(x) = \mathrm{d}\tilde{P}_X(x)/\mathrm{d}P_X(x)$—see Tibshirani et al. (2019). Introducing these weights to the PI computation, leads to a valid coverage for the CP intervals under covariate shift Tibshirani et al. (2019). We present the exact formula below.

2.2. Prediction Intervals

We restrict ourselves to the two common types of weighted conformal prediction methods: weighted split-CP Tibshirani et al. (2019) and JAW Prinster et al. (2022). Further works that explore time series data, group-based distribution shifts and robust validation under distribution shift can be found in Cauchois et al. (2024); Bhattacharyya and Barber (2024); Angelopoulos et al. (2023); Gibbs and Candes (2021).

Weighted Split-CP: Consider a hold-out set $\mathcal{Z} = \{(X_i, Y_i)\}_{i=1,\dots,n}$, which is independent of the training set used for model training. Over the hold-out set $\mathcal{Z}$, we collect samples of the error $e(X_i, Y_i)$ as defined in Equation (1). The PI then follows from including those values of y in the prediction set that result in an error smaller than a given quantile of these error samples.

We represent with δ_x a point mass at value x. We then place point masses at the errors $e(X_i, Y_i)$ and scale them by the weight functions defined as

$$\begin{aligned} p_i^w(x) &= \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j) + w(x)}, \ i = 1, \cdots, n, \\ p_{n+1}^w(x) &= \frac{w(x)}{\sum_{j=1}^{n} w(X_j) + w(x)}, \end{aligned} \tag{4}$$

were $w(X)$ is the likelihood ratio defined earlier. Summing up the weighted point masses, provides the empirical error distribution $\sum_{i=1}^{n} p_i^w(x)\delta_{e(X_i,Y_i)} + p_{n+1}^w(x)\delta_\infty$. Recall that this scaling with weights is what leads to weight-exchangeability and subsequently coverage properties given in Theorem 1. We take the $(1-\alpha)$-th quantile of this error distribution

using the operator $Q^+_{1-\alpha}\{\cdots\}$. The PI for weighted split-CP then reads

$$\mathcal{T}(x;\alpha) = \mu(x) \pm Q^+_{1-\alpha}\{p^w_i(x)\delta_{e(X_i,Y_i)}\}. \tag{5}$$

For notational simplicity, we suppress the dependence of $\mathcal{T}$ on $\mathcal{Z}$. In the special case where $\tilde{P}_X = P_X$, the weights reduce to $p^w_i(x) = p^w_{n+1}(x) = 1/(n+1)$, which leads to the original split-CP method Lei et al. (2018).

JAWs: Instead of a hold-out set, for statistical efficiency, JAWs considers a leave-one-out approach on the training data set—a k-fold cross-validation based methodology is also feasible Barber et al. (2022). Let μ_{-i} represent a model trained on all training points other than the i-th one. Then the error samples are collected via $e_{-i}(X_i, Y_i) = |Y_i - \mu_{-i}(X_i)|$. To derive a PI for JAWs from that of split-CP, we replace μ and e by μ_{-i} and e_{-i}, respectively. This leads to

$$\begin{aligned}\mathcal{T}(x;\alpha) = \Big[&Q^-_\alpha\left\{p^w_i(x)\delta_{\mu_{-i}(x)-e_{-i}(X_i,Y_i)}\right\},\\ &Q^+_{1-\alpha}\left\{p^w_i(x)\delta_{\mu_{-i}(x)+e_{-i}(X_i,Y_i)}\right\}\Big].\end{aligned} \tag{6}$$

Unified notation: For simplicity, we collectively express the above two prediction intervals as

$$\mathcal{T}(x;\alpha) = \left[Q^-_\alpha\left\{p^w_i(x)\delta_{V^-_i(X)}\right\}, Q^+_{1-\alpha}\left\{p^w_i(x)\delta_{V^+_i(X)}\right\}\right], \tag{7}$$

where $V^+_i(X)$ and $V^-_i(X)$ are the upper and lower bounds of intervals that include the error samples, respectively, and read

$$V^\pm_i(X) := \mu_\square(X) \pm e_\square(X_i, Y_i). \tag{8}$$

The function Q^- is the same as Q^+ but with a delta-mass placed at $-\infty$, which ensures that we also consider the lower bounds of the interval expressed via $V^-_i(X)$. The placeholder ($\square$) could either be empty or $-i$ for split-CP and JAW, respectively.

Desirable properties: The following properties are noteworthy. Firstly, for exchangeable datasets $w = 1$, we recover the standard un-weighted split-CP and the Jackknife+ intervals Barber et al. (2021); Papadopoulos et al. (2002). Secondly, the PIs are nested

$$\mathcal{T}(X;\alpha_1) \subseteq \mathcal{T}(X;\alpha_2), \quad \forall \alpha_1 \geq \alpha_2. \tag{9}$$

Lastly, the PIs for both the weighted split-CP and JAW have the coverage property, which we recall below—see Tibshirani et al. (2019); Barber et al. (2022) for further details.

Theorem 1 (Lower-bound) *Tibshirani et al. (2019); Prinster et al. (2022) Under the assumptions: a) data under co-variate shift in the sense of Equation (3); and b)$\tilde{P}_X$ is absolutely continuous with respect to P_X, the prediction interval $\mathcal{T}(X;\alpha)$ resulting from weighted split-CP and JAW satisfy*

$$\mathbb{P}(Y \in \mathcal{T}(X;\alpha)) \geq 1 - c\alpha, \tag{10}$$

where c equals 1 and 2 for split-CP and JAW, respectively.

Theorem 2 (Upper-bound) *Lei et al. (2018) If (X_i, Y_i) are i.i.d and $e(X_i, Y_i)$ have a continuous joint distribution for $i = 1, \cdots, n$, then—in addition to the lower-bound—the split-CP prediction interval $\mathcal{T}(X;\alpha)$ satisfies*

$$\mathbb{P}(Y \in \mathcal{T}(X;\alpha)) \leq 1 - \alpha + 2/(n+2). \tag{11}$$

3. The Inverse Conformal Prediction Algorithm

We propose a CP based approach to solve the risk-assessment problem, and present its theoretical properties. We name our algorithm as InvCP (Inverse Conformal Prediction), as it applies the inverse of CP to compute the coverage level instead of a prediction interval. algorithm 1 summarizes the proposed method that the rest of the section discusses.

3.1. Solution to the risk-assessment problem

Recall that to solve the risk-assessment problem using PIs, for any test input $X = x$ and a calibration set $\mathcal{Z}$, we seek a miscoverage $\alpha(X, \mathcal{Z})$ such that the PI, at the test input, is the largest possible but contained inside the interval $\mathcal{I}(x) = [\mu(x) - \tau(x), \mu(x) + \tau(x)]$. Since the PIs are nested Equation (9), we ensure that the PI is the largest possible by taking a minimum over all α. Enforcing that this largest possible interval is contained inside $\mathcal{I}(X)$ provides

$$\alpha\left(X, \mathcal{Z}\right) := \min_{\alpha'}\{\alpha^{'} : \mathcal{T}(X; \alpha') \subseteq \mathcal{I}(X)\}, \tag{12}$$

We average $\alpha(X, \mathcal{Z})$ over a hold-out set $\mathcal{Z}^{\gamma} := \{X_{i_\gamma}\}_{\{i_\gamma=1,\ldots,m\}}$, and approximate the failure probability $\mathbb{P}(e(X, Y) \geq \tau(X))$ via

$$\alpha_{\mathcal{I}}^{m} := \frac{1}{m} \sum_{X \in \mathcal{Z}^{\gamma}} \alpha(X, \mathcal{Z}). \tag{13}$$

The set $\mathcal{Z}^{\gamma}$ is independent of the calibration set $\mathcal{Z}$ and the training set, and requires only the input information. In the theorem below, we use this independence property to derive a convergence result for α_I^m and we also prove the conservativeness of our method. Later in Subsection 3.2 we discuss how to compute $\alpha(X, \mathcal{Z})$.

Theorem 3 (Conservativeness) *Assume that the prediction interval in the definition for $\alpha(X, \mathcal{Z})$ (given in Equation (12)) has the coverage property as given in Theorem 1. Let $\alpha_{\mathcal{I}}^{m}$ be the estimator defined in Equation (13). Then, the following holds: (i) the probability $\mathbb{P}(Y \notin \mathcal{I}(X))$ has the upper-bound*

$$\mathbb{P}(Y \notin \mathcal{I}(X)) \leq c\mathbb{E}_{\mathcal{Z},X}\left[\alpha_{\mathcal{I}}^{m}\right], \tag{14}$$

where c equals 1 and 2 for weighted split-CP and JAW, respectively; (ii) as $m \to \infty$ and for all $\mathcal{Z}$, the estimator $\alpha_{\mathcal{I}}^{m}$ converges, in probability, to the expected value $\mathbb{E}_X[\alpha(X, \mathcal{Z})]$.

Proof See Appendix A.2. ■

Theorem 4 (Accuracy) *Under the same assumption as in Theorem 2, as the size value of n increases, we have the convergence property*

$$\mathbb{E}_{\mathcal{Z},X}\left[\alpha_{\mathcal{I}}^{m}\right] \overset{n\to\infty}{\longrightarrow} \mathbb{P}(Y \notin \mathcal{I}(X)). \tag{15}$$

Proof See Appendix A.3. ■

Remark 5 (Conservative bound for JAW) *Since we estimate* $\mathbb{P}(Y \notin \mathcal{I}(X))$ *using* $\alpha_{\mathcal{I}}^m$*, ideally, we expect the bound* $\mathbb{P}(Y \notin \mathcal{I}(X)) \leq \mathbb{E}[\alpha_{\mathcal{I}}^m]$*. As Theorem 3 dictates, this is true for split-CP. However, for JAW, we get the conservative bound* $\mathbb{P}(Y \notin \mathcal{I}(X)) \leq 2\mathbb{E}[\alpha_{\mathcal{I}}^m]$*. This is an artifact of the coverage property of JAW (and also Jackknife+), which reads* $\mathbb{P}(Y \in \mathcal{T}(X;\alpha)) \leq 2\alpha$*; see Theorem 1. Nonetheless, in experiments, we observe* $c \approx 1$*, which is aligned with previous studies Prinster et al. (2022); Tibshirani et al. (2019).*

Remark 6 (Connection to previously proposed bounds) *We compare our result to that propsed by Prinster et al. (2022). Firstly, Prinster et al. (2022) provides the bound* $\mathbb{P}(Y \notin \mathcal{I}(X)) \leq c\alpha(X, \mathcal{Z})$*. The latter holds for PIs with conditional coverage, which is usually not the case for CP. Our analysis changes this bound to* $\mathbb{P}(Y \notin \mathcal{I}(X)) \leq c\mathbb{E}_{\mathcal{Z},X}[\alpha(X, \mathcal{Z})]$ *that holds for a CP technique that provides marginal coverage. Secondly, for* $m = 1$*, the estimator* $\alpha_{\mathcal{I}}^m$ *is a high-variance approximation for* $\mathbb{E}_{\mathcal{Z},X}[\alpha_{\mathcal{I}}^m]$ *and is the same as that proposed in Prinster et al. (2022). Therefore, the estimator in Prinster et al. (2022) is a special case of ours. Our estimator better captures the randomness in* $\alpha(X, \mathcal{Z})$ *via a hold-out set.*

3.2. Computation of $\alpha(X, \mathcal{Z})$

To find an $\alpha(X, \mathcal{Z})$ that satisfies Equation (12), we proceed as in Prinster et al. (2022). We find two PIs that have their left and right endpoints, respectively, inside $\mathcal{I}(X)$. We then set $\alpha(X, \mathcal{Z})$ as the maximum of the coverage levels of these PIs. The nested property of the PIs (see Equation (9)) ensures that this solves Equation (12)—Appendix A elaborates further. The details are as follows. Starting with the right endpoint, we find a $\alpha^+(X)$ such that the right endpoint of $\mathcal{T}(X;\alpha^+(X))$ (given by $Q^+_{1-\alpha^+(X)}\left\{p_i^w(x)\delta_{V_i^+(X)}\right\}$) is inside $\mathcal{I}(X)$. This amounts to summing those point masses that lie outside of the right endpoint of $\mathcal{I}(X)$. The final result reads

$$\alpha^+(X) = \sum_{i=1}^{n} p_i^w(X)\mathbb{1}\{\mu(X) + \tau(X) \leq V_i^+(X)\}.$$

We now compute $\alpha^-(X)$ such that the left endpoint of $\mathcal{T}(X;\alpha^+(X))$ (given by $Q^-_{1-\alpha^+(X)}\left\{p_i^w(x)\delta_{V_i^+(X)}\right\}$) is inside $\mathcal{I}(X)$. With a similar computation as above, we find

$$\alpha^-(X) = \sum_{i=1}^{n} p_i^w(X)\mathbb{1}\{V_i^-(X) \leq \mu(X) - \tau(X)\}.$$

We set

$$\alpha\left(X, \mathcal{Z}\right) = \max(\alpha^-(X), \alpha^+(X)). \tag{16}$$

The InvCP algorithm in algorithm 1 summarises our computation of $\alpha_{\mathcal{I}}^m$.

Remark 7 (Simplifications) *InvCP algorithm further simplifies for specific cases. For weighted split-CP intervals—using Equation (7) and Equation (8)—*$\alpha^+(X)$ *simplifies as*

$$\alpha^+(X) = \sum_{i=1}^{n} p_i^w(X)\mathbb{1}\{\tau(X) \leq e(X_i, Y_i)\}.$$

Algorithm 1: Inverse Conformal Prediction (InvCP)

Input: a calibration data set $\mathcal{Z} = \{(X_i, Y_i)\}_{1 \le i \le n}$; α-calibration set $\mathcal{Z}^\gamma = \{X_{i_\gamma}\}_{1 \le i_\gamma \le m}$; a pre-defined interval function $\mathcal{I}(X) = [a_-(X), a_+(X)]$; a weight function p_i^w.

for $i_\gamma = 1$ ***to*** m **do**
  for $i = 1$ ***to*** n **do**
    Compute $V_i^{\pm}(X_{i_\gamma})$ and $p_i^w(X_{i_\gamma})$;
  end
  Compute $\alpha(X_{i_\gamma}, \mathcal{Z})$ using Equation (16);
end

Output: $1 - \alpha_{\mathcal{I}}^m$, where $\alpha_{\mathcal{I}}^m = \frac{1}{m} \sum_{X_{i_\gamma} \in \mathcal{Z}^\gamma} \alpha(X_{i_\gamma}, \mathcal{Z})$.

Similar expression could be obtained for $\alpha^-(X)$. For the unweighted split-CP, $\alpha^+(X)$ further simplifies to $\frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{\tau(X) \le e(X_i, Y_i)\}$. Appendix A provides more details for the case where the tolerance level $\tau(X)$ is fixed for all X values.

The takeaway is that for the above cases, both $\alpha^+(X)$ and $\alpha^-(X)$ can be found by counting the errors in the calibration set that exceed the tolerance level $\tau(X)$ at the input X.

4. Experimental results

We compare the performance of the InvCP approach to other approaches presented in the previous works in solving the risk-assessment problem. We consider both the iid and covariate shift setting. An open source implementation of our work could be found here.

4.1. Risk-assessment methods and datasets

Risk-assessment methods: We consider two deep neural networks, NN-1 and NN-2, which have one and two outputs, respectively. The first output of both approximates the mean of $Y|X = x$. The additional output of NN-2 approximates the variance of the same. The width and depth are chosen such that the training error drops to a sufficiently low value—see Appendix B for further results. We compare the following approaches: CP-S (split CP), CP-CV (CP with cross-validation), CP-SW (weighted CP-S), and CP-CVW (weighted CP-CV). We choose 10-folds for both CP-CV and CP-CVW. For NN-1, we also consider the Res-Gauss method, which fits a Gaussian distribution over samples of the error $e(X, Y)$ collected over a hold-out set $\mathcal{Z}$—see Khosravi et al. (2011b). The miss-coverage results from the CDF of the Gaussian distribution. For NN-2, we consider mean variance estimation (MVE) that assumes a Gaussian distribution for $Y|X = x$ with the mean and the variance resulting from the two outputs of NN-2. The miss-coverage follows from the CDF of this Gaussian—see Nix and Weigend (1994) for further details. Neither Err-Gauss nor MVE are weighted for the covariate shift setting.

Datasets: Following Pearce et al. (2018), we consider five datasets from the UCI repository Dua and Graff (2017): Kin8, Combine Cycle Power Plant, Naval Propulsion, California Housing and Wine Quality White. The error tolerance $\tau(X)$ is set as $\epsilon * \mu(X)$, where $\epsilon \in \mathbb{R}^+$. Similar to Tibshirani et al. (2019), we apply covariate shift via exponential tilting, which resamples the testset with a likelihood ratio of $w(x) \propto \exp(x^T \beta)$. Both β and

Datasets	CP-S	CP-SW	CP-CV	CP-CVW	Res-Gauss
Kin-8 (iid)	0.01	0.01	-0.19	-0.19	**-0.03**
Kin-8 (shift)	0.02	**-0.001**	-0.25	-0.19	0.01
Combine Cycle Power Plant (iid)	**0.001**	0.001	-0.16	-0.16	-0.08
Combine Cycle Power Plant (shift)	-0.05	**-0.01**	-0.32	-0.24	-0.11
Naval Propulsion (iid)	**-0.001**	-0.001	-0.53	-0.53	-0.1
Naval Propulsion (shift)	0.02	**-0.01**	-0.6	-0.5	-0.06
California Housing (iid)	**-0.02**	-0.02	-0.20	-0.20	-0.18
California Housing (shift)	-0.1	**-0.02**	-0.29	-0.21	-0.27
Wine Quality White (iid)	**-0.01**	-0.01	-0.26	-0.26	-0.02
Wine Quality White (shift)	-0.04	**-0.04**	-0.51	-0.32	-0.04

Table 1: Dev of different methods for NN-1. Method with the best accuracy is shown in bold. Recall that Dev ≤ 0 implies that the method is conservative. All the conservative methods are highlighted in blue.

ϵ are chosen such that we have similar magnitude of miss-coverages over all the datasets. Appendix B outlines further details related to the selection of β and ϵ.

Splitting of the dataset: Each dataset is split into three parts: (i) training set over which the model is trained; (ii) calibration set $\mathcal{Z}$ used to compute scores for the split-CP methods; and (iii) calibration set $\mathcal{Z}^\gamma$ over which we compute $\alpha^m_{\mathcal{I}}$ in Equation (13). We shuffle the set $\mathcal{Z} \cup \mathcal{Z}^\gamma$ and perform a random split of $\mathcal{Z}$ and $\mathcal{Z}^\gamma$ repeatedly for 100 iterations. We estimate $\mathbb{E}_{\mathcal{Z},\mathcal{Z}^\gamma}(\alpha^m_{\mathcal{I}})$ by averaging $\alpha^m_{\mathcal{I}}$ over these iterations.

Reference: Since the true failure probability is unavailable, as our reference, we consider an empirical counting-based miss-coverage $\alpha^m_{\mathcal{I},\text{Emp}} := \sum_{X\in\mathcal{Z}^\gamma} \mathbb{1}(Y \notin \mathcal{I}(X))/\text{m}$. To estimate $\mathbb{E}_{\mathcal{Z}^\gamma}(\alpha^m_{\mathcal{I},\text{Emp}})$, we average $\alpha^m_{\mathcal{I},\text{Emp}}$ over the hundred samples of $\mathcal{Z}^\gamma$—see above. We label this approach as Emp. The deviation of a risk-assessment method from the empirical value then reads

$$\text{Dev} := \mathbb{E}_{\mathcal{Z}^\gamma}(\alpha^m_{\mathcal{I},\text{Emp}}) - \mathbb{E}_{\mathcal{Z},\mathcal{Z}^\gamma}(\alpha^m_{\mathcal{I}}). \tag{17}$$

Thus, the accuracy of a risk-assessment method would be |Dev|, and a method would be conservative if

$$\text{Conservativeness: Dev} \leq 0. \tag{18}$$

4.2. Discussion on results

Accuracy comparison for NN-1: Table 1 summarizes the results. Apart from Kin8 (iid), CP-based methods are the most accurate for both iid and shifted datasets. For Kin8 (iid), Res-Gauss provides the best results while being conservative; however, unlike CP-based methods, its conservativeness cannot be theoretically guaranteed. CV-based methods always provide conservative results but are not the most accurate.

Accuracy comparison for NN-2: Table 2 summarize the results. Apart from Kin8 (shift), CP-based methods are the most accurate for both iid and shifted datasets. For Kin8 (shift), MVE provides the best results but is not conservative. Similar to NN-1, CV-based

Datasets	CP-S	CP-SW	CP-CV	CP-CVW	MVE
Kin-8 (iid)	**0.01**	0.01	-0.10	-0.10	0.04
Kin-8 (shift)	-0.02	-0.01	-0.23	-0.15	**0.008**
Combine Cycle Power Plant (iid)	**0.003**	0.003	-0.15	-0.15	-0.06
Combine Cycle Power Plant (shift)	-0.05	**-0.01**	-0.30	-0.19	-0.16
Naval Propulsion (iid)	**-0.001**	-0.001	-0.10	-0.10	-0.15
Naval Propulsion (shift)	0.32	**-0.006**	-0.20	-0.12	0.05
California Housing (iid)	**0.007**	0.007	-0.20	-0.20	-0.05
California Housing (shift)	-0.09	**0.012**	-0.16	-0.22	-0.05
Wine Quality White (iid)	**-0.008**	-0.008	-0.05	-0.05	0.01
Wine Quality White (shift)	0.05	**-0.02**	-0.12	-0.04	0.03

Table 2: Accuracy of different methods for NN-2. Recall that Dev ≤ 0 implies that the method is conservative. All the conservative methods are highlighted in blue.

methods always provide conservative results but are not the most accurate. Although MVE uses the entire training set for mean and variance computation, it could inaccurately estimate the mis-coverage significantly.

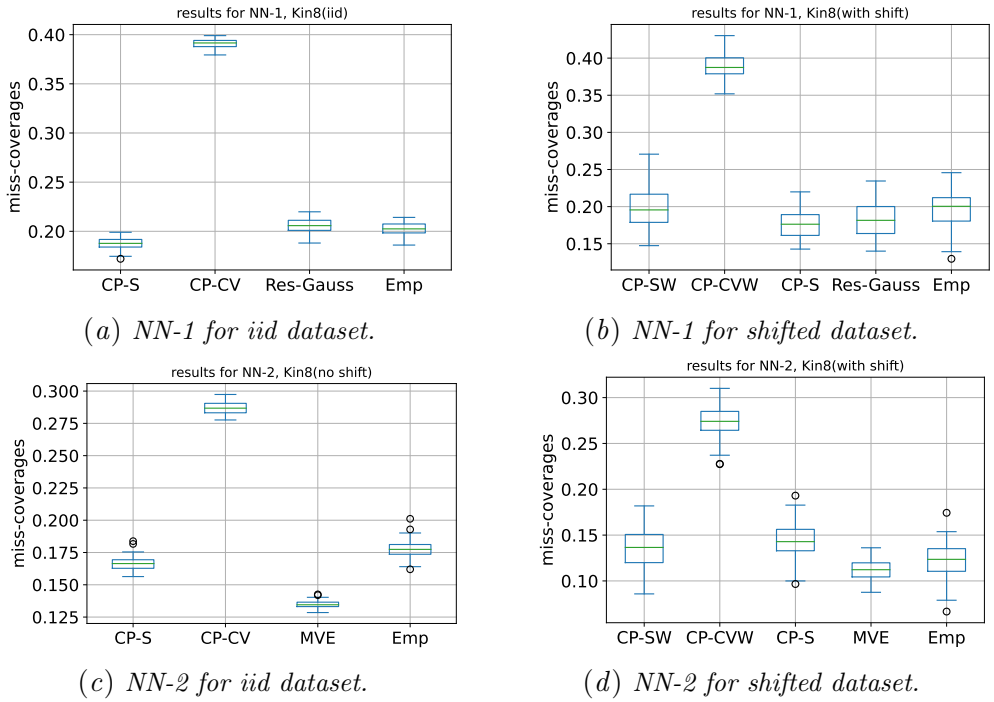
Observe that the unweighted CP-S is conservative and the most accurate for iid dataset. However, for shifted dataset, it losses most of its accuracy and is not conservative. This demonstrates the importance of introducing weights while dealing with covariate shift.

Loss of conservativeness: Table 1 and Table 2 demonstrate that for some datasets, CP-based methods do not maintain their conservativeness. Since for these datasets the miss-coverage from CP methods is very close to that from Emp, we attribute this loss of conservativeness to the finite sample randomness of the hold-out sets $\mathcal{Z}$ and $\mathcal{Z}^\gamma$.

Variation of miss-coverage levels: For the Kin8 dataset and the NN-1 architecture, variation of miss-coverages over the hold-out sets $\mathcal{Z}$ and $\mathcal{Z}^\gamma$ are shown in $2(a)$ and $2(b)$. CV-based methods have the lowest variance however, as explained earlier, they are not the most accurate. The weighted CP methods have a higher variance compared to their non-weighted versions because weights decrease the effective calibration set size—Appendix B provides further elaboration.

Similar results as above but for the NN-2 architecture are shown in $2(c)$ and $2(d)$. Observe that MVE has the lowest variance because to estimate the mean and the variance, it uses the entire training set, which is much larger than the hold-out set $\mathcal{Z}$. However, despite the low variance, it is not the most accurate method. Similar results for other datasets could be found in Appendix B.

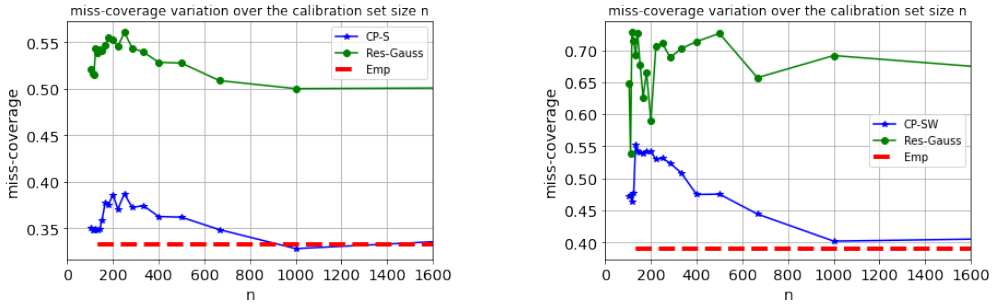
Convergence with n: Both the split-CP and Res-Gauss methods rely on a calibration set. We study the influence of the size of the calibration set over the accuracy of our risk-assessment methods. We consider the California Housing dataset since it has the largest number of datapoints and thus we can choose a large value of m (i.e. the size of the hold-out set $\mathcal{Z}^\gamma$. $3(a)$ and $3(b)$ present the results. Under both iid and shifted data, Res-Gauss's accuracy doesn't improve after a certain value of n. Recall that Res-Gauss assumes a Gaussian distribution for $Y|X = x$, which could be the reason for its limited accuracy. If the underlying data distribution is different from a Gaussian then, irrespective of the value of n, Res-Gauss could only provide limited accuracy. In contrast, CP-S makes no such assumption



(*a*) *NN-1 for iid dataset.* (*b*) *NN-1 for shifted dataset.*

(*c*) *NN-2 for iid dataset.* (*d*) *NN-2 for shifted dataset.*

Figure 2: Miss-coverage variation for Kin8 dataset over the holdout sets ($\mathcal{Z}$ and $\mathcal{Z}^\gamma$).



(*a*) *Results for iid-data.* (*b*) *Results for shifted data.*

Figure 3: Variation of miss-coverage for the California Housing dataset over the size of the calibration set.

on the data distribution, and its accuracy improves as n increases. These findings also validate the convergence results established in Theorem 3 and Theorem 4.

5. Limitations and future work

The current paper proposes the InvCP algorithm, an interval-generation based method for risk assessment of regression models. We consider JAW and weighted split conformal prediction intervals as they provide theoretical coverages under covariate shift. Other conformal prediction intervals such as Conformalized Quantile Regression Romano et al. (2019) or Conformalizing Bayes Angelopoulos and Bates (2021) can also be applied to our framework following the proposed InvCP algorithm. However, their theoretical properties under covariate shift are unclear and require further research. Furthermore, the recent development of adaptive conformal inference under arbitrary distribution shifts Gibbs and Candes (2021) provides the potential of conducting risk assessment in an online fashion. Deriving conservative risk assessment methods under arbitrary shifts is important as many ML models are used in fast-changing areas such as finance and economics, where the market and customers behaviours can shift abruptly. In the future, we aim to extend our algorithm to improve its adaptibility in these more challenging data scenarios.

6. Conclusions

We formalized the problem of risk assessment to estimate the failure probability of a regression model. We showed how conformal prediction-based prediction interval techniques can be used for risk assessment under different data settings—exchangeability and covariate shift. We theoretically prove that the InvCP approach is conservative and we validate our theoretical findings with computational experiments that use deep neural networks.

References

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

Anastasios N. Angelopoulos, Emmanuel J. Candes, and Ryan J. Tibshirani. Conformal pid control for time series prediction. *arXiv preprint arXiv:2307.16895*, 2023.

Andreas Auer, Martin Gauch, Daniel Klotz, and Sepp Hochreiter. Conformal prediction for time series with modern hopfield networks. *arXiv preprint arXiv:2303.12783*, 2023.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49:486–507, 2021.

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

Joao A. Bastos. Conformal prediction of option prices. *Expert Systems with Applications*, 245:123087, 2024. ISSN 0957-4174.

Dimitris Bertsimas and Agni Orfanoudaki. Algorithmic insurance. *arXiv preprint arXiv:2106.00839*, 2021.

Aabesh Bhattacharyya and Rina Foygel Barber. Group-weighted conformal prediction. *arXiv preprint arXiv:2401.17452*, 2024.

Maxime Cauchois, Suyash Gupta, Alnur Ali, and John C. Duchi. Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, 0(0): 1–66, 2024. doi: 10.1080/01621459.2023.2298037.

Michael R Cohen, Roger W Anderson, Richard M Attilio, Laurence Green, Raymond J Muller, and Jane M Pruemer. Preventing medication errors in chemical chemotherapy. *American Journal of Health-System Pharmacy*, 53(7):737–746, 1996.

MacKay David. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992. doi: 10.1162/neco.1992.4.3.448.

Dheeru Dua and Casey Graff. Uci machine learning repository, 2017. http://archive.ics.uci.edu/ml.

Mary Feng, Gilmer Valdes, Nayha Dixit, and Timothy D Solberg. Machine learning in radiation oncology: opportunities, requirements, and needs. *Frontiers in oncology*, 8:110, 2018.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Howard Gurney. How to calculate the dose of chemotherapy. *British journal of cancer*, 86(8): 1297–1302, 2002.

Elizabeth Huynh, Ahmed Hosny, Christian Guthier, Danielle S. Bitterman, Steven E. Petit, Daphne A. Haas-Kogan, Benjamin Kann, Hugo J. W. L. Aerts, and Raymond H. Mak. Artificial intelligence in radiation oncology. *Nature Reviews Clinical Oncology*, 17(12):771–781, December 2020. ISSN 1759-4774. doi: 10.1038/s41571-020-0417-8.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110, 03 2021. doi: 10.1007/s10994-021-05946-3.

Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F. Atiya. Lower upper bound estimation method for construction of neural network-nased prediction intervals. *IEEE Transactions on Neural Networks*, 22(3):337–346, 2011a. doi: 10.1109/TNN.2010.2096824.

Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F Atiya. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on neural networks*, 22(9):1341–1356, 2011b.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2015.

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

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. *IEEE International Conference on Neural Networks*, 1994.

Daniel Nolte, Souparno Ghosh, and Ranadip Pal. Efficient normalized conformal prediction and uncertainty quantification for anti-cancer drug sensitivity prediction with deep regression forests. *arXiv preprint arXiv:2402.14080*, 2024.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. *Lecture notes in computer science*, pages 345–356, 2002.

Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: a distribution-free, ensembled approach. *Proceedings of Machine Learning Research*, 80:4075–4084, 2018.

Drew Prinster, Anqi Liu, and Suchi Saria. Jaws: auditing predictive uncertainty under covariate shift. *Advances in Neural Information Processing Systems*, 35:35907–35920, 2022.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2008.

Sashank J. Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift corection. *AAAI Conference on Artificial Intelligence*, 8, 2015.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Dennis Ulmer, Lotta Meijerink, and Giovanni Cinà. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. *Machine Learning for Health*, pages 341–354, 2020.

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

Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995.

Appendix A. Method

A.1. Why taking the maximum works?

We choose $\alpha(X, \mathcal{Z})$ as the maximum of $\alpha^+(X)$ and $\alpha^-(X)$. Taking the maximum requires justification. The misscoverage $\alpha^+(X)$ ensures that the right endpoint of $\mathcal{T}(X; \alpha^+(X))$ is inside of $\mathcal{I}(X)$. This does not necessarily guarantee that $\mathcal{T}(X; \alpha^+(X)) \subseteq \mathcal{I}(X)$. The left endpoint of $\mathcal{T}(X; \alpha^+(X))$ might still be outside of $\mathcal{I}(X)$—likewise for $\alpha^-(X)$. Taking a maximum works because of the nested property of PIs (see Equation (9)), which provides $\mathcal{T}(X; \alpha(X, \mathcal{Z})) \subseteq \mathcal{T}(X; \alpha^+(X)) \cap \mathcal{T}(X; \alpha^-(X))$. Since the left and right endpoints of $\mathcal{T}(X; \alpha^-(X))$ and $\mathcal{T}(X; \alpha^+(X))$ are already inside $\mathcal{I}(X)$, respectively, we find $\mathcal{T}(X; \alpha(X, \mathcal{Z})) \subseteq \mathcal{I}(X)$.

A.2. Proof for Theorem 3.1

Since $\mathcal{T}(X; \alpha(X, \mathcal{Z}))$ is included in $\mathcal{I}(X)$, we find

$$\mathbb{P}(Y \notin \mathcal{I}(X) | \mathcal{Z} = z, X = x) \leq \mathbb{P}(Y \notin \mathcal{T}(X; \alpha(X, \mathcal{Z})) | \mathcal{Z} = z, X = x). \tag{19}$$

where the randomness is over $Y|X$. Marginalising the above relation with respect to the calibration set $\mathcal{Z}$ and the input X provides

$$\mathbb{P}(Y \notin \mathcal{I}(X)) \leq \mathbb{P}(Y \notin \mathcal{T}(X; \alpha_{\mathcal{I}})) \leq c\alpha_{\mathcal{I}}, \tag{20}$$

where

$$\alpha_{\mathcal{I}} := \mathbb{E}_{\mathcal{Z},X}[\alpha(X, \mathcal{Z})], \tag{21}$$

and the last inequality in the above expression, follows from the coverage property of CP (Theorem 1).

By taking an expectation of Equation (13) with respect to the distribution of X and Z, and applying the definition in Equation (21), we can prove that $\alpha^m_{\mathcal{I}}$ is an unbiased estimator for $\alpha_{\mathcal{I}}$. In Equation (20), replacing $\alpha_{\mathcal{I}}$ by the expected value of its estimator $\alpha^m_{\mathcal{I}}$, we find that

$$\mathbb{P}(Y \notin \mathcal{I}(X)) \leq c\mathbb{E}_{\mathcal{Z},X}\left[\alpha^m_{\mathcal{I}}\right]. \tag{22}$$

Furthermore, from the law of large numbers, as $m \to \infty$ and $\forall \mathcal{Z}$, we find $\alpha^m_{\mathcal{I}} \xrightarrow{P} \mathbb{E}_X[\alpha(X, \mathcal{Z})]$.

A.3. Proof for Theorem 3.2

Now we move to prove the accuracy of split-CP under i.i.d. assumption. For given X, Z, find

$$\tilde{\alpha}(X, \mathcal{Z}) := \max_{\alpha'} \{\alpha' : \mathcal{I}(X) \subseteq \mathcal{T}(X; \alpha')\} \tag{23}$$

Note that under i.i.d data assumption, all the weights $p^w_i(x) = p^w_{n+1}(x) = 1/(n+1)$, and

$$\tilde{\alpha}(X, \mathcal{Z}) \geq \alpha(X, \mathcal{Z}) - 1/(n+1). \tag{24}$$

To see above, let $\alpha_0 = \alpha(X, \mathcal{Z}) - 1/(n+1)$, then $Q^+_{1-\alpha_0}\left\{1/(n+1)\delta_{V_i^-(X)}\right\} > Q^+_{1-\alpha(X,\mathcal{Z})}\left\{1/(n+1)\delta_{V_i^-(X)}\right\}$, hence $\mathcal{T}(X; \alpha_0) \supset \mathcal{T}(X; \alpha(X, \mathcal{Z}))$. As $\mathcal{T}(X; \alpha(X, \mathcal{Z}))$ is the largest prediction interval that is within $\mathcal{I}(X)$, we get $\mathcal{T}(X; \alpha_0) \supseteq \mathcal{I}(X)$. According to the definition of $\tilde{\alpha}(X, \mathcal{Z})$ in Equation (23), we get Equation (24).

Next, by the definition above, we get

$$\mathbb{P}(Y \notin \mathcal{I}(X) | \mathcal{Z} = z, X = x) \geq \mathbb{P}(Y \notin \mathcal{T}(X; \tilde{\alpha}(X, \mathcal{Z})) | \mathcal{Z} = z, X = x). \tag{25}$$

Marginalising the above relation with respect to the calibration set $\mathcal{Z}$ and the input X, and utilizing the probability bound of split-CP, we get

$$\mathbb{P}(Y \notin \mathcal{I}(X)) \geq \mathbb{P}(Y \notin \mathcal{T}(X; \tilde{\alpha}_{\mathcal{I}})) \geq \tilde{\alpha}_{\mathcal{I}} - \frac{2}{n+2} \tag{26}$$

$$\geq \alpha_{\mathcal{I}} - \frac{1}{n+1} - \frac{2}{n+2}, \tag{27}$$

Combing with Equation (20) where $c = 1$ for split-CP, we have that $\alpha_{\mathcal{I}}$ converges to $\mathbb{P}(Y \notin \mathcal{I}(X))$ as $n \to \infty$.

A.4. Algorithm details

Length invariant symmetric intervals: We consider an interval $\mathcal{I}(X)$ of the form $\mathcal{I}(X) = [\mu(X) - \tau, \mu(X) + \tau]$, where $\tau > 0$. This interval is symmetric around $\mu(X)$, and its length doesn't change with X. Furthermore, we assume that the data is exchangeable i.e., there is no co-variate shift and, thus, the likelehood ratio $w = 1$. For such a case, as shown below, the miscoverage $\alpha(X, Z)$ defined in based on a symmetric prediction interval around $\mu(X)$, e.g., Split-CP, is independent of X for any given $\mathcal{Z}$. Consequently, no α-hold-out set is required to collect samples of $\alpha(X, \mathcal{Z})$, i.e. access to a training and calibration set is sufficient.

We further elaborate on the above claim. Under exchangeable data, the weights read $p_i^w(X) = \frac{1}{n+1}$ for all $i \in \{1, \ldots, n\}$. Applying these weights to Equation (3.2), we find

$$\alpha(X, \mathcal{Z}) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{e_i \geq \tau\} \tag{28}$$

where e is the error $e(x, y) = |y - \mu(x)|$. The sum $\frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}\{e_i \geq \tau\}$ is X-independent and hence, for a given calibration set, $\alpha(X, \mathcal{Z})$ is also X-independent.

Remark 8 (Connection to earlier work-continued) *For the case discussed above, since $\alpha(X, Z)$ is X-independent, our estimator is the same estimator as that proposed in Prinster et al. (2022). Note that even for this case, the bound $\mathbb{P}(Y \in \mathcal{I}(X)) \geq 1 - c\alpha(X, \mathcal{Z})$ derived in Prinster et al. (2022) would hold only if the randomness from $\mathcal{Z}$ is ignored.*

Appendix B. Experimental Results

This section provides further details on the experimental results conducted in the main text.

B.1. Details of the datasets

We split each dataset into three parts: training dataset, calibration dataset and test dataset. Training dataset is used to train the ML model. Calibration dataset is used to compute the scores for split-CP methods, and the mean and the variance for Err-Gauss method. Test dataset is the holdout set denoted by $\mathcal{Z}^{\gamma}$ in Equation (13). Table 3 presents the sizes of these different sets.

For covariate shift, we consider exponential tilting where the testing distribution $\tilde{P}_X$ is proportional to $\exp(x^T\beta)P_X$, where P_X is the training distribution. Table 3 presents the values of β for different datasets. The test set under covariate shift is sampled (with replacement) from the original test set by weighting the probabilities with $\exp(x^T\beta)$. Under co-variate shift, the effective number of calibration points change to (see Reddi et al. (2015))

$$n_{eff} = \left(\sum_{i=1}^{n} |w(X_i)|\right)^2 / \left(\sum_{i=1}^{n} |w(X_i)|^2\right).$$

Therefore, while applying the unweighted risk-assessment methods (CP-S, Res-Gauss and MVE) to the covariate shift setting—for a fair comparison with weighted techniques—we reduce the number of calibration points to n_{eff}. Table 3 presents the different values of n_{eff}.

As our interval $\mathcal{I}(X)$, we set the tolerance $\tau(X)$ to $\tau(X) = \epsilon * \mu(X)$. Table 4 presents the values of ϵ chosen. These values were choosen such that we get similar miss-coverage values (in order or magnitude) over different datasets.

Datasets	Total Points	Features	train/calib/test set size	β	n_{eff}
K8	8912	8	5734/1000/1458	$[-1, 0, \ldots, 0, 4]$	81
CCPP	9568	4	6698/1000/1870	$[-1, 0, \ldots, 0, 6]$	45
NP	11934	16	8354/1000/2580	$[-1, 0, \ldots, 0, 6]$	101
CH	20640	8	14448/1000/5192	$[-1, 0, \ldots, 1, 0]$	246
WW	4898	11	1469/500/2929	$[-1, 0, \ldots, 1, 0]$	42

Table 3: *Dataset information—abbreviations defined in Subsection 4.1*

Datasets	ϵ
K8	0.15
CCPP	0.01
NP	0.05
CH	0.2
WW	0.15

Table 4: *Error tolerance over datasets—abbreviations defined in Subsection 4.1*

B.2. Details of the models

Recall that we have two models NN-1 and NN-2, which have one and two outputs, respectively. The first output of both approximates the mean of $Y|X = x$. The additional output of NN-2 approximates the variance of the same. For all datasets except Naval Propulsion, NN1 and NN2 have two hidden layers with 64 and 16 units, respectively. For Naval Propulsion, to better approximate the data, we added another hidden layer with 64 units. Both the models are trained using an Adam's optimizer with a learning rate of 0.01 and a batch-size of 64. Log of the MSE error for a first few epochs and for NN1 is shown in Figure 4. Similar training curves were obtained for NN2.

B.3. Results over additional datasets

We present the results for data sets Combine Cycle Power Plant (CCPP), California Housing (CH), Naval Propulsion (NP) and Wine Quality White (WW) in this section.

For CCPP data set, the results under iid and shift data settings using model NN-1 and NN-2 are shown in Figure 6(a)-6(d). In all experiments, all methods generate conservative mis-coverage estimates, while CP-S and CP-SW are the most accurate methods. CP-CV and CP-CVW have smaller variances comparing with the split-CP methods.

For California Housing data set, the results are shown in Figure 7(a)-7(d). CP-S and CP-SW are the most accurate methods in the i.i.d and shifted data sets, respectively. The unweighted method CP-S provides non-conservative results on the shifted data. Res-Gauss method has higher variance and less accuracy. MVE has smaller variance, but it cannot maintain conservative.

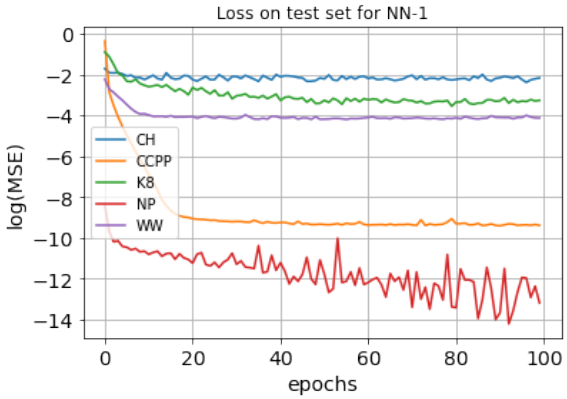
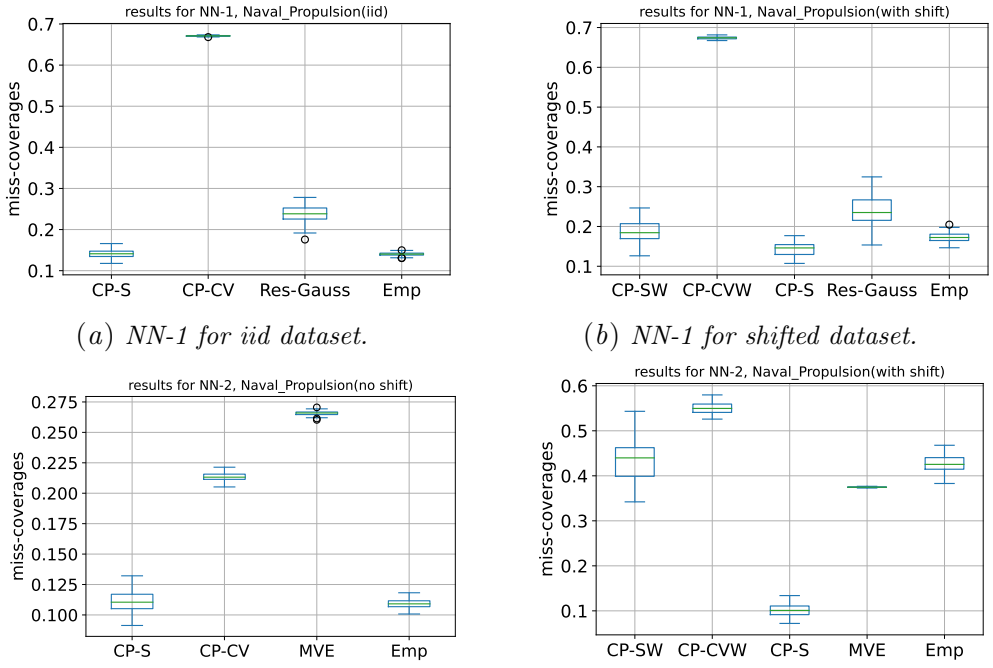


Figure 4: Log(MSE) on test set for NN-1.

For Wine Quality White data set, the results are shown in Figure 8(a)-8(d). For NN-1 model, CP-S/CP-SW and Res-Gauss provide comparable accurate results. For NN-2 model, only CP based methods provide conservative estimates on the shifted dataset.



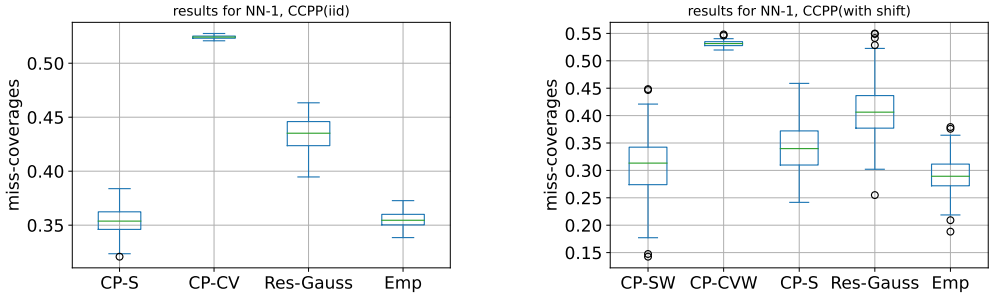
(a) *NN-1 for iid dataset.*

(b) *NN-1 for shifted dataset.*

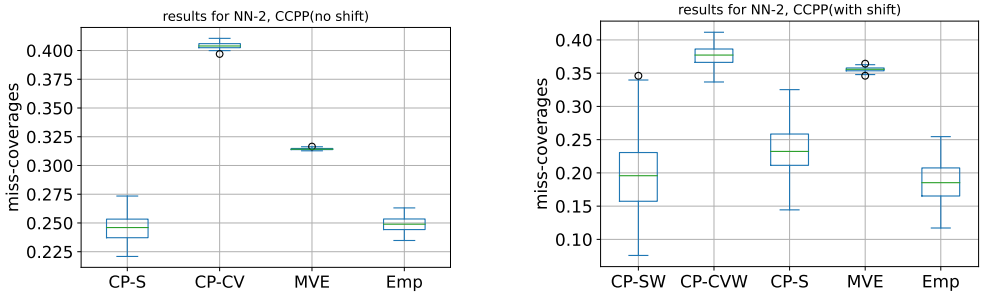
(c) *NN-2 for iid dataset.*

(d) *NN-2 for shifted dataset.*

Figure 5: Miss-coverage variation for Naval Propulsion dataset over the holdout sets ($\mathcal{Z}$ and $\mathcal{Z}^{\gamma}$).

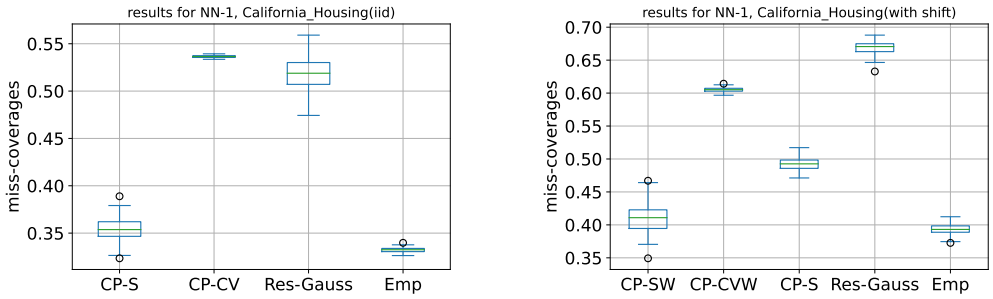


(a) *NN-1 for iid dataset.* (b) *NN-1 for shifted dataset.*

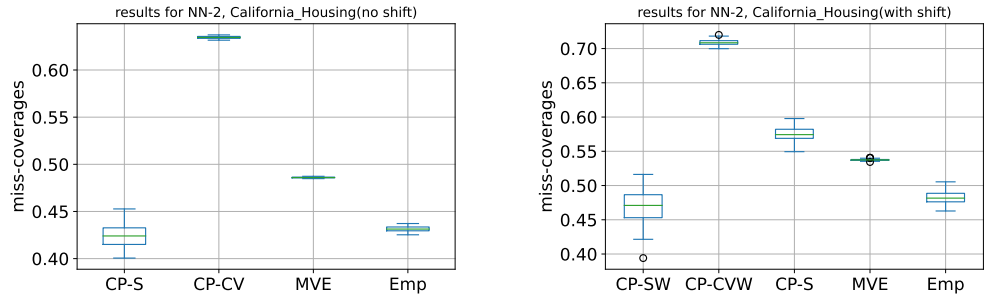


(c) *NN-2 for iid dataset.* (d) *NN-2 for shifted dataset.*

Figure 6: Miss-coverage variation for Combusion Cycle Power Plant dataset over the holdout sets ($\mathcal{Z}$ and $\mathcal{Z}^\gamma$).

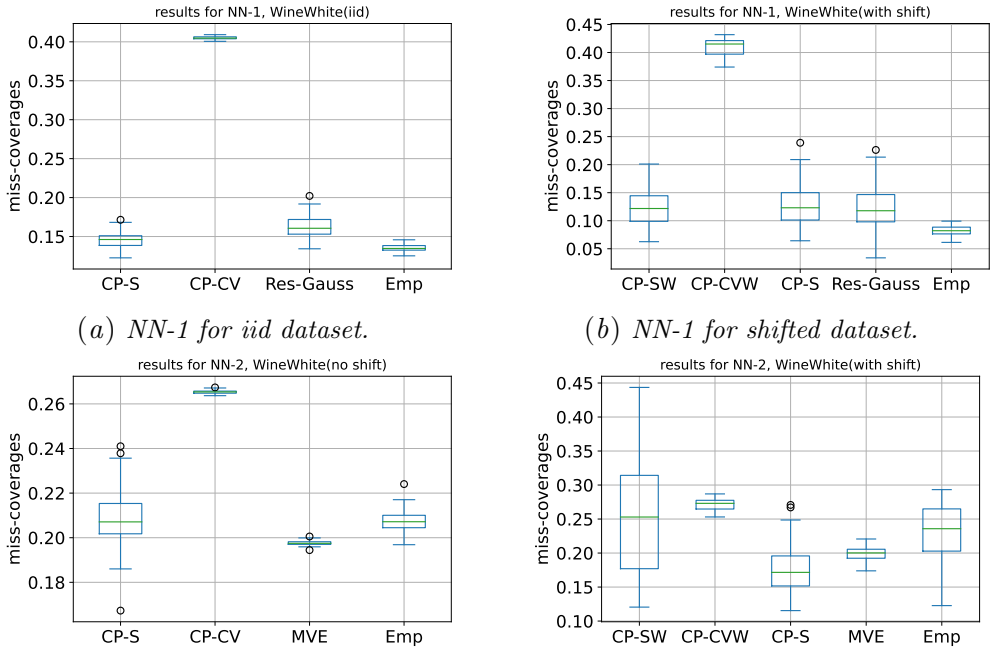


(a) *NN-1 for iid dataset.* (b) *NN-1 for shifted dataset.*



(c) *NN-2 for iid dataset.* (d) *NN-2 for shifted dataset.*

Figure 7: Miss-coverage variation for California Housing dataset over the holdout sets ($\mathcal{Z}$ and $\mathcal{Z}^\gamma$).



(a) *NN-1 for iid dataset.*

(b) *NN-1 for shifted dataset.*

(c) *NN-2 for iid dataset.*

(d) *NN-2 for shifted dataset.*

Figure 8: Miss-coverage variation for Wine Quality White dataset over the holdout sets ($\mathcal{Z}$ and $\mathcal{Z}^\gamma$).