

On Training-Conditional Conformal Prediction and Binomial Proportion Confidence Intervals

Anonymous authors

Paper under double-blind review

Abstract

Estimating the expectation of a Bernoulli random variable based on N independent trials is a classical problem in statistics, typically addressed using Binomial Proportion Confidence Intervals (BPCI). In the control systems community, many critical tasks—such as certifying the statistical safety of dynamical systems—can be formulated as BPCI problems.

Conformal Prediction (CP), a distribution-free technique for uncertainty quantification, has gained significant attention in recent years and has been applied to various control systems problems, particularly to address uncertainties in learned dynamics or controllers. A variant known as training-conditional CP was recently employed to tackle the problem of safety certification.

In this note, we highlight that the use of training-conditional CP in this context does not provide valid safety guarantees. We demonstrate why CP is unsuitable for BPCI problems and argue that traditional BPCI methods are better suited for statistical safety certification.

1 Introduction

Uncertainty quantification is a critical aspect in fields where predictions influence safety and performance guarantees, such as in control systems. Probabilistic guarantees, including those derived from the theory of Probably Approximately Correct (PAC) learning, play an important role in providing bounds on the accuracy of predictions under limited training data.

Conformal Prediction (CP) is one of the approaches that has gained visibility due to its ability to provide valid prediction sets without requiring strong distributional assumptions. A distinctive characteristic of CP is that, rather than providing a point prediction of the variable of interest, it provides set predictions with a valid bound on the probability that the predicted set contains the true variable [Vovk et al. \(2005\)](#). This technical note focuses on a specific formulation of CP known as training-conditional CP [Vovk \(2012\)](#). However, existing applications in areas such as safety verification for dynamical systems have shown limitations in the interpretation of these guarantees. In particular, recent works [Chilakamarri et al. \(2024\)](#); [Lin & Bansal \(2024\)](#); [Vincent et al. \(2024\)](#) have applied training-conditional CP to safety verification problems in control systems. While promising, these applications have misinterpreted the implications of CP’s set prediction framework, especially in cases where the underlying data can be modeled as Bernoulli random variables. This paper aims to rigorously analyze these limitations and provide an alternative framework for interpreting PAC-based guarantees in such contexts. In [Section 2](#) we recall existing methods for estimating the expectation of a Bernoulli random variable. In [Section 3](#) we introduce the formalism of training-conditional CP, followed by a detailed analysis of its PAC guarantees. In [Section 4](#) we present a special case of interest, where the nonconformity measure corresponds to an indicator function, leading to Bernoulli-distributed conformity scores. We demonstrate that the PAC guarantees derived from this setting are unsuitable for estimating the expectation of a Bernoulli random variable.

2 Binomial Proportion Confidence Intervals

Consider a setting where there are $N + 1$ i.i.d. Bernoulli random variables (r.v.) $R_1, R_2, \dots, R_N, R_{N+1}$ with parameter b , i.e. $R_i \sim \text{Bern}_b$ and $\Pr_{\text{Bern}_b}(R_i = 1) \doteq b$. Given a realization of the first N r.v., the problem is to estimate an interval of values for the probability that the $N + 1$ -th variable will be equal to 1, or in other words we want to estimate the parameter b . This is a very well studied problem and it is known in the literature under the name of Binomial Proportion Confidence Intervals (BPCI), see [Dean & Pagano \(2015\)](#) for a survey. We give below a quick overview of the setting.

Define the new r.v. $Y \doteq \sum_{i=1}^N R_i$. It is well known that Y has a binomial distribution $Y \sim \text{Bin}(N, b)$, defined by $\Pr_{\text{Bin}(N, b)}(Y = y) \doteq \binom{N}{y} b^y (1 - b)^{N-y}$ for $y \in \mathbb{Z}_{[0, N]}$, where $\mathbb{Z}_{[0, N]}$ denotes the integers $0, 1, \dots, N$. Let $\check{b} : \mathbb{Z}_{[0, N]} \rightarrow [0, 1]$ and $\hat{b} : \mathbb{Z}_{[0, N]} \rightarrow [0, 1]$ be two random variables serving as interval estimators. The coverage probability of the interval estimator $[\check{b}, \hat{b}]$ for $Y \sim \text{Bin}(N, b)$ is defined as

$$\rho(b, \check{b}, \hat{b}) \doteq \Pr_{\text{Bin}(N, b)}(\check{b}(Y) \leq b \leq \hat{b}(Y)). \quad (1)$$

In the expression above b is fixed and it's the true parameter of the binomial distribution describing Y . Note that \check{b} and \hat{b} are a transformation of the same random variable Y . This expression can also be rewritten equivalently as

$$\rho(b, \check{b}, \hat{b}) = \sum_{y \in I} \Pr_{\text{Bin}(N, b)}(Y = y),$$

where $I \doteq \{y \in \mathbb{Z}_{[0, N]} : \check{b}(y) \leq b \leq \hat{b}(y)\}$. For $\alpha \in (0, 1)$ an interval estimator $[\check{b}, \hat{b}]$ is a *conservatively valid* (sometimes also called 'exact' or 'secure') $1 - \alpha$ confidence interval if the coverage probability $\rho(b, \check{b}, \hat{b})$ is greater or equal to $1 - \alpha$ for all the values of b . An example of a conservatively valid interval estimator is given by the Clopper-Pearson method [Clopper & Pearson \(1934\)](#), see [Dean & Pagano \(2015\)](#) for more estimators.

Before concluding this section, we rewrite (1) in an equivalent form that is more commonly found in the literature on Probably Approximately Correct (PAC) bounds. First, note that $\Pr_{\text{Bern}_b}(R_{N+1} = 1) = b$. Second, since Y is a r.v. obtained as a transformation of the i.i.d. Bernoulli random variables R_1, \dots, R_N , the probability of any event $M \subseteq \mathbb{Z}_{[0, N]}$, $\Pr_{\text{Bin}(N, b)}(Y \in M)$ can be equivalently described by $\Pr_{\text{Bern}_b}^N(\{(r_1, \dots, r_N) : \sum_{i=1}^N r_i \in M\})$, where $\Pr_{\text{Bern}_b}^N$ is the product probability measure induced by the N i.i.d. Bernoulli random variables. Hence, we rewrite (1) as

$$\Pr_{\text{Bin}(N, b)}(\check{b}(Y) \leq b \leq \hat{b}(Y)) = \Pr_{\text{Bern}_b}^N \left(\check{b} \left(\sum_{i=1}^N R_i \right) \leq \Pr_{\text{Bern}_b}(R_{N+1} = 1) \leq \hat{b} \left(\sum_{i=1}^N R_i \right) \right). \quad (2)$$

We will use this form of the coverage probability to draw a comparison with the guarantees given by training-conditional CP.

3 Training-conditional Conformal Prediction

In this section we introduce the basic concepts of training-conditional CP [Vovk \(2012\)](#). Let $(\mathbf{Z}, \mathcal{F}, \mathbb{P})$ be a probability space, and consider $L + 1$ i.i.d. random variables (r.v.) $Z'_1, \dots, Z'_M, Z_1, \dots, Z_N$ and Z_{N+1} with $L = N + M$. Let Z'_i for $i = 1, \dots, M$ be the *training set* and Z_i for $i = 1, \dots, N$ be the *calibration set*. Note that Z_{N+1} is not part of either set. We use the lower case of a r.v. to denote a realization. Our considerations hold also for the case where $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ where \mathbf{X} and \mathbf{Y} represent a measurable feature space and label space respectively and each $z \in \mathbf{Z}$ may be written as $z = (x, y)$ where $x \in \mathbf{X}$ is some feature and $y \in \mathbf{Y}$ a label. For clarity we omit the exact structure of \mathbf{Z} . An *Inductive Nonconformity M-measure* (INM) is a measurable function $A : \mathbf{Z}^M \times \mathbf{Z} \rightarrow \mathbb{R}$. While no additional requirements are needed for A , intuitively an effective INM will assign a high real number to any element in \mathbf{Z} that does not conform to a training set (in \mathbf{Z}^M). An *Inductive Nonconformal Predictor* (INP) is a set predictor defined as

$$\Gamma^\epsilon(z_1, \dots, z_N, z'_1, \dots, z'_M) \doteq \{z \in \mathbf{Z} : p^z > \epsilon\}, \quad (3)$$

where $\epsilon \in [0, 1]$ is the *significance level*, the *p*-values are defined as

$$p^z \doteq \frac{|\{i : R_i \geq R^z\}| + 1}{N + 1}, \quad (4)$$

and

$$R_i \doteq A((z'_1, \dots, z'_M), z_i), \quad R^z \doteq A((z'_1, \dots, z'_M), z), \quad (5)$$

are the *nonconformity scores*.

In the following, when it is clear from the context we omit the arguments of the INP and write Γ^ϵ instead of $\Gamma^\epsilon(z_1, \dots, z_N, z'_1, \dots, z'_M)$. Intuitively, z belongs to the INP Γ^ϵ if there are strictly more than $\lfloor \epsilon(N + 1) - 1 \rfloor$ elements R_i in the calibration set with a higher (worse) or equal nonconformity score than R^z . It is easy to see that $\epsilon' < \epsilon''$ implies that $\Gamma^{\epsilon''} \subseteq \Gamma^{\epsilon'}$.

Theorem 1 (Vovk (2012)) Choose $\epsilon, E \in [0, 1]$, fix the training set $Z'_1 = z'_1, \dots, Z'_M = z'_M$, and consider the event

$$S_E \doteq \{(z_1, \dots, z_N) \in \mathbf{Z}^N : \mathbb{P}(Z_{N+1} \in \Gamma^\epsilon(z_1, \dots, z_N, z'_1, \dots, z'_M)) \geq 1 - E\} \quad (6)$$

in the σ -algebra \mathcal{F}^N of the product probability space $(\mathbf{Z}^N, \mathcal{F}^N, \mathbb{P}^N)$. It holds that

$$\mathbb{P}^N(S_E) \geq 1 - \delta, \quad (7)$$

where $\delta \doteq \text{Bin}_{N,E}(\lfloor \epsilon(N + 1) - 1 \rfloor)$ is the cumulative binomial distribution with N trials and probability of success E .

The quantities δ and $1 - E$ are sometimes referred to as the *confidence* and *coverage probability* (which is not the coverage probability mentioned in Sec. 2). Theorem 1 is to be understood in the following way. Given two values ϵ and E , for the given training set, the event S_E is the subset of \mathbf{Z}^N containing all the tuples (z_1, \dots, z_N) such that the INP Γ^ϵ contains a realization of Z_{N+1} with probability at least $1 - E$, or, in other words, Γ^ϵ returns a subset of \mathbf{Z} of measure at least $1 - E$. By (7) the measure of this set of tuples S_E is at least $1 - \delta$, where δ depends on ϵ , β and N . This form of guarantees where a double layer of nested probabilities is present is called Probably Approximately Correct (PAC). Moreover, this is a distribution-free result, that is, it holds independently of \mathbb{P} . Finally, we remark that ϵ and E are chosen *a priori*, that is they can not be defined as random variables depending on a realization of the calibration set.

4 A Special Case of Interest

In this section we draw a parallel between the BPCI and training-conditional CP and show the fundamental difference between the two approaches.

Let the INM be an indicator function for the set $Q \subset \mathbf{Z}$, that is

$$A((z'_1, \dots, z'_M), z) \doteq \begin{cases} 1 & \text{if } z \in Q, \\ 0 & \text{if } z \in \bar{Q}. \end{cases} \quad (8)$$

Typically Q depends on z'_1, \dots, z'_M ; for instance, in the case of binary classification problems the training set may be used to train a parametrized function to assign one of the two labels to all $z \in Q$, as in Support Vector Machines. We omit this dependency here, since Theorem 1 assumes that the training set is given. Hence, for a fixed training set the nonconformity scores of the calibration set are i.i.d. Bernoulli distributed with parameter b , i.e. $R_i \sim \text{Bern}_b$. Is it possible to derive a conservatively valid confidence interval for the parameter b from the calibration set? We proceed with an example.

Example 1 Suppose that the calibration set has size 2, i.e. $N = 2$. Up to reindexing, there are three distinct outcomes.

Case 1: With probability $(1 - b)^2$ we have $z_1, z_2 \notin Q$, resulting in nonconformity scores $R_1 = R_2 = 0$. We construct the prediction set Γ^ϵ following its definition (3). For all z in Q we have that $R^z = 1$, resulting in

$|\{i \leq 2 : R_i \geq R^z\}| = 0$ and $p^z = \frac{1}{3}$, meaning z has the highest (worst) nonconformity score. Whereas for all z in \bar{Q} we have that $R^z = 0$ resulting in $p^z = 1$. The inclusion of z in the predicted set Γ^ϵ depends on the significance level ϵ . We distinguish two situations, either $\epsilon \in [\frac{1}{3}, 1)$ or $\epsilon \in [0, \frac{1}{3})$. If $\epsilon \in [\frac{1}{3}, 1)$ then any $z \in Q$ is excluded from Γ^ϵ since $\frac{1}{3} \leq \epsilon$. On the other hand, any $z \in \bar{Q}$ is included in Γ^ϵ , since $1 > \epsilon$. To summarize if $\epsilon \in [\frac{1}{3}, 1)$ then $\Gamma^\epsilon = \bar{Q}$. If $\epsilon \in [0, \frac{1}{3})$ then $\Gamma^\epsilon = \mathbf{Z}$, since any $z \in Q \cup \bar{Q} = \mathbf{Z}$ has a sufficiently high p -value.

Case 2: With probability $2b(1-b)$ we have $(z_1 \in Q \wedge z_2 \in \bar{Q})$ or $(z_2 \in Q \wedge z_1 \in \bar{Q})$ hence $R_1 \cup R_2 = \{0, 1\}$. If $R^z = 1$ then $p^z = \frac{2}{3}$, whereas if $R^z = 0$ then $p^z = 1$. We conclude that if $\epsilon \in [0, \frac{2}{3})$ then $\Gamma^\epsilon = \mathbf{Z}$, if $\epsilon \in [\frac{2}{3}, 1)$ then $\Gamma^\epsilon = \bar{Q}$.

Case 3: With probability b^2 we have $z_1, z_2 \in Q$ and $R_1 = R_2 = 1$. If $R^z = 1$ then $p^z = 1$, but also if $R^z = 0$ then $p^z = 1$. Then for any significance level $\epsilon \in [0, 1)$ it holds $\Gamma^\epsilon = \mathbf{Z}$.

Now, fix $E \in [0, 1]$ and consider any $\epsilon \in [\frac{2}{3}, 1)$. A straightforward application of Theorem 1 tells us that

$$\mathbb{P}^2(S_E) \geq E^2. \quad (9)$$

Let us further inspect the above result. For the chosen significance level $\epsilon \in [\frac{2}{3}, 1)$, $\Gamma^\epsilon = \mathbf{Z}$ with probability b^2 (from Case 3) and $\Gamma^\epsilon = \bar{Q}$ with probability $1 - b^2$ (from Case 1 and 2)¹. Theorem 1 is a distribution-free result and as such it holds for all values of b . We distinguish two situations, $b \leq E$ and $b > E$:

1) Suppose that $b \leq E$ (and hence $1 - b \geq 1 - E$). If $z_1, z_2 \in Q$ (Case 3) we have that $\mathbb{P}(Z_{N+1} \in \Gamma^\epsilon) = \mathbb{P}(\mathbf{Z}) = 1 \geq 1 - E$, hence $Q \times Q \subseteq S_E$. If at least one of z_1 and z_2 belongs to \bar{Q} (Case 1 and 2) we have that $\mathbb{P}(Z_{N+1} \in \Gamma^\epsilon) = \mathbb{P}(\bar{Q}) = 1 - b \geq 1 - E$, hence $\bar{Q} \times \bar{Q} \subseteq S_E$, which means that $S_E = \mathbf{Z}^2$. Trivially, $\mathbb{P}^2(S_E) = \mathbb{P}^2(\mathbf{Z}^2) = 1 \geq E^2$.

2) On the other hand, suppose that $b > E$ (and hence $1 - b < 1 - E$). If $z_1, z_2 \in Q$ (Case 3), as before, $\mathbb{P}(Z_{N+1} \in \Gamma^\epsilon) = \mathbb{P}(\mathbf{Z}) = 1 \geq 1 - E$, and once again $Q \times Q \subseteq S_E$. If at least one of z_1 and z_2 belongs to \bar{Q} (Case 1 and 2) then $\mathbb{P}(Z_{N+1} \in \Gamma^\epsilon) = \mathbb{P}(\bar{Q}) = 1 - b < 1 - E$, hence such z_1 and z_2 can't belong to S_E by definition. Then $\mathbb{P}^2(S_E) = \mathbb{P}^2(Q \times Q) = b^2 \geq E^2$.

Theorem 1 holds for both cases $b \leq E$ and $b > E$, since $\mathbb{P}^2(\mathbf{Z}^2) = 1 \geq E^2$ and $\mathbb{P}^2(Q \times Q) = b^2 \geq E^2$ respectively. Assume that the true value of b is greater than value of E that we fixed (and smaller than 1) and that the calibration set returns $R_1 = 0$ and $R_2 = 1$. What can we say about b ?

For the given calibration set and significance level the INP predicts $\Gamma^\epsilon = \bar{Q}$, hence it is tempting to say that $\mathbb{P}^2(\mathbb{P}(\bar{Q}) \geq 1 - E) = \mathbb{P}^2(1 - b \geq 1 - E) \geq E^2$, or equivalently $\mathbb{P}^2(b \leq E) \geq E^2$: recalling (2), we may conclude that $[0, E]$ is a E^2 confidence interval for b . But this is clearly not true: since we assumed that $b > E$ the interval $[0, E]$ will never contain the parameter b (note that none of the arguments of $\mathbb{P}^2(\cdot)$ depends on (z_1, z_2) in the preceding statement, unlike (2)). We conclude from this example that this is not a viable path to obtain a PAC bound for b comparable to (2).

The example above leads us to the following remark and main message of this technical note.

Remark 1 Theorem 1 is a statement on the correctness of the set predictor Γ^ϵ . Adopting the frequentist point of view, it is a statement on how often the set predictor Γ^ϵ constructed from N samples attains the desired coverage level $1 - E$ for a new realization of Z_{N+1} . In other words, since we don't know the true value b , if $b > E$ the INP is correct (it attains the desired coverage level) only when $\Gamma^\epsilon = \mathbf{Z}$ (but this is uninformative), and it is incorrect when $\Gamma^\epsilon = \bar{Q}$; if instead $b \leq E$, the INP is always correct. Since b is precisely the quantity we are looking to estimate, and therefore unknown, this is not a valid reasoning to estimate it. See the appendix for a graphical representation.

Remark 1 can be extended to any situation where the nonconformity score can take only a finite set of values (as the binary case), therefore defining a classification problem. Essentially training-conditional CP provides a methodology to construct a set predictor, guaranteed to be attain the desired coverage level with a minimum confidence. The predictor can output tight sets (corresponding to a low number of classes) for

¹If Γ^ϵ predicts \bar{Q} it implies that the nonconformity score of Z_{N+1} is predicted to be 0, whereas if it predicts \mathbf{Z} then all we know is that the nonconformity score of Z_{N+1} is predicted to be in $\{0, 1\}$ which is not useful.

‘good’ calibration data and loose sets (corresponding to a high number of classes) for ‘poor’ calibration data; on average the probability that the calibration data returns a predictor attaining the coverage level of $1 - E$ is not smaller than $1 - \delta$. Depending on the parameters’ choice ϵ and E , we have shown that the $1 - \delta$ confidence may be achieved simply by choosing the entire sample space (corresponding to all the classes) sufficiently often, see also Figure 1. However, this does not provide useful information on the probability of seeing one specific class, which is instead what (2), and in general BPCI methods.

4.1 A Note on Safety Verification for Dynamical Systems

Recent papers have used training-conditional conformal prediction, particularly Theorem 1, to provide PAC guarantees on the safety of control systems with neural network-based controllers [Chilakamarri et al. \(2024\)](#); [Lin & Bansal \(2024\)](#), and more broadly, on the safety of autonomous systems [Vincent et al. \(2024\)](#). In this section we show that these papers follow the reasoning outlined in Section 4, and are therefore incorrect. Below, we follow the notation used in [Lin & Bansal \(2024\)](#), but the same applies to the other works.

Consider a dynamical system defined by $\dot{x} = f(x)$ where $x \in X \subseteq \mathbb{R}^n$, a fixed time horizon $T \in \mathbb{R}_{>0}$. Denote by $\xi_x(\tau)$ for $\tau \in [0, T]$ the state trajectory of the system at time τ when initialized at x (for simplicity we assume that the solution to the differential equation exists and is unique)². Let $X_A \subset X$ represent a set of undesirable states, and consider the cost function defined as

$$J(x) \doteq \min_{\tau \in [0, T]} d(\xi_x(\tau)),$$

where $d : X \rightarrow \mathbb{R}$ is such that $d(x) \leq \gamma \iff x \in X_A$ and $d(x) > \gamma \iff x \in X \setminus X_A$ with $\gamma \in \mathbb{R}$. An instructive example for the discussion is below is to choose $\gamma = 0$ and $d : X \rightarrow \{0, 1\}$, with $d = 0 \iff x \in X_A$ and $d = 1 \iff x \in X \setminus X_A$, but the same applies for any different choice. In other words, J assigns a positive real number to a point $x \in X$ if and only if the state trajectory from x never intersects with X_A . Let $(X, \mathcal{F}, \mathbb{P})$ be a probability space. To quantify the safety of this system in probabilistic terms, it is of interest estimating the probability of drawing a state x such that $J(x) > 0$. In other words, we want to estimate $\mathbb{P}(\{x : J(x) > 0\})$. Let x_1, \dots, x_N be N values in X : in [Lin & Bansal \(2024\)](#) the authors define the nonconformity score as $R_i \doteq J(x_i)$ for $i = 1, \dots, N$, and are therefore interested in estimating $\mathbb{P}(\{x : R^x > 0\})$. However, this is equivalent to defining a nonconformity measure as

$$A(x) \doteq \begin{cases} 1 & \text{if } x \in X_A, \\ 0 & \text{if } x \in X \setminus X_A, \end{cases} \quad (10)$$

and we have shown that this line of reasoning is not suitable for estimating the parameter b of a Bernoulli r.v. $R_i \sim \text{Bern}_b$ given N i.i.d. realizations of it. The authors in [Chilakamarri et al. \(2024\)](#) rely on the framework developed by [Lin & Bansal \(2024\)](#), and therefore incur in the same problem. Finally, in ([Vincent et al., 2024](#), Theorem 1) the authors re-derive Theorem 1, originally derived by Vovk in [Vovk \(2012\)](#). They explicitly make the claim that training-conditional CP reduces to the Clopper-Pearson confidence interval if the underlying i.i.d. random variables are Bernoulli distributed (see Sec. Proofs-D therein), yet we have disproved their claim.

5 Conclusion

In this note we examined existing methodologies to use training-conditional CP for statistical safety verification, a problem that can be reduced to estimating the expectation of a Bernoulli random variable. While training-conditional CP remains a powerful tool for uncertainty quantification we have shown that it is not appropriate for BPCI problems. Specifically, we discussed the correct interpretation of the guarantees for confidence intervals and the PAC-style guarantees of training-conditional CP. We do not rule out the possibility that a different formulation of CP could be applied to BPCI problems. This is left for future work.

²In the original paper the trajectory ξ depends on a learned controller and depends on a training set Z'_1, \dots, Z'_M . For clarity we omit this dependence here, since the training set is given and is fixed.

References

- Vamsi Krishna Chilakamarri, Zeyuan Feng, and Somil Bansal. Reachability analysis for black-box dynamical systems. *arXiv preprint arXiv:2410.07796*, 2024.
- Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.
- Natalie Dean and Marcello Pagano. Evaluating confidence interval methods for binomial proportions in clustered surveys. *Journal of Survey Statistics and Methodology*, 3(4):484–503, 2015.
- Albert Lin and Somil Bansal. Verification of neural reachable tubes via scenario optimization and conformal prediction. In *6th Annual Learning for Dynamics & Control Conference*, pp. 719–731. PMLR, 2024.
- Joseph A Vincent, Aaron O Feldman, and Mac Schwager. Guarantees on robot system performance using stochastic simulation rollouts. *IEEE Transactions on Robotics*, 2024.
- Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

A Appendix

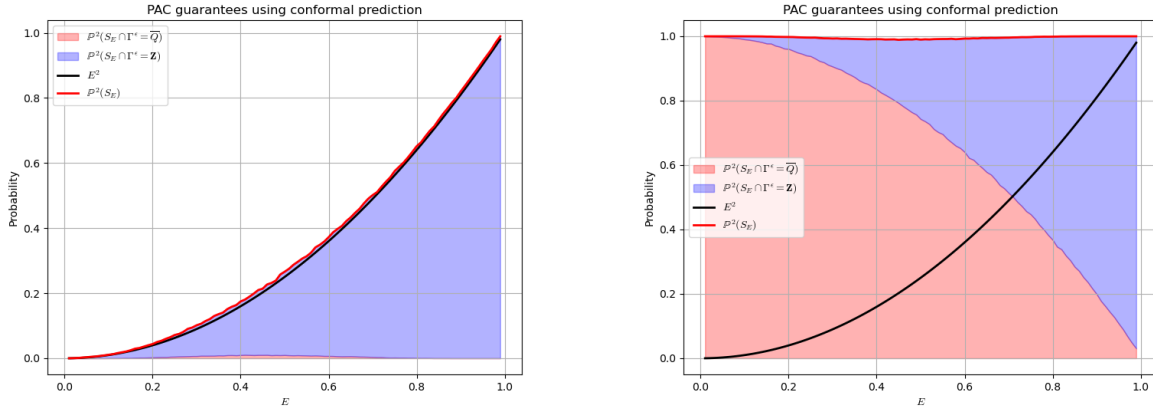


Figure 1: On the left the curves resulting from $b_1 \leq E$, on the right the curves resulting from $b_2 > E$, for different values of E .

We validate empirically (9) as follows, see also Figure 1. We define a list of values for E by $E_q = 0.01 + 0.01 * q$ for $q = 0, \dots, 98$. For every value of E_q we consider an underlying Bernoulli distribution with parameter $b_{1,q} = E_q - \alpha E_q < E_q$ (left figure) and an underlying Bernoulli distribution with parameter $b_{2,q} = E + \alpha E_q > E_q$ (right figure) with $\alpha = 0.005$. The significance level ϵ is set to $2/3$. We draw $n_{\text{cal}} = 5 \cdot 10^4$ pairs of calibration points $\{z_1^{(i)}, z_2^{(i)}\}_{i=1}^{n_{\text{cal}}}$. For every pair of calibration points $z_1^{(i)}, z_2^{(i)}$ we construct the resulting INP as $\Gamma_{(i)}^\epsilon \doteq \Gamma^\epsilon(\dots, z_1^{(i)}, z_2^{(i)})$, draw $n_{\text{test}} = 5 \cdot 10^4$ test points $\{z_{N+1}^{(j)}\}_{j=1}^{n_{\text{test}}}$ and compute the empirical frequency $\hat{g}_i = \frac{|\{j=1, \dots, n_{\text{test}} : z_{N+1}^{(j)} \in \Gamma_{(i)}^\epsilon\}|}{n_{\text{test}}}$ as an approximation for $\mathbb{P}(Z_{N+1} \in \Gamma_{(i)}^\epsilon)$; finally we compute $\hat{h} = \frac{|\{i=1, \dots, n_{\text{cal}} : \hat{g}_i \geq 1-E\}|}{n_{\text{cal}}}$ as an approximation to $\mathbb{P}^2(S_E)$ shown in the plots as the solid red line. The solid black line represents the curve given by E^2 , which remains always below the red line in both plots, as expected. The area shaded in blue represents the fraction of the \hat{g}_i 's for which the INP $\Gamma_{(i)}^\epsilon$ is equal to

\mathbf{Z} , whereas the area shaded in red represents the fraction of the \hat{g}_i 's for which the INP $\Gamma_{(i)}^\epsilon$ is equal to \overline{Q} and \hat{g}_i is greater or equal than $1 - E$. It is visible in the left plot that the only reason why the solid red line (approximating $\mathbb{P}^2(S_E) = b_1^2$) is above E^2 is that the INP is allowed to predict the entire set \mathbf{Z} . In contrast, on the right the solid red line approximates $\mathbb{P}^2(S_E) = 1$ since any pair of $z_1^{(i)}, z_2^{(i)}$ results in a prediction $\Gamma_{(i)}^\epsilon$ satisfying $\mathbb{P}(Z_{N+1} \in \Gamma_{(i)}^\epsilon) \geq 1 - E$; accordingly, for a fixed E the area shaded in red covers approximately $1 - b_2^2$ of the 'Probability' axis and the area shaded in blue approximately b_2^2 .