UPPSALA
UNIVERSITET

IT mDA 25 021

Degree project 30 credits

July, 2025

Uncertainty Quantification for Receiver Operating Characteristic Curves

Wenhan Zhou

Master's Programme in Data Science

UPPSALA
UNIVERSITET

Uncertainty Quantification for Receiver Operating Characteristic Curves

Wenhan Zhou

Abstract

Quantifying uncertainty is fundamental to reliable inference in statistical learning, especially in binary classification where models estimate probabilities for class membership. This thesis investigates the construction of Confidence Bands (CBs) for Receiver Operating Characteristic (ROC) curves, which are widely used to evaluate classifier performance. We establish a mathematical connection between ROC curves and conditional Cumulative Distribution Functions (CDFs), enabling the problem of ROC curve uncertainty quantification to be addressed through established statistical techniques for CDFs.

We develop and compare three approaches for constructing CBs for ROC curves: bootstrap CBs, Beta-Binomial CBs, and Monte Carlo CBs. The bootstrap method generates empirical CDFs through resampling, providing consistency with classical concentration inequalities but is limited by sampling bias in the calibration set. The Beta-Binomial approach adopts a Bayesian framework, constructing posterior distributions over conditional CDFs by incorporating confusion matrix elements, thus reducing sampling bias but introducing sensitivity to prior choice. Our primary contribution, the Monte Carlo CBs method, leverages the Probability Integral Transform and properties of order statistics to provide distribution-free, finite-sample guarantees for the uncertainty in ROC curves, overcoming both sampling bias and reliance on asymptotic approximations.

We formally analyze the theoretical properties of these methods, demonstrating that Monte Carlo CBs uniquely satisfy conditional coverage, finite-sample validity, distribution-free behavior, and variance adaptivity. Empirical validation on both synthetic and real-world datasets confirms that Monte Carlo CBs maintain their theoretical guarantees across varying sample sizes and significance levels. This work advances the statistical theory of uncertainty quantification for functional objects and provides practical tools for robust ROC curve analysis in binary classification.

Faculty of Science and Technology

Uppsala University, Place of publication Uppsala

Supervisor: Antonio Horta Ribeiro Subject reader: Dave Zachariah

Examiner: Filip Malmberg

Contents

List of Figures

1 Introduction

In statistical learning, quantifying uncertainty is crucial for reliable decision-making. This is particularly important in binary classification tasks, where models estimate the probability of an instance belonging to a specific class. However, these probability estimates are often unreliable, a model output of $\hat{p} = 0.7$ cannot necessarily be interpreted as the instance having a 70% chance of belonging to the positive class [GPSW17].

Evaluating binary classifiers often relies on the ROC curve, which plots the TPR against the FPR at various threshold settings. A key insight, and one central to this thesis, is that both TPR and FPR can be represented as CDFs, see Section 2. This mathematical framework allows us to approach the problem of ROC curve uncertainty quantification through the lens of CDF uncertainty estimation, a well-established area in statistical theory [Mas90, BK89, DW22, SP16, CB12].

1.1 Aim of the Thesis

The primary aim of this thesis is to investigate, develop, and compare statistical techniques for quantifying uncertainty in CDFs. This exploration is particularly motivated by the fact that the ROC curves can be represented as conditional CDFs, see Section 2. By focusing on the uncertainty quantification of CDFs, we aim to build a principled foundation for understanding the reliability of ROC curve estimates.

The central goal of this work is:

> **Developing statistical techniques that can provide theoretically sound, practically applicable, and robust uncertainty quantification methods for ROC curves.**

To address this, we will explore methods for constructing CBs for ROC curves by leveraging uncertainty quantification techniques for their underlying CDF representations. The methods developed or analyzed in this thesis are intended to satisfy several desirable properties, ensuring that the resulting uncertainty estimates are both meaningful and trustworthy:

- **Conditional Coverage**: Any new ROC curve from a test set lies within the CBs with high probability.
- **Finite Sample Guarantee**: The CBs should be valid for any finite sample size of the calibration set, including very small where existing methods fail.

- **Distribution-Free**: The CBs should not rely on strong distributional or model assumptions. Furthermore, they should be invariant to the specific datasets and classifier.
- **Variance-adaptive**: The width of the CBs should adapt to the uncertainty of the ROC curve for all thresholds.

1.2 Challenges in CDF Uncertainty Quantification

Despite the importance of uncertainty quantification for CDFs in general and ROC curves in particular, constructing CBs that satisfy the properties mentioned above remains challenging. Several fundamental difficulties explain this research gap:

- **Functional objects**: Both CDFs and ROC curves are functional objects rather than point estimates, requiring methods that can provide uniform guarantees across the entire domain.
- **Functional uncertainty with single realization**: ROC curves represent classifier performance over an entire dataset rather than individual predictions, and each dataset produces only one ROC curve. This creates the challenge of quantifying uncertainty about a functional object with effectively one observation, making standard uncertainty quantification methods inadequate.
- **Distributional assumptions**: Traditional statistical approaches often rely on strong distributional assumptions that may not hold in the general case.
- **Finite sample size**: The number of samples available for calibration is often limited, making it challenging to rely on asymptotic properties.

These challenges have limited the development of rigorous methods for constructing CBs with well-defined statistical properties.

1.3 Applications in Healthcare

The theoretical framework developed in this thesis has significant practical applications, particularly in healthcare. For example, in cardiovascular disease detection, which caused an estimated 17.9 million deaths in 2019, representing 32% of all global deaths [Wor24]. Machine learning models are increasingly applied to ECG data to predict patient outcomes [GOD+23, LRP+21, Hab22, GGL+22].

In such high-stakes clinical applications, different points on the ROC curve represent different trade-offs between sensitivity (detecting all positive cases) and specificity (avoiding false alarms), which directly translate to trade-offs between medical costs and patient risks. Reliable uncertainty quantification through CBs can help clinicians make more informed decisions about these trade-offs.

1.4 Methodology and Approach

To overcome the challenges in ROC curve uncertainty quantification, we explore five techniques: **conformal prediction**, **bootstrap**, **Bayesian approaches**, **concentration inequalities**, and **GoF tests**. Crucially, these methods operate as post-processing steps, relying solely on the model's predicted probabilities $\hat{p}_i$ and the corresponding true labels y_i, obtained from a calibration set, without requiring refitting the original classifier.

By leveraging the CDF representation of ROC curves, we can apply established statistical techniques for CDF uncertainty quantification to address our specific problem.

1.5 Contributions

The thesis makes several contributions:

- It identified the uncertainty quantification of ROC curves as a problem of CDF uncertainty quantification, see Section 2.
- It explores the possibility of using a Bayesian framework to construct CBs for the ROC curve, see Section 4.3.2.
- It improves upon previous state-of-the-art methods in GoF tests for constructing CBs for CDFs [DW22, SP16] by achieving conditional coverage and finite sample guarantee, see Section 4.3.3.

This work provides not only theoretical advancements in uncertainty quantification for binary classification but also practical tools that can be applied in various domains where reliable decision-making under uncertainty is crucial. As mentioned previously, the methods developed are general and can be applied to any domain where binary classification and uncertainty quantification for CDFs are relevant.

For readers interested in the code, it is available at https://github.com/SunAndClouds/masters-thesis.git.

2 Background

This section provides the necessary mathematical background that serve as building blocks for our methodology. We begin with foundational concepts of binary classification performance metrics, followed by an exploration of uncertainty quantification approaches that will be applied to ROC curves.

2.1 Cumulative Distribution Functions

The concept of a CDF is fundamental to probability theory and statistics, providing a complete description of the probability distribution of a random variable. This section formally defines the true CDF, its empirical counterpart derived from data, and explores the relationships between ordered random variables and these CDFs.

2.1.1 The True CDF

Let X be a random variable, assumed to be drawn from a specific, underlying (but possibly unknown) probability distribution. Its true CDF, denoted as $F(t)$, is defined as the theoretical probability that X takes on a value less than or equal to t:

$$F(t) = \mathbb{P}(X \leq t).$$

This function $F(t)$ fully characterizes the probability distribution of X. It is a theoretical construct representing this underlying probability law. In practical scenarios, $F(t)$ is often unknown and needs to be estimated from observed data. A CDF $F(t)$ possesses the following key mathematical properties:

- **Non-decreasing**: For any $t_1 < t_2$, $F(t_1) \leq F(t_2)$.
- **Right-continuous**: For any t, $\lim_{h \to 0^+} F(t + h) = F(t)$.
- **Limits**: $\lim_{t \to -\infty} F(t) = 0$ and $\lim_{t \to \infty} F(t) = 1$.

2.1.2 The Empirical CDF

Given i.i.d. samples $\{X_1, \ldots, X_n\}$ drawn from the distribution whose true CDF is $F(t)$, the ECDF, denoted as $F_n(t)$, is defined as the proportion of observations in the sample

that are less than or equal to t:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[X_i \leq t],$$

where $\mathbb{1}[\cdot]$ is the indicator function. The ECDF $F_n(t)$ is a piece-wise constant function that serves as a non-parametric estimator of the unknown true CDF $F(t)$. The Glivenko-Cantelli theorem, states that $F_n(t)$ converges uniformly to $F(t)$ almost surely as the sample size $n \to \infty$. That is,

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{a.s.} 0 \quad \text{as} \quad n \to \infty.$$

We can view the ECDF $F_n(t)$ from a single dataset as one instance from the many possible ECDFs that could be computed from samples of size n. While the true CDF $F(t)$ represents the expected value of such ECDFs, any individual ECDF $F_n(t)$ provides a single estimate. The Glivenko-Cantelli theorem assures us that this single estimate becomes increasingly accurate as the sample size n grows.

2.1.3 Order Statistics

When analyzing samples from a distribution, it is often useful to consider the **order statistics**, which are the sample values arranged in ascending order. Given our sample $\{X_1, X_2, \ldots, X_n\}$, the ordered values are denoted as $\{X_{(1)}, X_{(2)}, \ldots, X_{(n)}\}$, where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$.

Order statistics provide critical information about the sample distribution, including extremes, quantiles, and the overall shape. The i-th order statistic $X_{(i)}$ represents the i-th smallest value in the sample, and its properties are essential for both theoretical and practical considerations.

2.1.4 Relations Between Order Statistics and CDFs

We now explore the connections between order statistics and CDFs.

Relation to the True CDF: The PIT is a result in probability theory that connects random variables to uniform distributions through their CDFs. If X is a continuous random variable with a strictly increasing CDF $F(t)$, then the random variable $Y = F(X)$ follows a standard uniform distribution.

Formally, for any $u \in [0, 1]$:

$$\mathbb{P}(F(X) \leq u) = \mathbb{P}(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u$$

which is precisely the CDF of a uniform distribution on the interval between 0 and 1.

When applied to order statistics, the PIT provides a crucial theoretical foundation. Consider a set of n i.i.d. samples $\{X_1, X_2, \ldots, X_n\}$ from a continuous distribution with CDF F. If we apply the PIT to each observation, defining $Y_i = F(X_i)$, then $\{Y_1, Y_2, \ldots, Y_n\}$ are i.i.d. uniform random variables on the interval between 0 and 1.

When these transformed uniform variables are arranged in ascending order, the resulting order statistics $\{Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}\}$, where $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$, follow a Beta distribution [Run12]:

$$Y_{(i)} \sim \text{Beta}(i, n - i + 1).$$

Relation to the Empirical CDF: Using order statistics, the ECDF can be expressed in a piecewise constant manner:

$$F_n(t) = \begin{cases} 0, & \text{if } t < X_{(1)} \\ \frac{i}{n}, & \text{if } X_{(i)} \leq t < X_{(i+1)} \text{ for } i = 1, 2, \ldots, n-1 \\ 1, & \text{if } t \geq X_{(n)} \end{cases}$$

This formulation highlights a crucial property of the ECDF: it is a step function that "jumps" by $\frac{1}{n}$ at each observed data point $X_{(i)}$. These jumps occur precisely at the order statistics of the sample, making the values $F_n(X_{(i)}) = \frac{i}{n}$ for each order statistic. The ECDF thus explicitly depends on the order statistics, which define both the locations of the jumps in the function and the values of the function at those points.

2.1.5 Adaptive Thresholds and Fixed Grids

When working with ECDFs, there are two main approaches to choosing evaluation points, each with distinct advantages:

Adaptive thresholds approach: The ECDF is evaluated at the observed order statistics $\{X_{(1)}, X_{(2)}, \ldots, X_{(n)}\}$. This approach captures exactly where the "jumps" in the CDF occur and is particularly useful for discrete or highly skewed distributions. The CDF values at these points are simply $F_n(X_{(i)}) = \frac{i}{n}$. Alternatively, we can plot the ECDF using the following formula:

$$F_n(X_{(i)}) = \left(X_{(i)}, \frac{i}{n} \right)$$

for all $i = 1, 2, \ldots, n$. This approach is accurate as it measures the exact points where the CDF changes its value. However, it might introduce comparison challenges as each ECDF will be evalauted at different values.

Fixed grid approach: The ECDF is evaluated at predefined, equally spaced points $\mathcal{T} = \{t_1, t_2, \ldots, t_T\}$ that form a uniform grid over some interval. The primary motivation for this approach is ensuring consistent dimensionality across different samples. This is benificial when employing bootstrap, as the resampled dataset will likely contain duplicates, resulting in fewer unique threshold values if using adaptive thresholds. This inconsistency hinders computing statistics (like variance or quantiles) across bootstrap samples, as it requires evaluation at identical threshold points. Another advantage of fixed grids is that they ensure coverage over the entire theoretical domain, which adaptive thresholds might miss if extreme values are not present in a particular sample.

2.2 Confusion Matrix Elements

Let $\mathcal{D} = \{(\hat{p}_i, y_i)\}_{i=1}^n$ be a dataset where $\hat{p}_i$ represents the predicted probabilities from a claffier and $y_i \in \{0, 1\}$ denotes the binary labels. To obtain binary predictions, we apply a threshold $t \in [0, 1]$ such that the predicted label is $\hat{y}_i = \mathbb{1}[\hat{p}_i > t]$.

The performance of a binary classifier is typically evaluated using elements of the confusion matrix. The TPs count instances correctly predicted as positive, defined as:

$$\text{TP}(t) = \sum_{i=1}^{n} \mathbb{1}[\hat{p}_i > t] \cdot \mathbb{1}[y_i = 1].$$

Similarly, FPs count instances incorrectly predicted as positive:

$$\text{FP}(t) = \sum_{i=1}^{n} \mathbb{1}[\hat{p}_i > t] \cdot \mathbb{1}[y_i = 0].$$

FNs represent instances incorrectly predicted as negative:

$$\text{FN}(t) = \sum_{i=1}^{n} \mathbb{1}[\hat{p}_i \leq t] \cdot \mathbb{1}[y_i = 1].$$

Finally, TNs count instances correctly predicted as negative:

$$\text{TN}(t) = \sum_{i=1}^{n} \mathbb{1}[\hat{p}_i \leq t] \cdot \mathbb{1}[y_i = 0].$$

These confusion matrix elements can be elegantly expressed in terms of ECDFs. Let $F_{n_1}(t)$ and $F_{n_0}(t)$ be the conditional ECDFs of the predicted probabilities for positive

and negative instances, respectively:

$$F_{n_1}(t) = \frac{1}{n_1} \sum_{i=1}^{n} \mathbb{1}[\hat{p}_i \leq t] \cdot \mathbb{1}[y_i = 1]$$
$$F_{n_0}(t) = \frac{1}{n_0} \sum_{i=1}^{n} \mathbb{1}[\hat{p}_i \leq t] \cdot \mathbb{1}[y_i = 0]$$

where $n_1 = \sum_{i=1}^{n} \mathbb{1}[y_i = 1]$ and $n_0 = \sum_{i=1}^{n} \mathbb{1}[y_i = 0]$ are the numbers of positive and negative instances, respectively. Using these ECDFs, we can rewrite the confusion matrix elements as:

$$\begin{aligned} \text{TP}(t) &= n_1 \cdot (1 - F_{n_1}(t)) \\ \text{FP}(t) &= n_0 \cdot (1 - F_{n_0}(t)) \\ \text{FN}(t) &= n_1 \cdot F_{n_1}(t) \\ \text{TN}(t) &= n_0 \cdot F_{n_0}(t) \end{aligned}$$

From these elements, we derive two critical performance metrics. The TPR, represents the proportion of actual positives correctly identified:

$$\text{TPR}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)} = \frac{n_1 \cdot (1 - F_{n_1}(t))}{n_1} = 1 - F_{n_1}(t)$$

The FPR, represents the proportion of actual negatives incorrectly classified as positives:

$$\text{FPR}(t) = \frac{\text{FP}(t)}{\text{FP}(t) + \text{TN}(t)} = \frac{n_0 \cdot (1 - F_{n_0}(t))}{n_0} = 1 - F_{n_0}(t).$$

In the following section, we will see that the ROC curve can be formulated in terms of these conditional ECDFs.

2.3 ROC Curve

ROC curves provide a graphical representation of a binary classifier's performance across various threshold settings. They can be formulated in terms of conditional ECDFs, establishing a strong theoretical connection between classifier evaluation and statistical theory.

Let $F_{n_1}(t)$ and $F_{n_0}(t)$ be the ECDFs of the predicted probabilities for positive and negative instances, respectively, as defined earlier. The ROC curve can then be defined as a parametric curve through a functional relationship with these ECDFs:

$$\text{ROC}(t) = (1 - F_{n_0}(t), 1 - F_{n_1}(t)),$$

for all $t \in [0, 1]$. An alternative formulation can be obatined by a variable substitution $t' = 1 - F_{n_0}(t)$, yielding:

$$\text{ROC}(t) = 1 - F_{n_1}\left(F_{n_0}^{-1}(1-t)\right).$$

Each point on the ROC curve corresponds to a specific threshold value, with the curve typically starting at the point (0,0) for the strictest threshold ($t = 1$) and ending at (1,1) for the most lenient threshold ($t = 0$).

2.3.1 Properties of ROC Curves

A key insight into ROC curves is that they satisfy the defining properties of a CDF, albeit in a two-dimensional parametric representation. The ROC curve plots the TPR, $\text{TPR}(t) = 1 - F_{n_1}(t)$, against the FPR, $\text{FPR}(t) = 1 - F_{n_0}(t)$, as a threshold t varies over its range. The ROC curve exibits properties similar to a standard CDF:

1. **Boundedness:** Since F_{n_0} and F_{n_1} are CDFs, their values lie in $[0, 1]$. Consequently, both FPR and TPR are also bounded within $[0, 1]$. The ROC curve is therefore confined to the unit square $[0, 1] \times [0, 1]$.

2. **Monotonicity (Non-decreasing):** As the threshold t decreases, $F_{n_0}(t)$ and $F_{n_1}(t)$ are non-decreasing. This implies that $\text{FPR}(t)$ and $\text{TPR}(t)$ are non-decreasing functions of decreasing t. When plotting TPR versus FPR, this ensures that the curve is non-decreasing; moving rightward (increasing FPR) never results in moving downward (decreasing TPR).

3. **Limits/Endpoints:** As $t \to 1$ (strictest threshold), $F_{n_0}(t) \to 0$ and $F_{n_1}(t) \to 0$. Thus, $\text{FPR}(t) \to 1$ and $\text{TPR}(t) \to 1$. As $t \to 0$ (most lenient threshold), $F_{n_0}(t) \to 1$ and $F_{n_1}(t) \to 1$. Thus, $\text{FPR}(t) \to 0$ and $\text{TPR}(t) \to 0$.

4. **Continuity:** The continuity of the ROC curve path depends directly on the continuity of the underlying score distributions F_{n_0} and F_{n_1}. If they are continuous, the ROC curve is continuous. If they are discrete or empirical, the ROC curve is piecewise constant, analogous to ECDFs being step functions (and right-continuous).

These characteristics demonstrate that an ROC curve, representing the trade-off between TPR and FPR, behaves mathematically like CDFs conditioned on the labels.

2.4 DKW Inequality

The DKW inequality, in its tight form established by Massart [Mas90], states that the probability that the ECDF deviates from the true CDF by less than δ at any point is at least $1 - \varepsilon$:

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \leq \delta_{1-\varepsilon}\right) \geq 1 - \varepsilon,$$

where $\delta_{1-\varepsilon}$ is given by:

$$\delta_{1-\varepsilon} = \sqrt{\frac{\log(2/\varepsilon)}{2n}}.$$

The corresponding CB is given by:

$$[L_n(t), U_n(t)] = [F_n(t) - \delta_{1-\varepsilon}, F_n(t) + \delta_{1-\varepsilon}],$$

yielding the following theoretical guarantee:

$$\mathbb{P}\left(\forall t \in \mathbb{R} : F(t) \in [L_n(t), U_n(t)]\right) \geq 1 - \varepsilon.$$

This is a foundational result for constructing finite-sample, distribution-free, uniform CBs for the true CDF. These properties make its application to ROC curve analysis, and subsequent refinements, of particular interest.

2.5 Beta Distribution

The Beta distribution plays a central role in our approach to modeling predicted probabilities. It is a flexible continuous probability distribution defined on the interval $p \in [0, 1]$, making it naturally suited for modeling probabilities or proportions. This inherent range aligns perfectly with the nature of predicted probabilities in binary classification, where values represent confidence levels between 0 and 1.

The distribution is parameterized by two concentration parameters, denoted as $\alpha_i > 0$ and $\beta_i > 0$. Here, p represents a general probability parameter, while the subscript in the parameters α_i and β_i allows for parameterization specific to different contexts or observations. The PDF of a Beta distribution is given by:

$$f(p; \alpha_i, \beta_i) = \frac{p^{\alpha_i - 1}(1-p)^{\beta_i - 1}}{B(\alpha_i, \beta_i)},$$

where $B(\alpha_i, \beta_i)$ is the Beta function that serves as a normalizing constant:

$$B(\alpha_i, \beta_i) = \int_0^1 p^{\alpha_i - 1}(1-p)^{\beta_i - 1} dp = \frac{\Gamma(\alpha_i)\Gamma(\beta_i)}{\Gamma(\alpha_i + \beta_i)}$$

with $\Gamma(\cdot)$ representing the gamma function.

The parameters α_i and β_i control the shape of the distribution. When $\alpha_i = \beta_i$, the distribution is symmetric around 0.5, resembling a balanced spread of probability. When $\alpha_i > \beta_i$, the distribution skews to the right, indicating a higher likelihood of larger values near 1. Conversely, when $\alpha_i < \beta_i$, it skews to the left, favoring smaller values near 0. Larger values of both parameters result in a more concentrated distribution, reflecting greater certainty around a specific value of p, much like a peak narrowing around a central point.

Key statistical properties of the Beta distribution provide insight into its central tendencies. The expected value, or mean, is given by:

$$\mathbb{E}[p] = \frac{\alpha_i}{\alpha_i + \beta_i}.$$

The variance of the Beta distribution is given by:

$$\text{Var}(p) = \frac{\alpha_i \beta_i}{(\alpha_i + \beta_i)^2(\alpha_i + \beta_i + 1)}.$$

Additionally, the CDF of the Beta distribution, which computes probabilities over intervals, is:

$$F(t; \alpha_i, \beta_i) = \int_0^t \frac{p^{\alpha_i - 1}(1-p)^{\beta_i - 1}}{B(\alpha_i, \beta_i)} dp.$$

2.5.1 Beta-Binomial Conjugacy

When modeling the uncertainty of aggregate counts rather than individual observations, the Beta distribution serves as a conjugate prior for the Binomial likelihood functionals. If we model our prior belief about a probability parameter p as Beta(α_i, β_i), and we observe k successes in n trials, then the posterior becomes:

$$\begin{aligned}\mathbb{P}(p|\mathcal{D}) &\propto \binom{n}{k} p^k (1-p)^{n-k} \frac{p^{\alpha_i - 1}(1-p)^{\beta_i - 1}}{B(\alpha_i, \beta_i)} \\ &\propto p^{\alpha_i + k - 1}(1-p)^{\beta_i + n - k - 1}\end{aligned}$$

This is proportional to a Beta$(\alpha_i + k, \beta_i + n - k)$ distribution. This conjugacy property enables straightforward sequential Bayesian updating. The parameters α_i and β_i can be interpreted as pseudo-counts or prior observations, with α_i representing the prior number of successes and β_i the prior number of failures.

2.5.2 Asymptotic Normality

For large values of the shape parameters α_i and β_i, the Beta distribution exhibits an important asymptotic behavior: it converges to a normal distribution. This property is particularly relevant for understanding the behavior of Beta-based confidence intervals as sample sizes increase. Specifically, as α_i and β_i become large, the Beta distribution $\text{Beta}(\alpha_i, \beta_i)$ can be approximated by:

$$\text{Beta}(\alpha_i, \beta_i) \approx \mathcal{N}\left(\frac{\alpha_i}{\alpha_i + \beta_i}, \frac{\alpha_i \beta_i}{(\alpha_i + \beta_i)^2(\alpha_i + \beta_i + 1)}\right)$$

In the context of Bayesian statistics, this approximation is formalized by the Bernstein-von Mises theorem, which states that under certain regularity conditions, the posterior distribution converges to a normal distribution centered at the MLE as the sample size increases [Was20].

For a binomial likelihood with parameter p and a Beta prior $\text{Beta}(\alpha_{\text{prior}}, \beta_{\text{prior}})$, after observing k successes in n trials, the posterior distribution is $\text{Beta}(\alpha_{\text{prior}} + k, \beta_{\text{prior}} + n - k)$. As n becomes large, this posterior approaches:

$$\text{Beta}(\alpha_{\text{prior}} + k, \beta_{\text{prior}} + n - k) \approx \mathcal{N}\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

where $\hat{p} = k/n$ is the MLE of p.

This asymptotic normality has important implications for uncertainty quantification in binary classification. When using Beta distributions to model conditional CDFs for positive and negative classes, the resulting confidence intervals at each threshold will, for large sample sizes, behave similarly to those constructed using normal approximations.

2.6 Bootstrap

Bootstrap methods provide powerful, distribution-free approaches for quantifying uncertainty in statistical estimates. In the context of ECDFs and their functionals, bootstrap offers a practical solution for performing uncertainty quantification without requiring parametric assumptions about the underlying data distribution.

Bootstrap works by approximating the empirical distribution of a statistic by resampling with replacement from the original dataset, typically a calibration dataset. This resampling creates multiple bootstrap samples, each containing n_{cal} instances sampled with

replacement from $\mathcal{D}_{\text{cal}}$. For each bootstrap sample $\mathcal{D}_{\text{cal}}^{(b)}$, we compute the corresponding ECDFs.

Let $F_n^{(b)}(t)$ denote the ECDF computed from the b-th bootstrap sample. Once we have generated B bootstrap CDFs, we can construct pointwise confidence intervals for each point in the domain. For a given confidence level $1-\varepsilon$ and point t in the domain, these intervals are defined as:

$$\hat{q}_{\varepsilon/2}(t) = \text{Quantile}_{\varepsilon/2}(\{F_n^{(b)}(t)\}_{b=1}^{B}), \quad \hat{q}_{1-\varepsilon/2}(t) = \text{Quantile}_{1-\varepsilon/2}(\{F_n^{(b)}(t)\}_{b=1}^{B}).$$

These quantiles represent the lower and upper bounds at each point in the domain of the CDF.Beyond the direct quantile-based approach, bootstrap methods also enable variance estimation for complex statistics without making strong distributional assumptions. The bootstrap estimate of variance for a CDF is given by:

$$\text{Var}\left(F_n^{(b)}(t)\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(F_n^{(b)}(t) - \bar{F}_B(t)\right)^2$$

where $\bar{F}_B(t) = \frac{1}{B}\sum_{b=1}^{B} F_n^{(b)}(t)$ is the mean across all bootstrap samples.

Under appropriate regularity conditions and as both the sample size n and number of bootstrap samples B increase, the distribution of the standardized bootstrap estimates $\sqrt{n}(F_n^{(b)}(t) - F_n(t))$ approaches a normal distribution. This asymptotic normality enables an alternative construction of confidence intervals:

$$\left[F_n(t) - z_{1-\varepsilon/2}\cdot\sqrt{\text{Var}\left(F_n^{(b)}(t)\right)}, F_n(t) + z_{1-\varepsilon/2}\cdot\sqrt{\text{Var}\left(F_n^{(b)}(t)\right)}\right]$$

where $z_{1-\varepsilon/2}$ is the $(1-\varepsilon/2)$-quantile of the standard normal distribution. Bootstrap methods for CDFs offer several important theoretical guarantees:

1. **Consistency**: As both the sample size and number of bootstrap samples increase, the bootstrap confidence intervals converge to the true confidence intervals of the population parameters.

2. **Distribution-free**: They make no assumptions about the underlying distribution of the data.

3. **Asymptotic normality**: With a large number of bootstrap samples, the bootstrap distribution can be approximated by a normal distribution, enabling parametric inference in non-parametric settings.

The consistency property of bootstrap is particularly important and can be understood through the lens of the Glivenko-Cantelli theorem. We can break down the convergence into three distinct steps:

Convergence of ECDF to True CDF: The Glivenko-Cantelli theorem states that the empirical CDF (ECDF), $F_n(t)$, converges uniformly to the true CDF, $F(t)$, almost surely as the sample size $n \to \infty$. This is expressed as:

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{a.s.} 0 \quad \text{as} \quad n \to \infty.$$

Convergence of Bootstrap ECDF to ECDF: In the context of bootstrap, by resampling with replacement from the original dataset, the bootstrap samples generate ECDFs, $F_n^{(b)}(t)$, which, by the same Glivenko-Cantelli theorem, converge uniformly to the ECDF $F_n(t)$ as the number of bootstrap samples $B \to \infty$. This can be written as:

$$\sup_{t \in \mathbb{R}} |F_n^{(b)}(t) - F_n(t)| \xrightarrow{a.s.} 0 \quad \text{as} \quad B \to \infty.$$

Convergence of Bootstrap ECDF to True CDF: Since $F_n(t)$ converges to $F(t)$, we can apply the triangle inequality to establish that the bootstrap sample CDF $F_n^{(b)}(t)$ will converge to the true CDF $F(t)$ as both B and n approach infinity. This overall convergence is formally expressed as:

$$\sup_{t \in \mathbb{R}} \left|F_n^{(b)}(t) - F(t)\right| \leq \sup_{t \in \mathbb{R}} \left|F_n^{(b)}(t) - F_n(t)\right| + \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow{a.s.} 0, \quad (1)$$

as $B, n \to \infty$. This step-by-step convergence demonstrates that as our calibration dataset grows larger and we generate more bootstrap samples, our CBs will increasingly reflect the true uncertainty in the estimation.

When applying these bootstrap techniques to ROC curves, the same mathematical principles apply, just to the transformed functions rather than the CDFs directly.

2.7 Conformal Prediction

Conformal prediction is a statistical framework for constructing prediction sets that provide rigorous, finite-sample coverage guarantees, regardless of the underlying data distribution. Unlike traditional methods that yield point estimates, conformal prediction produces a prediction set for a new observation. This property makes conformal prediction a valuable tool for uncertainty quantification in a wide range of applications, including our context of functional objects and ROC curves.

At the core of conformal prediction is the concept of a *non-conformity score*. This is a function

$$s : \mathcal{P} \times \mathcal{Y} \to \mathbb{R},$$

where $\mathcal{P}$ is the set of possible predicted probabilities and $\mathcal{Y}$ is the set of possible true labels. The non-conformity score $s(\hat{p}_i, y_i)$ quantifies how different a data point $(\hat{p}_i, y_i)$ is relative to previously observed data $(\hat{p}_j, y_j)$. Larger values of s indicate greater non-conformity, meaning the pair $(\hat{p}_i, y_i)$ is less typical relative to the reference.

A central assumption of conformal prediction is *exchangeability*, which generalizes the concept of i.i.d. data. A sequence of data points $(\hat{p}_i, y_i)$ is exchangeable if, for any permutation π of the indices $1, \ldots, n$, the joint probability satisfies

$$\mathbb{P}(z_1, \ldots, z_n) = \mathbb{P}(z_{\pi(1)}, \ldots, z_{\pi(n)}),$$

where $z_i = (\hat{p}_i, y_i)$. This invariance to permutations ensures that the order of observations does not affect their joint distribution. Exchangeability is a weaker assumption than i.i.d., making conformal prediction applicable in broader scenarios while still maintaining its statistical guarantees.

The importance of exchangeability lies in its role in enabling the coverage guarantees of conformal prediction. Specifically, it ensures that the non-conformity scores computed from the data,

$$\{s(\hat{p}_i, y_i)\}_{i=1}^{n+1},$$

are themselves exchangeable. This property is crucial for the validity of generalization for a future prediction, as it allows us to make probability statements about a new observation. The following theorem formalizes this guarantee.

Theorem 2.1 (Conformal Calibration Coverage Guarantee [AB22])**.** *Let* $\{(\hat{p}_i, y_i)\}_{i=1}^{n+1}$ *be exchangeable pairs of datapoints. Define* $\hat{q}_{1-\varepsilon}$ *as:*

$$\hat{q}_{1-\varepsilon} = \inf\left(q : \frac{|i : s(\hat{p}_i, y_i) \leq q|}{n} \geq \frac{\lceil (n+1)(1-\varepsilon) \rceil}{n}\right), \tag{2}$$

and let

$$C(\hat{p}_i) = \{y : s(\hat{p}_i, y) \leq \hat{q}_{1-\varepsilon}\} \tag{3}$$

be the prediction set. Then, for any significance level $\varepsilon \in (0, 1)$*,*

$$\mathbb{P}(y_{n+1} \in C(\hat{p}_{n+1})) \geq 1 - \varepsilon. \tag{4}$$

3 Related Works

This section reviews a diverse range of literature that informs our approach to constructing CBs for ROC curves. We begin by examining established statistical techniques, including parametric methods, which are one of the first attempts to quantify uncertainty in ROC curves. Subsequently, we explore various bootstrap approaches for generating CBs. The discussion then moves to the promising yet less developed application of conformal prediction to ROC analysis, highlighting its potential for finite-sample guarantees. Following this, we examine methods specifically aimed at achieving uniform coverage for ROC curves, such as multiple hypothesis testing corrections and fixed-width bands. We then explore work on GoF testing as it relates to CB construction. Finally, we discuss variance-adaptive concentration inequalities, as these mathematical tools for constructing CBs on CDF-like objects provide important theoretical foundations for our approach.

3.1 Parametric Approaches

Early approaches to ROC curve confidence estimation were primarily parametric, assuming specific distributions for the underlying data. The binormal assumption for test scores enables analytical expressions for confidence intervals and bands [Dem12, CM04].

While these parametric approaches offer computational efficiency, they rely on strong modeling assumptions that may not hold for complex high-dimensional data. Our approach aims to overcome this limitation by developing methods that maintain validity regardless of the underlying data distribution and model of choice.

3.2 Bootstrap Methods

Bootstrap methods, whose general principles and mathematical foundations are detailed in Section 2.6 of the Background, have been widely used for uncertainty quantification in ROC curve analysis. Here, we focus on how these methods have been specifically applied to constructing CBs for ROC curves and CDFs, and discuss their limitations in this context.

The idea of using bootstrap to construct uniform CBs for CDFs themselves, rather than just specific statistics derived from them, has a strong precedent. Notably, Bickel and Krieger [BK89] investigated the construction of CBs for a CDF using bootstrap.

They proposed this as an alternative to the standard Kolmogorov-Smirnov CB, which is known to be conservative, especially when the underlying distribution F is discrete. Their method involves computing the ECDF F_n from the original sample and then, for each bootstrap sample, computing its ECDF $F_n^{(b)}$. These samples are then used in conjuction with F_n to estimate the Kolmogorov statistic:

$$\hat{s}_n^{(b)} = \sup_{t \in \mathbb{R}} \sqrt{n} |F_n^{(b)}(t) - F_n(t)|. \tag{5}$$

The bootstrap CB is then given by $F_n(t) \pm \hat{q}_{1-\varepsilon}$, where $\hat{q}_{1-\varepsilon}$ is the appropriate quantile of $\hat{s}_n^{(b)}$. Bickel and Krieger demonstrated that their bootstrap CB has the correct coverage probability asymptotically and, through simulations, showed that it works well for small samples and outperforms the conservative approach, particularly for distributions that have small carriers. Given that the TPR and FPR components of an ROC curve are essentially conditional CDFs, the work of Bickel and Krieger provides a way to construct bootstrap CBs for ROC curves. In Section 5, we reproduce their results and show that the bootstrap CB for ROC curves is consistent, serving as a reasonable baseline for our main method, see Section 4.3.3.

However, the performance of bootstrap in general is bottlenecked by the number of calibration samples n_{cal}. In finite sample settings, there will always be an unquantifiable uncertainty in terms of the bias between the empirical and the true distribution. This makes the bootstrap method unreliable in practice where the calibration set is non representative relative to the test set.

3.3 Conformal Prediction

Despite the theoretical advantages of conformal prediction for uncertainty quantification, particularly its finite-sample guarantees and distribution-free nature, its application to ROC curves remains relatively limited. Zheng et al. [ZYS24] represent one notable attempt, but a key step in their method relies on kernel density estimation for constructing prediction sets, making it dependent on bandwidth selection, which introduces a dependency on the characteristics of the dataset and classifier.

Another attempt to apply conformal prediction for functionals is given by Diquigiovanni et al. [DFV21]. They demonstrate that given an estimator that finds an approximation of a set of functionals $\{f_1(t), ..., f_n(t)\}$, it is possible to construct a CB that bounds at least $\lfloor n(1-\varepsilon) \rfloor$ of the functionals with probability $1-\varepsilon$. They propose non-conformity scores based on the Kolmogorov-Smirnov statistic that measures the deviation between each of the functionals, normalized by two different quantities:

- Standard deviation of the set of functionals.

- Maximum deviation of the set of functionals.

They demonstrate the validity of the CBs if multiple functionals are available during calibration. However, their method is not directly applicable in our case as we only have access to n_{cal} samples consisting of individual points of predicted probabilities and corresponding labels. This will thus only yield a single functional, thus a single non-conformity score, making the quantile computation non-feasable. However, we did find their standard deviation based non-conformity score to be a good choice for our method, see Section 4.3.3.

3.4 Methods for Uniform Coverage

Ensuring CBs provide simultaneous coverage across all thresholds of the ROC curve is crucial. Several methods address this challenge, often applied in conjunction with techniques like bootstrap.

3.4.1 Multiple Hypothesis Testing Correction

A common strategy is to adjust pointwise significance levels using multiple testing corrections. The simplest approach, Bonferroni correction, adjusts the significance level ε to ε/T for T thresholds, but this is often overly conservative, leading to excessively wide bands. Less conservative methods like the Holm-Bonferroni [Hol79] or the Benjamini-Hochberg procedure [BH95] offer improvements but still tend towards conservatism and control different error rates rather than directly guaranteeing uniform coverage probability for the band itself.

3.4.2 Fixed-Width Bands

Another approach aiming for simultaneous coverage is the fixed-width bands method [MPR05]. This non-parametric technique constructs the confidence band by displacing the entire empirical ROC curve both "northwest" and "southeast" along a specific, predefined slope (often related to the ratio of positive to negative examples). The resulting band has a constant width when measured along this chosen slope. The required width is determined empirically using bootstrap resampling: numerous ROC curves are generated from bootstrap samples, and the width is chosen such that a desired proportion (e.g., $1 - \varepsilon$) of these bootstrap curves fall entirely within the band constructed around the original empirical curve.

While methods like multiple testing corrections and fixed-width bands aim for uniform coverage, they often rely on adjustments or empirical width calculations based on initial estimations. This contrasts with methods like DKW inequality which aim to directly construct bands with a guaranteed coverage probability $1 - \varepsilon$ by considering the entire function or using concentration inequalities, potentially yielding bounds that better reflect the true uncertainty structure.

3.5 Goodness-of-Fit Testing

Building upon the principle of adapting to local variance, several GoF tests have been developed [DW22, SP16, CB12, BM23]. Variance-adaptive methods aim to create tighter and more informative CBs by incorporating this local variability, making GoF tests more powerful, especially in the tails.

Early work by Chicheportiche and Bouchaud [CB12] specifically targeted the tail behavior of distributions by developing a weighted Kolmogorov-Smirnov test. While primarily providing an *asymptotic perspective* for GoF, their approach highlighted the necessity of "accounting for the tails," a crucial aspect of variance adaptivity.

Stepanova and Pavlenko [SP16] proposed a broad class of GoF test statistics based on sup-functionals of weighted empirical processes, employing Erdős-Feller-Kolmogorov-Petrovski upper-class functions as weights. Their work includes the construction of CBs derived from these tests, which are inherently variance-adaptive and were found to perform well. Dumbgen and Wellner [DW22] introduced an improved approach that refines existing methods by utilizing multi-scale testing techniques and refined laws of the iterated logarithm to construct both GoF tests and corresponding CBs.

While these methods are powerful and result in bands that adapt to local variance, their tightest theoretical guarantees and the practical determination of critical values often rely on *asymptotic theory* while the ability to generalize to a future prediction is not guaranteed.

The key distinctions from the above described methods to the goal of this thesis are:

- **Purpose of the Guarantee**: The GoF paradigm typically aims to test a hypothesis about the entire CDF (is F_n a good estimate of F?) or provide CBs that contain the true CDF F with high probability. In contrast, this thesis seeks to construct bands for an ROC curve such that a *test ECDF* will fall within these bands with a prespecified probability.
- **Nature of Theoretical Guarantees**: Many of the variance-adaptive GoF meth-

ods [DW22, SP16, CB12], while conceptually powerful for assessing fit, establish their most precise results and practical utility through *asymptotic theory*. This thesis, however, prioritizes non-asymptotic, finite-sample guarantees. Such guarantees ensures that the stated confidence level holds for the given dataset size without relying on large-sample approximations.

3.6 Variance-Adaptive Concentration Inequalities

An advancement in variance-adaptive concentration inequalities is the variance-dependent DKW inequality proposed by Bartl and Mendelson [BM23]. For F and F_n, they proved that there exists absolute constants c_0 and c_1 such that:

$$\delta^2 \geq c_0 \frac{\log \log n}{n} \tag{6}$$

and

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} \frac{|F_n(t) - F(t)|}{\sigma(t)} \leq \delta\right) \geq 1 - \varepsilon, \tag{7}$$

where $\delta = \sqrt{\frac{\ln(2/\varepsilon)}{c_1 n}}$. This bound adapts to the local variance $\sigma^2(t) = F(t)(1 - F(t))/n$ at each point t, providing significantly tighter CBs than the standard DKW bound in regions where the variance is small. This is particularly relevant for ROC curves, which often have regions of both high and low variance near the tails of the CDF. We draw inspiration from this work when constructing the non-conformity score in Section 4.3.3. While Bartl and Mendelson establish the theoretical form of this variance-adaptive bound and prove its optimality up to multiplicative constants, they do not determine the exact values of c_0 and c_1. This limitation makes direct application challenging in practice.

4 Uncertainty Quantification for CDFs

As outlined in the introduction, our goal is to develop a method that satisfies the following properties:

1. Conditional coverage.
2. Finite sample guarantee.
3. Distribution-free.

4. Variance-adaptive.

We present several methodologies for constructing CBs for CDFs, building up to the main method proposed in this thesis, see Section 4.3.3.

The construction of ECDFs from finite datasets provides an approximation of the true CDF, which is a theoretical construct that cannot be directly observed. The true CDF represents the underlying distribution of the data, and its inclusion in our analysis is crucial for developing robust statistical methods. To address this, we consider a function space of CDFs, denoted as $\mathcal{F}$, where both empirical and theoretical CDFs are included. This space allows us to analyze the behavior of CDFs comprehensively, providing a framework for uncertainty quantification that includes the true CDF. In the following section, we extend the conformal prediction framework to functional objects, providing a theoretical foundation for constructing CBs for CDFs and ROC curves.

4.1 Functional Conformal Prediction

Conformal prediction, originally developed for constructing prediction sets for individual data points with guaranteed coverage probability, can be extended to handle functionals evaluated at specific thresholds through a framework known as FCP [DFV21]. This extension is particularly valuable for functionals derived from CDFs, such as ROC curves, where uncertainty quantification across the entire function is desired.

For CDFs and their functionals, we consider a domain of thresholds $t \in \mathcal{T}$ with finite number of elements. The goal of FCP is to construct a functional prediction set or CBs $C(F_n(t))$. This set should contain a future function across all thresholds simultaneously with a specified probability $1 - \varepsilon$, that is:

$$\mathbb{P}\left(\forall t \in \mathcal{T} : F_n^{(B+1)}(t) \in C(F_n(t))\right) \geq 1 - \varepsilon.$$

Unlike the point-wise prediction set in standard conformal prediction, which targets a single label y_{n+1} for a future datapoint, the CB in FCP encapsulates the entire functional behavior, providing a region where the true function is expected to lie. We denote the CB as $C(F_n(t))$ and it can be represented as $[L_n(t), U_n(t)]$.

Within this function space, we define the non-conformity measure for CDFs as follows:

$$s : \mathcal{F} \times \mathcal{F} \to \mathbb{R}.$$

This measure quantifies the difference between two functions, providing a basis for constructing CBs that contain the true CDF with a specified probability.

Several non-conformity scores can be used within this framework, each offering different theoretical guarantees and practical implications:

- **Threshold-Wise Quantile**: Defines a CB that bounds a specified percentage of points at each threshold:

$$s(F_n, F) = \text{Quantile}_{1-\varepsilon}\left(|F_n(t) - F(t)|\right). \tag{8}$$

- **Supremum Norm**: Captures the maximum deviation across thresholds:

$$s(F_n, F) = \sup_{t \in \mathcal{T}} |F_n(t) - F(t)|. \tag{9}$$

  This corresponds to a 1.0 quantile coverage.

- **Standardized Supremum Norm** [DFV21]: The supremum norm divided by the standard deviation of the ground truth function:

$$s(F_n, F) = \sup_{t \in \mathcal{T}} \frac{|F_n(t) - F(t)|}{\sigma(t)}. \tag{10}$$

- **Standardized L2 Norm**: Measures the average deviation over the domain. This is a more representative measure of how the functionals are distributed as it is closely related to standard deviation:

$$s(F_n, F) = \sqrt{\sum_{t \in \mathcal{T}} \left(\frac{F_n(t) - F(t)}{\sigma(t)}\right)^2}. \tag{11}$$

Other functional norms or distance metrics can also be considered depending on the specific kind of theoretical guarantees desired. We will focus on the normalized supremum norm as it provides CBs that are simultaneously valid across all thresholds.

4.1.1 Dataset-Level Exchangeability

When applying conformal prediction to functionals derived from datasets, we encounter a fundamental challenge: the standard notion of exchangeability, which traditionally applies to individual data points, must be extended to entire datasets and the functional objects derived from them. This extension is necessary because CDFs, ROC curves, and other statistical functionals are computed from complete datasets rather than individual observations.

Consider a dataset $\mathcal{D} = \{(\hat{p}_i, y_i)\}_{i=1}^n$ where each data point $(\hat{p}_i, y_i)$ is drawn independently and identically from a distribution P. For FCP, we typically split $\mathcal{D}$ into a training set $\mathcal{D}_{\text{train}}$, a calibration set $\mathcal{D}_{\text{cal}}$, and a test set $\mathcal{D}_{\text{test}}$. The concept of dataset-level exchangeability becomes particularly relevant when we consider multiple datasets or samples used to compute functional objects.

To formalize this extension, we first observe that we are working with functionals evaluated at a finite set of thresholds $\mathcal{T}$. This discretization allows us to view each functional object $F_n(t)$ as a vector in $\mathbb{R}^{|\mathcal{T}|}$, which simplifies the theoretical treatment while retaining the practical advantages of the functional perspective.

Two datasets $\mathcal{D}^{(a)}$ and $\mathcal{D}^{(b)}$ are exchangeable if their joint distribution is invariant to permutation. When these datasets are used to compute functional objects such as ECDFs, $F_n^{(a)}(t)$ and $F_n^{(b)}(t)$, the resulting functional objects inherit this exchangeability property. This inheritance is crucial because it allows us to apply the theoretical guarantees of conformal prediction to functional objects.

The mathematical foundation for this inheritance lies in the properties of ECDFs. The expectation of any ECDF is precisely the true underlying CDF:

$$\mathbb{E}[F_n(t)] = F(t).$$

This means that any ECDFs generated from the same distribution are i.i.d. realizations centered around the true CDF, and are therefore exchangeable. This property extends directly to the non-conformity scores derived from these ECDFs, which is essential for establishing the validity of our CBs.

Specifically, if $F_n^{(1)}(t), F_n^{(2)}(t), \ldots, F_n^{(B)}(t), F_n^{(B+1)}(t)$ are ECDFs derived from exchangeable datasets drawn from the same distribution, then these functional objects themselves form an exchangeable sequence. This property is the cornerstone that allows us to apply conformal prediction theory to functional spaces.

In practice, various sampling techniques can be used to generate exchangeable functional objects while maintaining this theoretical framework. Methods such as **bootstrap sampling**, **Bayesian methods**, or **Monte Carlo simulation** can all produce exchangeable functional objects as long as the sampling mechanism preserves the exchangeability property. The specific choice of method affects how uncertainty is quantified, but the fundamental guarantee of exchangeability remains intact across these approaches.

Given a sequence of exchangeable functional objects, we can compute corresponding exchangeable non-conformity scores between each functional object and a reference (such as the ground truth or an approximation thereof) using an appropriate functional

distance metric:

$$s\left(F_n^{(b)}, F\right) = \sup_{t \in \mathcal{T}} \frac{|F_n^{(b)}(t) - F(t)|}{\sigma(t)}.$$

The $(1-\varepsilon)$-quantile of these scores determines the threshold for constructing CBs:

$$C\left(F\right) = \{F \in \mathcal{F} : s\left(F_n^{(b)}, F\right) \leq \hat{q}_{1-\varepsilon}\}.$$

This prediction set defines the CB by including all functions whose non-conformity score is below the threshold $\hat{q}_{1-\varepsilon}$, ensuring that the ground truth lies within this band with probability at least $1-\varepsilon$. Under the assumption of functional exchangeability, we can extend the rigorous marginal coverage guarantees of conformal prediction to functionals evaluated at specific points, as formalized in the following section.

4.1.2 Coverage Guarantee for Functional Conformal Prediction

Building on the exchangeability assumptions for functional objects and datasets, we can adapt the coverage guarantee from Theorem 2.1 to the context of FCP. The following theorem formalizes this extension, ensuring that CBs for functional data contain the true function with the desired probability.

Theorem 4.1 (Functional Conformal Coverage Guarantee [DFV21])**.** *Let* $\{(F_n^{(b)}, F)\}_{b=1}^{B+1}$ *be exchangeable pairs of functionals where* $F_n^{(b)}$ *are CDFs derived from the calibration dataset* $\mathcal{D}_{cal}$*. Define* $\hat{q}_{1-\varepsilon}$ *as:*

$$\hat{q}_{1-\varepsilon} = \inf\left(q : \frac{|b : s(F_n^{(b)}, F) \leq q|}{B} \geq \frac{\lceil (B+1)(1-\varepsilon) \rceil}{B}\right), \tag{12}$$

and let

$$C(F_n) = [F_n(t) - \hat{q}_{1-\varepsilon}, F_n(t) + \hat{q}_{1-\varepsilon}], \tag{13}$$

be the functional prediction set. Then, for any significance level $\varepsilon \in (0,1)$*,*

$$\mathbb{P}\left(F_n^{(B+1)} \in C(F_n)\right) \geq 1-\varepsilon, \tag{14}$$

or equivalently:

$$\mathbb{P}\left(\forall t \in \mathcal{T} : |F_n^{(B+1)}(t) - F(t)| \leq \hat{q}_{1-\varepsilon}\right) \geq 1-\varepsilon. \tag{15}$$

The proof is identical as in Theorem 2.1, see [AB22].

This theorem extends the marginal coverage guarantee of Theorem 2.1 to functional objects, providing a theoretically sound basis for constructing CBs for CDFs and ROC curves under the assumption of exchangeable functional objects and datasets.

Corollary 4.2 (Functional Variance-Adaptive Coverage Guarantee)**.** *Let*

$$\sigma(t) = \sqrt{Var(F(t))} > 0 \tag{16}$$

be the standard deviation of the true CDF at threshold $t \in \mathcal{T}$. Define the standardized non-conformity score as:

$$s(F_n^{(b)}, F) = \sup_{t \in \mathcal{T}} \frac{|F_n^{(b)}(t) - F(t)|}{\sigma(t)}. \tag{17}$$

Let $\hat{q}_{1-\varepsilon}$ be the $(1-\varepsilon)$-quantile be defined in the same way as in Theorem 4.1. Then for the prediciton set:

$$C(F_n(t)) = [F_n(t) - \hat{q}_{1-\varepsilon} \cdot \sigma(t), F_n(t) + \hat{q}_{1-\varepsilon} \cdot \sigma(t)], \tag{18}$$

we have that:

$$\mathbb{P}\left(\forall t \in \mathcal{T} : F(t) \in C(F_n(t))\right) \geq 1 - \varepsilon, \tag{19}$$

or equivalently:

$$\mathbb{P}\left(\forall t \in \mathcal{T} : |F_n^{(B+1)}(t) - F(t)| \leq \hat{q}_{1-\varepsilon} \cdot \sigma(t)\right) \geq 1 - \varepsilon. \tag{20}$$

Proof. The proof follows from Theorem 4.1, with a modified non-conformity score that accounts for variance at different thresholds.

Given the standardized non-conformity score defined as:

$$s(F_n^{(b)}, F) = \sup_{t \in \mathcal{T}} \frac{|F_n^{(b)}(t) - F(t)|}{\sigma(t)},$$

we compute the empirical $(1-\varepsilon)$-quantile $\hat{q}_{1-\varepsilon}$ of these scores:

$$\hat{q}_{1-\varepsilon} = \inf\left(q : \frac{|b : s(F_n^{(b)}, F) \leq q|}{B} \geq \frac{\lceil (B+1)(1-\varepsilon) \rceil}{B}\right).$$

By the exchangeability of the ECDF samples and applying Theorem 4.1, we have:

$$\mathbb{P}(s(F_n^{(B+1)}, F) \leq \hat{q}_{1-\varepsilon}) \geq 1 - \varepsilon$$

Expanding the non-conformity score definition, the event $s(F_n^{(B+1)}, F) \leq \hat{q}_{1-\varepsilon}$ means:

$$\sup_{t \in \mathcal{T}} \frac{|F_n^{(B+1)}(t) - F(t)|}{\sigma(t)} \leq \hat{q}_{1-\varepsilon}$$

This is equivalent to:

$$\forall t \in \mathcal{T} : \frac{|F_n^{(B+1)}(t) - F(t)|}{\sigma(t)} \leq \hat{q}_{1-\varepsilon}$$

Multiplying both sides by $\sigma(t)$, which is positive by definition:

$$\forall t \in \mathcal{T} : |F_n^{(B+1)}(t) - F(t)| \leq \hat{q}_{1-\varepsilon} \cdot \sigma(t)$$

This inequality is equivalent to stating that for all $t \in \mathcal{T}$:

$$F(t) \in [F_n(t) - \hat{q}_{1-\varepsilon} \cdot \sigma(t), F_n(t) + \hat{q}_{1-\varepsilon} \cdot \sigma(t)] = C(F_n(t))$$

Therefore:

$$\mathbb{P}\left(\forall t \in \mathcal{T} : F(t) \in C(F_n(t))\right) \geq 1 - \varepsilon$$

Which completes the proof of the corollary. □

4.2 General Framework for CBs Construction

Building on the concepts of FCP, we now establish a general framework for constructing CBs for CDFs. This framework provides a unified methodology that underlies various sampling approaches for uncertainty quantification of CDFs/ECDFs evaluated at specific thresholds, minimizing repetition while highlighting the common elements across methods.

In this framework, we consider the general problem of bounding the difference between two functional forms: a reference CDF and an observed or sampled CDF. It is important to note that the precise interpretation of these functions varies across methods. The reference CDF, denoted as $F(t)$, represents the true continuous CDF, which serves as the ground truth. In practice, it may be approximated by the ECDF computed from a calibration set (as in bootstrap methods) or other forms depending on the context. Similarly, $F_n^{(b)}(t)$ represents the b-th ECDF sample derived from sampling from the true CDF F, essentially resampling an approximation like $F_n(t)$ b times. For each $b = 1, \ldots, B$, we generate a sample $\mathcal{D}^{(b)} \sim \mathcal{D}$ by drawing pairs $(\hat{p}_i^{(b)}, y_i^{(b)})$ from the underlying dataset $\mathcal{D}$, which contains predicted probabilities from a classifier, from which the ECDF $F_n^{(b)}(t)$ is computed. The specific interpretations and sampling mechanisms will be clarified in the respective method subsections.

Non-Conformity Score: The foundation of our framework is the non-conformity score, which quantifies the discrepancy between two functional forms. For CDFs, we define the general form of the score as:

$$s(F_n^{(b)}, F) = \sup_{t \in \mathcal{T}} \frac{|F_n^{(b)}(t) - F(t)|}{\sigma(t)}, \tag{21}$$

where $\sigma(t)$ is a scaling factor that can be adjusted to create different types of CBs. Setting $\sigma(t) = 1$ yields uniform CBs, with equal width across all thresholds, while setting $\sigma(t)$ based on local variability (e.g., estimated standard deviation of F at t) yields variance-adaptive CBs, which are narrower in regions of low variability and wider in regions of high variability. Different methods may estimate or approximate $F(t)$ and $\sigma(t)$ differently, as detailed in their respective subsections.

Empirical Quantile Estimation: To construct CBs with a desired coverage probability $1 - \varepsilon$, we first compute a set of non-conformity scores $\{s_n^{(b)}\}_{b=1}^B$ from B sampled functional forms. The empirical quantile is then estimated as:

$$\hat{q}_{1-\varepsilon} = \text{Quantile}_{1-\varepsilon}\left(\{s_n^{(b)}\}_{b=1}^B\right), \tag{22}$$

which represents the threshold for the non-conformity score at the $(1 - \varepsilon)$ level. This quantile defines the width of the CB, with specific methods varying in how they generate the set of scores.

Prediction Set Construction: Using the estimated quantile, we construct the general form of the CB or prediction set as:

$$[L_n(t), U_n(t)] = [F_n(t) - \hat{\delta}_{1-\varepsilon}(t), F_n(t) + \hat{\delta}_{1-\varepsilon}(t)], \tag{23}$$

where $\hat{\delta}_{1-\varepsilon}(t) = \hat{q}_{1-\varepsilon} \cdot \sigma(t)$ adjusts the band width based on the quantile and scaling factor. Here, $F_n(t)$ typically represents an ECDF as an approximation of the true CDF $F(t)$, depending on the specific method.

Theoretical Coverage Guarantee: The CBs constructed using this framework provide a theoretical coverage guarantee:

$$\mathbb{P}\left(\forall t \in \mathcal{T}, F(t) \in [L_n(t), U_n(t)]\right) \geq 1 - \varepsilon, \tag{24}$$

which ensures that the sampled functional form lies within the band with high probability across all thresholds $t \in \mathcal{T}$. We state the coverage probability as $1 - \varepsilon$ to account for the conservative quantile estimation in finite samples and the two-sided construction of the CB, ensuring a robust guarantee as per Theorem 4.1. The exact interpretation of what $F_n^{(b)}(t)$ represents in terms of uncertainty (e.g., aleatory versus epistemic uncertainty) depends on the specific sampling method and will be discussed in the respective subsections.

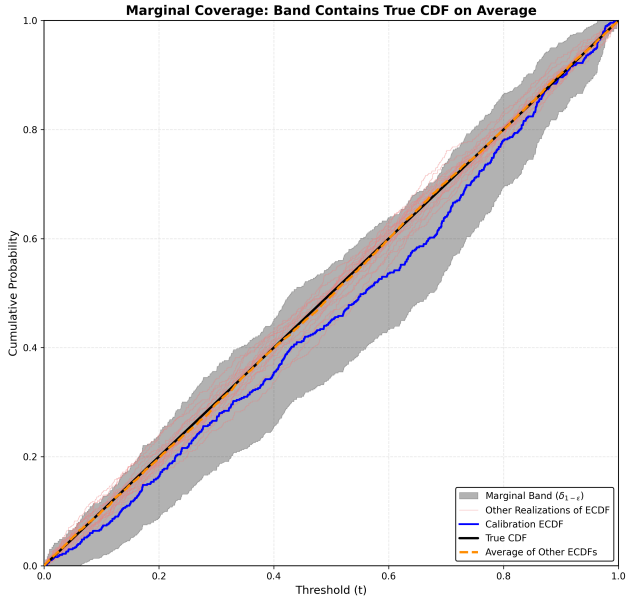

Figure 1 Illustration of the marginal coverage property of conformal prediction bands for ECDFs. The blue line represents the ECDF from a single calibration dataset. The grey shaded area is the 95% CB constructed around this calibration ECDF. While individual new ECDFs from the same data-generating process (shown in light red) are not guaranteed to lie entirely within the band, their average (dashed orange line) is well-contained. This demonstrates that the CB provides coverage on average, or "marginally", over repeated experiments.

In the following subsections, we will explore specific sampling approaches for generating multiple functional forms, such as Bootstrap, Bayesian, and Monte Carlo methods. Each method adapts this general framework to their unique characteristics while maintaining the core methodology for constructing CBs.

4.2.1 Extension to Future Test Set Coverage

In practical applications, we are often more interested in bounding the ECDF derived from a future dataset rather than the true underlying CDF itself. This motivates the extension of our CB framework to provide coverage guarantees for a test set ECDF, $F_{n_{\text{test}}}(t)$, using bounds derived from a calibration set.

While this extension may appear to introduce a more conservative bound, it appropriately captures the combined uncertainty between calibration and test ECDFs relative to

the true CDF, aligning with the practical objective of many statistical applications.

Non-Conformity Score for Test Set Coverage: We define the target non-conformity score for test set coverage as the maximum deviation between the calibration set ECDF and the test set ECDF, scaled by the local variability:

$$s_{\text{test}} = \sup_{t \in \mathcal{T}} \frac{|F_{n_{\text{cal}}}(t) - F_{n_{\text{test}}}(t)|}{\sigma(t)}. \tag{25}$$

Using the triangle inequality, we can bound this score by decomposing it into deviations from the true underlying CDF:

$$\begin{aligned} s_{\text{test}} &= \sup_{t \in \mathcal{T}} \frac{|F_{n_{\text{cal}}}(t) - F_{n_{\text{test}}}(t)|}{\sigma(t)} \\ &= \sup_{t \in \mathcal{T}} \frac{|(F_{n_{\text{cal}}}(t) - F(t)) + (F(t) - F_{n_{\text{test}}}(t))|}{\sigma(t)} \\ &\leq \sup_{t \in \mathcal{T}} \frac{|F_{n_{\text{cal}}}(t) - F(t)|}{\sigma(t)} + \sup_{t \in \mathcal{T}} \frac{|F(t) - F_{n_{\text{test}}}(t)|}{\sigma(t)}. \end{aligned}$$

Bounding via Exchangeability: Let us develop a rigorous approach to bound the maximum deviation between calibration and test ECDFs. Under the assumption of dataset-level exchangeability and equal sample sizes $n_{\text{cal}} = n_{\text{test}}$, we know that the deviations of the calibration ECDF from the true CDF and the test ECDF from the true CDF follow identical distributions.

Let us define:

$$\begin{aligned} s_{\text{cal}} &= \sup_{t \in \mathcal{T}} \frac{|F_{n_{\text{cal}}}(t) - F(t)|}{\sigma(t)}, \\ s_{\text{test}} &= \sup_{t \in \mathcal{T}} \frac{|F_{n_{\text{test}}}(t) - F(t)|}{\sigma(t)}. \end{aligned}$$

From Theorem 4.1, for any significance level $\varepsilon/2 \in (0, 1)$, we have:

$$\begin{aligned} \mathbb{P}(s_{\text{cal}} \leq \hat{q}_{1-\varepsilon/2}) &\geq 1 - \varepsilon/2, \\ \mathbb{P}(s_{\text{test}} \leq \hat{q}_{1-\varepsilon/2}) &\geq 1 - \varepsilon/2. \end{aligned}$$

By the union bound, the probability that both deviations are simultaneously bounded is:

$$\mathbb{P}(s_{\text{cal}} \leq \hat{q}_{1-\varepsilon/2} \text{ and } s_{\text{test}} \leq \hat{q}_{1-\varepsilon/2}) \geq 1 - \varepsilon. \tag{26}$$

Now, let us define our target score as the maximum deviation between the calibration and test ECDFs:

$$s_{\text{diff}} = \sup_{t \in \mathcal{T}} \frac{|F_{n_{\text{cal}}}(t) - F_{n_{\text{test}}}(t)|}{\sigma(t)} = \sup_{t \in \mathcal{T}} \frac{|(F_{n_{\text{cal}}}(t) - F(t)) - (F_{n_{\text{test}}}(t) - F(t))|}{\sigma(t)}.$$

Under the event that $s_{\text{cal}} \leq \hat{q}_{1-\varepsilon/2}$ and $s_{\text{test}} \leq \hat{q}_{1-\varepsilon/2}$, we can bound s_{diff} using the triangle inequality:

$$\begin{aligned}
s_{\text{diff}} &= \sup_{t \in \mathcal{T}} \frac{|(F_{n_{\text{cal}}}(t) - F(t)) - (F_{n_{\text{test}}}(t) - F(t))|}{\sigma(t)} \\
&\leq \sup_{t \in \mathcal{T}} \frac{|F_{n_{\text{cal}}}(t) - F(t)|}{\sigma(t)} + \sup_{t \in \mathcal{T}} \frac{|F_{n_{\text{test}}}(t) - F(t)|}{\sigma(t)} \\
&= s_{\text{cal}} + s_{\text{test}} \\
&\leq \hat{q}_{1-\varepsilon/2} + \hat{q}_{1-\varepsilon/2} \\
&= 2 \cdot \hat{q}_{1-\varepsilon/2}.
\end{aligned}$$

Therefore:

$$\mathbb{P}(s_{\text{diff}} \leq 2 \cdot \hat{q}_{1-\varepsilon/2}) \geq 1 - \varepsilon. \tag{27}$$

This bound provides a rigorous coverage guarantee while accounting for the combined uncertainty from both the calibration and test sets.

Adjusted Prediction Set: Based on this bound, we construct an adjusted CB for test set coverage:

$$[L_{n_{\text{cal}}}(t), U_{n_{\text{cal}}}(t)] = [F_{n_{\text{cal}}}(t) - 2 \cdot \hat{\delta}_{1-\varepsilon/2}(t), F_{n_{\text{cal}}}(t) + 2 \cdot \hat{\delta}_{1-\varepsilon/2}(t)], \tag{28}$$

where $\hat{\delta}_{1-\varepsilon/2}(t) = \hat{q}_{1-\varepsilon/2} \cdot \sigma(t)$ corresponds to the adjusted significance level $\varepsilon/2$ as derived in our rigorous approach. This adjusted CB is designed to bound a future test ECDF, $F_{n_{\text{test}}}(t)$, ensuring that it lies within the specified bounds with high probability. The adjusted band provides the following coverage guarantee:

$$\mathbb{P}\left(\forall t \in \mathcal{T}, F_{n_{\text{test}}}(t) \in [L_{n_{\text{cal}}}(t), U_{n_{\text{cal}}}(t)]\right) \geq 1 - \varepsilon. \tag{29}$$

The coverage probability remains $1 - \varepsilon$, reflecting the conservative nature of our bound due to the triangle inequality and the finite sample quantile estimation.

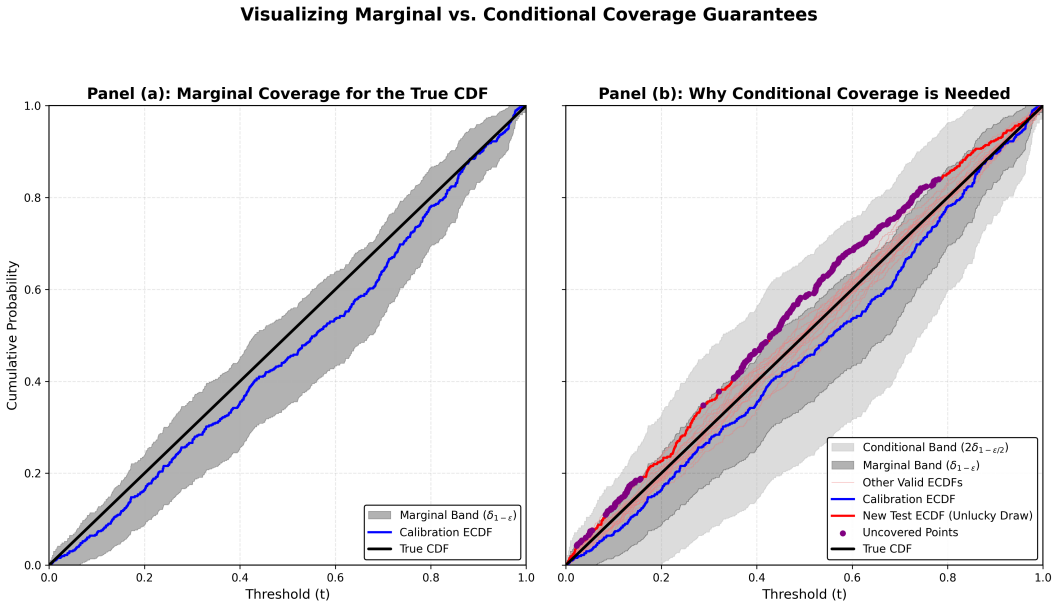

Figure 2 The CB in panel (a) is constructed based on the ECDF from a specific realization of the true CDF. It guarantees marginal coverage with band-width $\delta_{1-\varepsilon}$ which means that the true CDF is covered with probability at least $1 - \varepsilon$. In panel (b), several curves are being sampled from the same distribution. However, we might get an unlucky draw that is far away from the calibration ECDF the CBs was constructed on. The purple points highlights the points that are outside the marginal CBs, which means that the CB is not sufficiently wide to contain this unlucky curve. Therefore, we need to construct a conditional CB that is wider than the marginal CB, with band-wdith $2\delta_{1-\varepsilon/2}$ to ensure that the test set ECDF is covered with probability at least $1 - \varepsilon$.

Note that this extension assumes that $n_{\text{cal}} \leq n_{\text{test}}$ for the CBs to generalize to the test set. The intuition is that the band-width decreases as the sample size increases. If for instance, the CBs are constructed on a calibration set with $n_{\text{cal}} = 1000$, the CBs will be too narrow to contain the test curve with $n_{\text{test}} = 100$ which would have a larger variance due to a smaller sample size.

4.3 Sampling Approaches for Generating Multiple CDFs

Having established the general framework for constructing CBs, we now explore specific sampling methods to generate multiple functional forms from a calibration set. Each method adapts the general framework to its unique characteristics while maintaining the core methodology for non-conformity scores, quantile estimation, and CB

construction. We consider three approaches: bootstrap sampling, Bayesian posterior sampling, and Monte Carlo sampling.

4.3.1 Bootstrap CBs

In the bootstrap framework, the reference groud truth CDF is the ECDF $F_n(t)$, computed from the calibration set. The corresponding samples $F_n^{(b)}(t)$ are generated based on the ECDF. Using bootstrap sampling, we generate multiple instances of the ECDFs from one calibration set as follows:

1. Let $F_n(t)$ be the ECDF of the calibration set:

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\hat{p}_i > t], \tag{30}$$

   with $\mathcal{T} = \{t_1, t_2, \ldots, t_T\}$ being a fixed grid of T thresholds.

2. For $b = 1, \ldots, B$, draw a bootstrap sample $(\hat{p}_i^{(b)}, y_i^{(b)})$ with replacement from the calibration set:

$$(\hat{p}_i^{(b)}, y_i^{(b)}) \sim \mathcal{D}_{\text{cal}}. \tag{31}$$

3. Compute the resampled conditional ECDF for each bootstrap sample:

$$F_n^{(b)}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\hat{p}_i^{(b)} > t]. \tag{32}$$

   If we want to get the CBs for ROC curve, we condition the ECDF on the resampled true labels $y_i^{(b)}$.

4. Estimate the standard deviation $\sigma(t)$ using the sample standard deviation across bootstrap samples:

$$\sigma(t) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(F_n^{(b)}(t) - F_n(t) \right)^2}. \tag{33}$$

Following the general framework established earlier, we compute the non-conformity scores using the supremum distance scaled by the estimated standard deviation, determine the empirical quantile, and construct the CBs with the appropriate width adjustment. The resulting prediction set provides a coverage guarantee of at least $1 - \varepsilon$.

For test set coverage extension, the CBs can be doubled in width as described in the previous subsection, yielding a conservative bound that accounts for both calibration and test set uncertainty.

> The quantity $F_n^{(b)}(t)$ that the CBs $[L_n(t), U_n(t)]$ are capturing with probability at least $1 - \varepsilon$ is the **aleatory uncertainty** of the ECDF. However, the bootstrap approach cannot capture the **epistemic uncertainty** between the ECDF and true CDF due to its inherent limitation of resampling from a finite calibration set.

4.3.2 Beta-Binomial CBs

In the Bayesian framework, the reference CDF $F(t)$ is specified by the choice of priors and corresponding posterior distribution. This is typically approximated by the ECDF from the calibration set. The sampled functional forms $F_n^{(b)}(t)$ represent CDFs sampled from posterior distributions, derived as the b-th ECDF sample from the true CDF F. Bayesian methods offer an alternative framework for generating multiple CDFs from a single calibration set by sampling from a constructed posterior distribution. For each $b = 1, \ldots, B$, we generate a sample $\mathcal{D}^{(b)} \sim \mathcal{D}$ by drawing pairs $(\hat{p}_i^{(b)}, y_i^{(b)})$ from the underlying dataset $\mathcal{D}$, which contains predicted probabilities from a classifier, adapted to the posterior distribution.

The Beta-Binomial CBs tries to model the uncertainty of the ECDF by placing priors on the conditional CDFs at each threshold as:

$$\begin{aligned} \text{Baseline Prior:} \quad & F_{n_1}(t) \sim \text{Beta}(1, 1) \\ \text{Baseline Prior:} \quad & F_{n_0}(t) \sim \text{Beta}(1, 1) \end{aligned}$$

After observing counts of the confusion matrix elements at each threshold, the posteriors become:

$$\begin{aligned} \text{Posterior:} \quad & F_{n_1}(t)|\text{data} \sim \text{Beta}(\text{TP}(t) + 1, \text{FN}(t) + 1) \\ \text{Posterior:} \quad & F_{n_0}(t)|\text{data} \sim \text{Beta}(\text{FP}(t) + 1, \text{TN}(t) + 1) \end{aligned}$$

We sample from these posteriors to generate multiple conditional CDFs:

$$F_{n_1}^{(b)}(t) \sim \text{Beta}(\text{TP}(t) + 1, \text{FN}(t) + 1) \tag{34}$$

$$F_{n_0}^{(b)}(t) \sim \text{Beta}(\text{FP}(t) + 1, \text{TN}(t) + 1) \tag{35}$$

The remaining steps for constructing CBs follow the general framework, with the same theoretical guarantee of coverage at probability at least $1 - \varepsilon$.

The quantity $F_n^{(b)}(t)$ that the CBs $[L_n(t), U_n(t)]$ are capturing with probability at least $1 - \varepsilon$ is the **aleatory uncertainty** of the ECDF. However, Bayesian approaches are limited by the choice of prior, which is a subjective decision that affects the uncertainty quantification.

4.3.3 Monte Carlo CBs

We now propose the main method of the thesis, which we name as Monte Carlo CBs. The reference CDF $F(t)$ is the true continuous CDF, representing the ground truth. The sampled functional forms $F_n^{(b)}(t)$ are derived from Monte Carlo simulations using the PIT. This method addresses the limitations of both bootstrap (finite sample size) and Bayesian approaches (subjective prior choice) by leveraging the theoretical properties of ordered statistics.

Let $\{\hat{p}_{(1)}, \hat{p}_{(2)}, \ldots, \hat{p}_{(n)}\}$ be a sequence of ordered random variables with a continuous CDF F, and ECDF:

$$F_n(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}[\hat{p}_{(j)} \leq t].$$

By the PIT, we have:

$$Y_{(i)} \overset{d}{=} F(\hat{p}_{(i)}) \quad \forall i \in \{1, 2, \ldots, n\}$$

where $Y_{(i)}$ is the i-th ordered uniform random variable between 0 and 1. For the ECDF, we have by definition:

$$F_n(\hat{p}_{(i)}) = \frac{i}{n}.$$

This relationship indicates that i/n represents the ECDF evaluated at the specific thresholds $\hat{p}_{(i)}$, providing a direct link between the uniform distribution and the empirical distribution of the data.

Using these properties, we define a distribution-free estimator of the non-conformity score:

$$\sup_{t \in \mathcal{T}} \frac{|F_n(t) - F(t)|}{\sigma(t)} \overset{d}{=} \max_{i \in \{1,2,\ldots,n\}} \frac{|\frac{i}{n} - Y_{(i)}|}{\sigma(\hat{p}_{(i)})} \overset{def}{=} \hat{s}_n.$$

By the definition of the standard deviation, we have:

$$\sigma(\hat{p}_{(i)}) = \sqrt{\text{Var}(Y_{(i)})}$$

Since $Y_{(i)} \sim \text{Beta}(i, n+1-i)$, we have:

$$\text{Var}(Y_{(i)}) = \frac{i(n+1-i)}{(n+1)^2(n+2)}.$$

The standard deviation is thus:

$$\sigma(\hat{p}_{(i)}) = \sqrt{\frac{i(n+1-i)}{(n+1)^2(n+2)}}.$$

To estimate the empirical quantile, we perform B Monte Carlo simulations by sampling from the Beta distribution:

$$Y_{(i)}^{(b)} \sim \text{Beta}(i, n+1-i),$$

for $b = 1, 2, \ldots, B$, and compute the non-conformity scores:

$$\hat{s}_n^{(b)} = \max_{i \in \{1,2,\ldots,n\}} \frac{|\frac{i}{n} - Y_{(i)}^{(b)}|}{\sqrt{\frac{i(n+1-i)}{(n+1)^2(n+2)}}}.$$

Following the general framework, we estimate the empirical quantile and construct the CBs at the ordered points $\hat{p}_{(i)}$. The bounds are constructed around i/n, the ECDF value at threshold $\hat{p}_{(i)}$ as

$$[L_n(\hat{p}_{(i)}), U_n(\hat{p}_{(i)})] = [i/n - 2 \cdot \hat{\delta}_{1-\varepsilon/2}(\hat{p}_{(i)}), i/n + 2 \cdot \hat{\delta}_{1-\varepsilon/2}(\hat{p}_{(i)})]$$

for test set coverage extension. Now, leveraging the PIT, we have $i/n = F_n(\hat{p}_{(i)})$, which yields the following coverage guarantee:

$$\mathbb{P}\left(\forall i \in \{1, 2, \ldots, n\}, F_{n_{\text{test}}}(\hat{p}_{(i)}) \in [L_n(\hat{p}_{(i)}), U_n(\hat{p}_{(i)})]\right) \geq 1 - \varepsilon,$$

with the prediction set:

$$[L_n(\hat{p}_{(i)}), U_n(\hat{p}_{(i)})] = \left[F_n(\hat{p}_{(i)}) - 2 \cdot \hat{\delta}_{1-\varepsilon/2}(\hat{p}_{(i)}), F_n(\hat{p}_{(i)}) + 2 \cdot \hat{\delta}_{1-\varepsilon/2}(\hat{p}_{(i)})\right].$$

The significant advantage of Monte Carlo CBs is that it allows for simulating CBs for an arbitrary calibration set size n. This flexibility enables generalization to a test set with an arbitrary number of samples n_{test} by choosing an n during construction that matches the anticipated n_{test}, providing a powerful tool for uncertainty quantification across varying sample sizes.

Implementation of Monte Carlo CBs is straightforward using the Python code in Listing 1. Note that since $\hat{\delta}_{1-\varepsilon}(\hat{p}_{(i)})$ is a function of n, ε, and B, the corresponding CBs are inherently distribution-free, which we demonstrate experimentally in Section 5.

```python
import numpy as np

def get_delta(n, epsilon=0.05, B=1000):
    i = np.arange(1, n + 1)
    CDF = np.random.beta(i, n + 1 - i, size=(B, n))
    ECDF = i / n
    var = i * (n + 1 - i) / ((n + 1) ** 2 * (n + 2))
    sigma = np.sqrt(var)
    scores = np.max(np.abs(ECDF - CDF) / sigma, axis=1)
    confidence = np.ceil((B + 1) * (1 - epsilon)) / B
    q = np.quantile(scores, confidence)
    delta = q * sigma
    return delta
```

Listing 1: Monte Carlo CBs Implementation.

5 Experiments

In this section, we conduct experiments to compare the methods bootstrap CBs, Beta-Binomial CBs, and Monte Carlo CBs. We also empirically verify the theoretical properties of the methods by performing coverage tests.

5.1 Experimental Setup

For our experiments, we generated synthetic data using the `make_moons` dataset with $n = 10000$ samples and a noise parameter of 0.5. We split the data into three parts: 60% for training the logistic regression model, 20% for calibration used for constructing the CBs. Note that we have the relationship $n = n_{\text{train}} + n_{\text{cal}} + n_{\text{test}}$, with $n_{\text{cal}} = n_{\text{test}}$.

For testing, we use `make_moons` with the random seed being 12387, to generate $100 \times n_{\text{cal}}$ pair of (x_i, y_i) samples, then make a batched inference using the previously trained model. The probability predictions are organized into a matrix of dimension $100 \times n_{\text{cal}}$, representing 100 exchangeable test sets each with the same sample size as the calibration set.

All experiments used a confidence level of $1 - \varepsilon = 0.95$, corresponding to 95% CBs using Equation (23) unless otherwise specified. The implementation used a threshold grid of 500 equally spaced points between 0 and 1, unless otherwise specified. For bootstrap-based methods, we generated $B = 1000$ samples. All experiments were conducted with a fixed random seed 42 to ensure reproducibility.

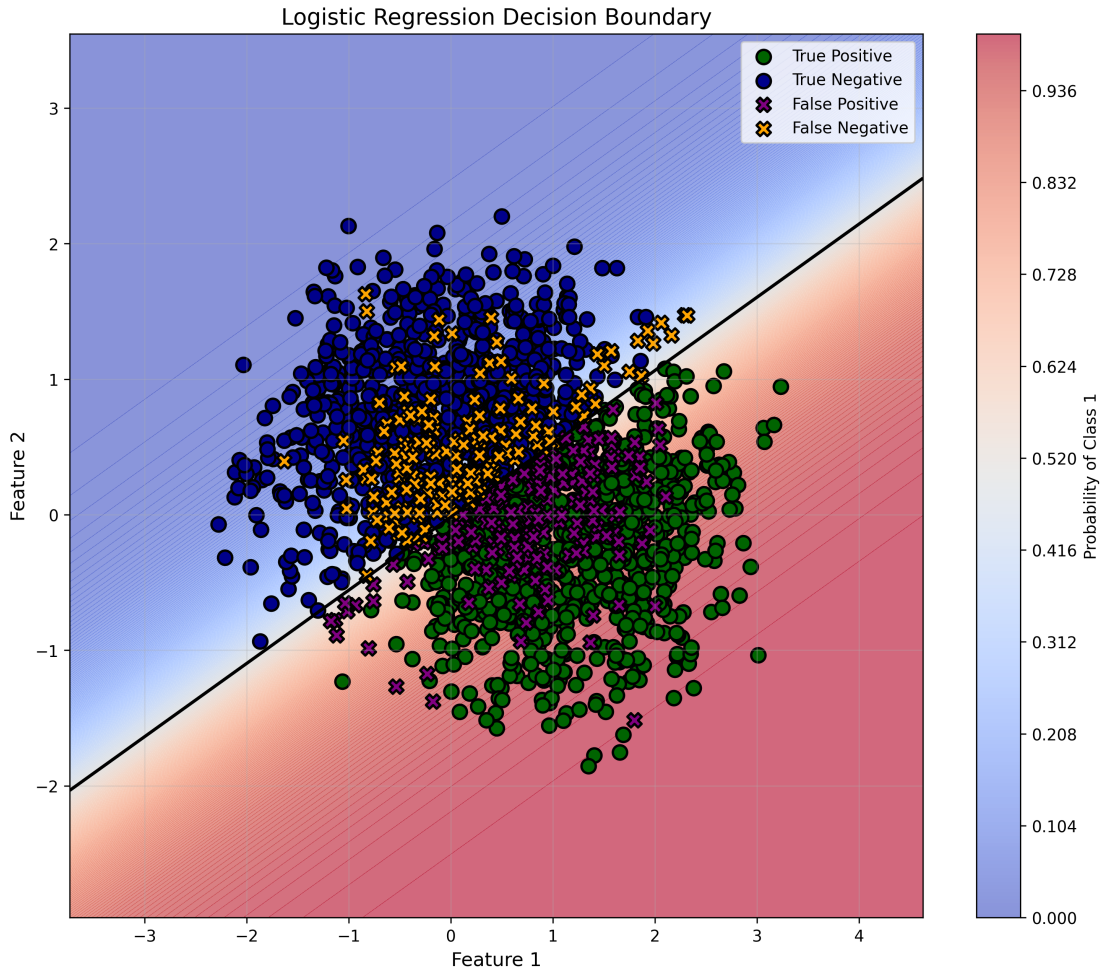

Figure 3 A logistic regression model trained on the `make_moons` dataset with 60% training. The markers and the decision boundary are created based on the calibration set.

5.2 Consistency of Bootstrap CBs

The theory of bootstrap, see Equation (1), suggests that the bootstrap CBs is consistent, that is, with a large n_{cal} and B, the CBs of bootstrap will be close to the true CBs. We will benchmark the bootstrap CBs against the DKW based CBs as it provides us with an explicit form of the CBs. A central question we aim to investigate is:

> **Does the bootstrap CBs demonstrate the theoretically expected convergence behavior with respect to sample size and significance level?**

We will set $\sigma_n(t) = 1$ for all $t \in [0, 1]$ to match the formulation of the DKW inequality:

$$\hat{\delta}_{1-\varepsilon}(t) = \hat{\delta}_{1-\varepsilon} \approx \sqrt{\frac{\ln(2/\varepsilon)}{2n_{\text{cal}}}}.$$

The first part is to fix $\varepsilon = 0.05$ and $B = 1000$, and then repeating the experiment for various values of $n \in \{500, 2000, 8000, 32000\}$ to study the relationship between BVACB and DKW based bounds, see Figure 4.

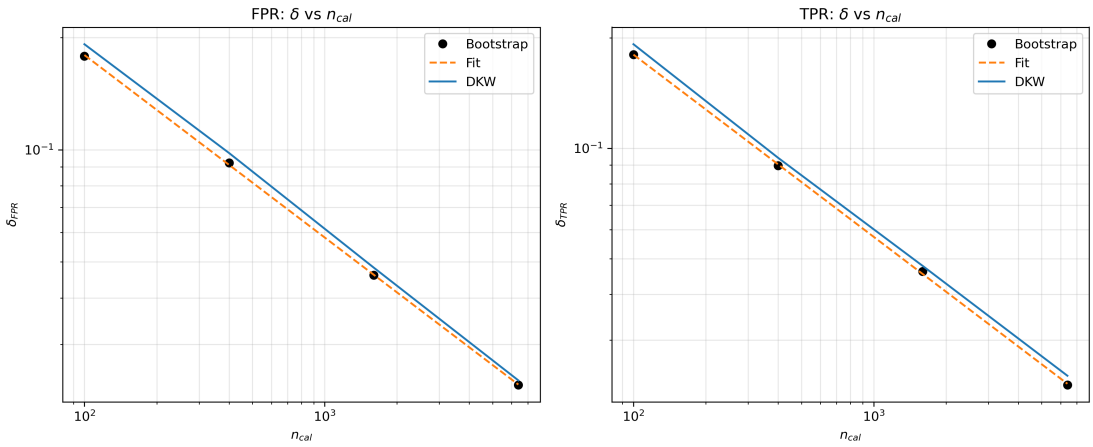

Figure 4 Studying the relationship $\hat{\delta}_{1-\varepsilon} \sim \Theta(n^{-1/2})$ for FPR and TPR respectively. The slope of the line fitted from the datapoints aligns well with the theoretical DKW reference, indicating an inverse square root law.

What we see is that the estimated δ_{FPR} and δ_{TPR} closely matches the theoretical values proposed by the DKW inequality. This is an indication that $\hat{\delta}_{1-\varepsilon} \sim \Theta(n^{-1/2})$ for bootstrap CBs.

Now, we perform a similar study by fixing $n = 10000$, but varying the significance level $\varepsilon \in \{0.64, 0.16, 0.04, 0.01\}$. A good fit relative to the corresponding outcome from the DKW inequality would indicate a $\hat{\delta}^2_{1-\varepsilon} \sim \Theta(\log(1/\varepsilon))$ relationship.

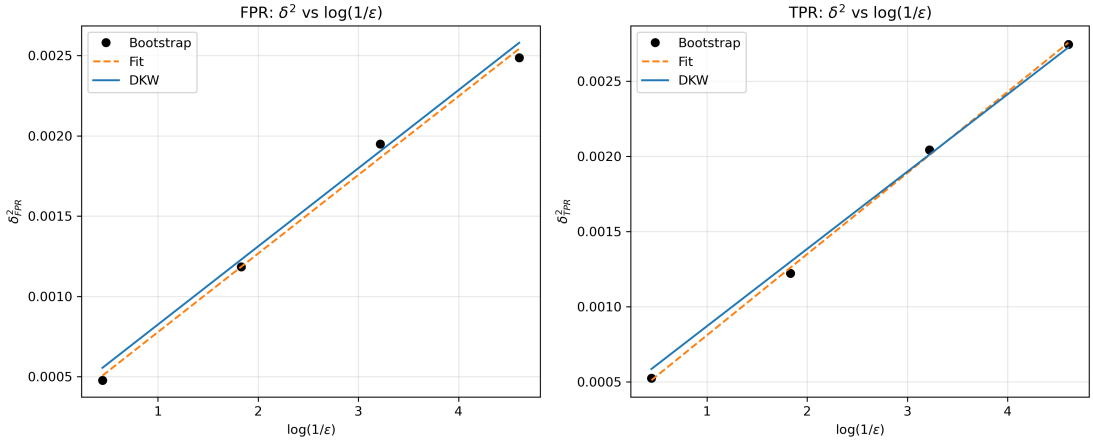

Figure 5 Studying the relationship $\hat{\delta}^2_{1-\varepsilon} \sim \Theta(\log(1/\varepsilon))$ for FPR and TPR respectively. The slope of the line fitted from the datapoints aligns well with the theoretical DKW reference, indicating a logarithmic law.

In conclusion, experiments show that across several orders of magnitudes, the bootstrap CBs is consistent for $\sigma_n(t) = 1$, aligning with the theoretical property of bootstrap, yielding a good approximation in the form of $\hat{\delta}_{1-\varepsilon} \approx \sqrt{\frac{\ln(2/\varepsilon)}{2n_{\text{cal}}}}$.

5.3 Comparing Monte Carlo with Bootstrap CBs

We now show that the Monte Carlo CBs is capable of modeling the epistemic uncertainty that arises from finite number of samples from the calibration set. We aim to answer the following question:

> **How does the Monte Carlo CBs perform against bootstrap CBs as the number of samples n is varying?**

To answer it, we set up three expreiments by varying the total number of samples n for the data generating process `make_moons`, this means that the corresponding training, calibration, and test set will also vary. As these sets changes, we expect that the distribution for the test set will vary, the goal is thus to verify whether if the Monte Carlo CBs will faithfully account for the distribution shifts due to increased sample size. Previously, we have shown that bootstrap is consistent with respect to the DKW inequality. We will assume that the consistency property generalizes for the variance-adaptive correspondence. Therefore, we should expect that the bounds from the Monte

Carlo CBs will have a similar behavior as the bootstrap CBs reference. We demonstrate the behavior of the Monte Carlo CBs in Figure 6.

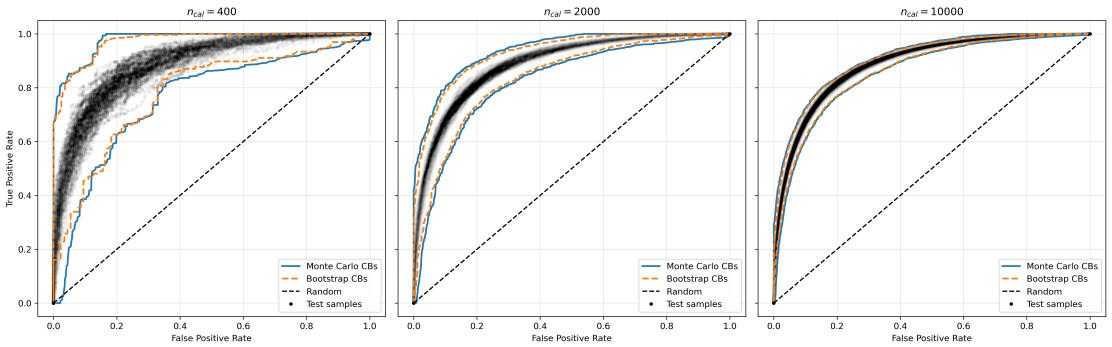

Figure 6 Comparing the Monte Carlo CBs with bootstrap CBs where the CBs are constructed using the calibration set. The point-clouds correspond to the PDF of the 100 test curves visualized as black point-clouds. We observe that both bootstrap CBs and Monte Carlo CBs are able to faithfully construct the CBs. The Monte Carlo CBs are consistently wider than bootstrap CBs because it accounts for the sampling bias from the calibration set. However, the difference between these two methods have a tendency to converge as the sample size increases, which is expected considering that the bootstrap CBs is consistent. This holds across two orders of magnitude of sample size, validating our theoretical analysis.

The consistent narrowness of bootstrap CBs compared to Monte Carlo CBs can be further explained through the triangle inequality. When considering the total uncertainty in predicting a test set CDF from a calibration set, we can decompose it as $|F_{\text{cal}} - F_{\text{test}}| \leq |F_{\text{cal}} - F_{\text{true}}| + |F_{\text{true}} - F_{\text{test}}|$, where F_{cal} is the calibration ECDF, F_{test} is the test ECDF, and F_{true} is the true underlying CDF. Bootstrap methodology inherently focuses only on the variability within the calibration set (aleatory uncertainty), effectively treating $|F_{\text{cal}} - F_{\text{true}}|$ as negligible or underestimating it. In contrast, Monte Carlo CBs account for both components of uncertainty, the epistemic uncertainty between the empirical and true CDFs, as well as the sampling variability. This fundamental difference in uncertainty quantification explains why Monte Carlo CBs yield wider bands, particularly at smaller sample sizes where the epistemic uncertainty is more pronounced. As sample size increases and $|F_{\text{cal}} - F_{\text{true}}| \to 0$, both methods naturally converge, as observed in the rightmost panel of Figure 6.

5.4 Monte Carlo CBs on ECG Data

In this part, we demonstrate how Monte Carlo CBs performs on real world data by conducting an experiment on ECG data. We then use a deep neural network to perform a binary classification task. The key question guiding this experiment is:

> **Can the Monte Carlo CBs method maintain its theoretical properties when applied to complex real-world data and sophisticated model architectures?**

We aim to convince the reader that Monte Carlo CBs works independently of the model class and dataset, only the predicted probabilities and the corresponding labels are needed. We visualize the CBs as well as the threshold-wise intervals in Figure 7 with the same data proportions as the `make_moons` dataset, e.g., 60% training, 20% calibration, and 20% test. The intervals are computed using

$$\hat{s}_n^{(b)}(t) = \frac{|F_n^{(b)}(t) - F(t)|}{\sigma(t)},$$

while the global CBs are computed using

$$\hat{s}_n^{(b)} = \sup_{t \in \mathcal{T}} \frac{|F_n^{(b)}(t) - F(t)|}{\sigma(t)}.$$

Since we have a limited calibration set, the point clouds are approximated using bootstrap for reference.

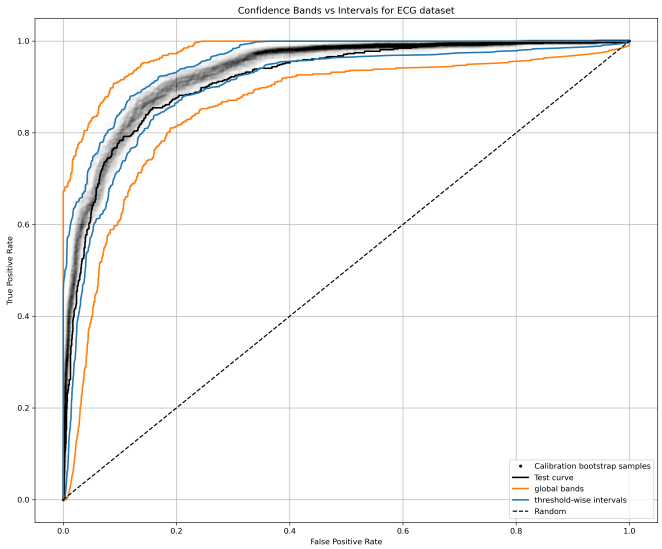

Figure 7 Comparing threshold-wise confidence intervals and global CBs for the Monte Carlo CBs applied on an ECG dataset with probabilities predicted by a deep neural network. The ROC curves are computed using the adaptive thresholds. The point clouds are approximated using bootstrap for reference. We observe that the threshold-wise intervals are much tighter than the global CBs, faithfully covering the bootstrap samples, but misses some parts of the test curve around FPR ≈ 0.39.

Experiments show that the Monte Carlo CBs is agnostic to the model class and dataset, which aligns with the theory. This is because the non-conformity score is distribution free as suggested in Section 4.3.3.

Furthermore, we see that the threshold-wise intervals roughly reflects the behaviour of the bootstrap samples. This is expected because of the consistency property of the bootstrap CBs, the sample size of $n_{\text{cal}} = 2000$ is sufficiently large to provide a good approximation of the true CDF. We can alternatively interpret the Monte Carlo quantile-based confidence intervals as a bootstrap-like method, but the intervals does not suffer from sampling bias from the calibration set.

5.5 Impact of Threshold Choice

Constructing a CDF involves the selection of thresholds used to map classifier scores to binary predictions. One option is to impose an external structure, such as a fixed number of evenly spaced thresholds spanning the potential score range. Alternatively, one can adopt a data-driven approach, using the distinct score values observed in the

dataset as the natural threshold points, e.g., the thresholds are defined according to the ordered random variables with $t_i = \hat{p}_{(i)}$. Through our experiments, we address the critical question:

> **How does the choice between fixed and adaptive thresholds affect the coverage properties of Monte Carlo CBs, particularly at the extremes of the distribution?**

We show that there is a difference in the performance of Monte Carlo CBs when using these two different threshold choices.

We begin by evaluating the performance of Monte Carlo CBs given a Beta distribution, this allows us to generate arbitrary number of i.i.d. test points as well as the true CDF for reference, see Figure 8.

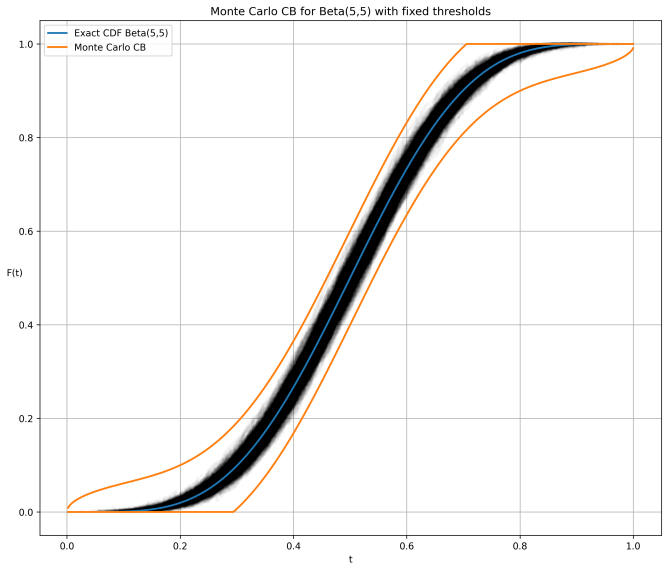

Figure 8 Plotting the Monte Carlo CBs for the Beta distribution with parameters $\alpha = 5$ and $\beta = 5$ with a fixed number of thresholds given by $t_i = \frac{i}{n+1}$ for $i = 1, 2, \ldots, n$ with $n = 500$. The black point clouds are simulated empirical CDFs based on samples from the Beta distribution.

We then compare Figure 8 with the corresponding results when using adaptive thresholds, see Figure 9.

A critical observation when comparing these two figures is the difference in how the CBs behave at the extremes of the interval. In Figure 8, using fixed thresholds, the

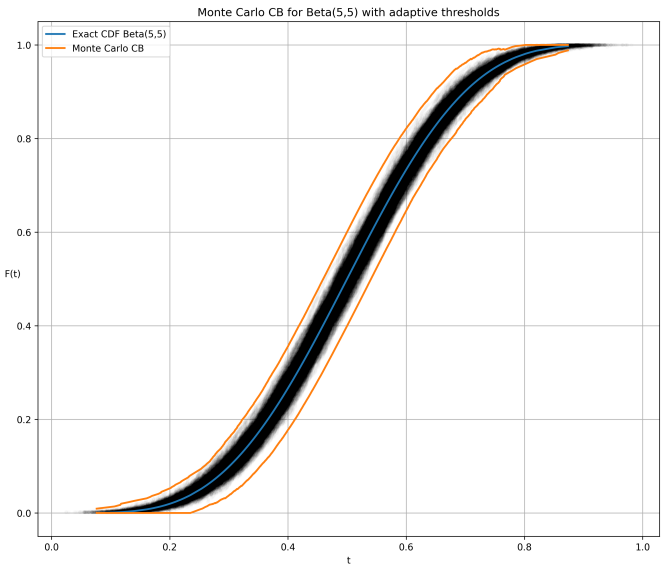

Figure 9 Plotting the Monte Carlo CBs for the Beta distribution with parameters $\alpha = 5$ and $\beta = 5$ with adaptive thresholds, e.g., $t_i = \hat{p}_{(i)}$. The black point clouds are simulated empirical CDFs based on samples from the Beta distribution.

orange CBs fully cover the entire $[0, 1]$ interval, extending all the way to both boundary points. In contrast, Figure 9 shows that with adaptive thresholds, the orange bands do not completely cover the extremes, they begin slightly above 0 and end slightly below 1.

This limitation of adaptive thresholds arises from their dependence on observed data points. Since adaptive thresholds are derived from order statistics ($t_i = \hat{p}_{(i)}$), they are inherently bounded by the minimum and maximum values observed in the calibration set. This means that for rare events in the tails of the distribution, where no samples might be observed in a finite dataset, the adaptive threshold approach provides no coverage guarantees. Fixed thresholds are predetermined to span the entire theoretical range of the random variable regardless of the observed data, ensuring coverage across the full domain. The trade-off is that fixed thresholds might allocate computational resources to regions with no data, while adaptive thresholds concentrate on regions where data is actually observed, potentially providing tighter bounds in those areas. However, this comes at the cost of potentially missing coverage in the extreme tails of the distribution.

5.6 Empirical Coverage Test for Monte Carlo CBs

We perform a coverage test to verify the theoretical properties of Monte Carlo CBs. The key question we aim to answer is:

> **How does the empirical coverage of Monte Carlo CBs compare to the expected theoretical coverage across different sample sizes and significance levels?**

To answer this question systematically, we conduct a comprehensive ablation study with the following varying variables:

- Sample sizes $n \in \{10, 25, 50, 400, 2000, 10000\}$
- Significance levels $\varepsilon \in \{0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2\}$

For each combination, we generate $B = 5000$ bootstrap samples and an equal number of test curves from a uniform distribution and construct CBs around the true CDF. We then measure what fraction of the ECDFs fall completely within these bands, giving us the empirical coverage rate.

In a properly calibrated CB, the empirical coverage should closely match the expected theoretical coverage of $1 - \varepsilon$. Any discrepancy between these values indicates potential over-coverage or under-coverage. Figure 10 presents the results as a heatmap of discrepancies between the empirical and expected coverage (empirical$-$expected). Several interesting patterns emerge:

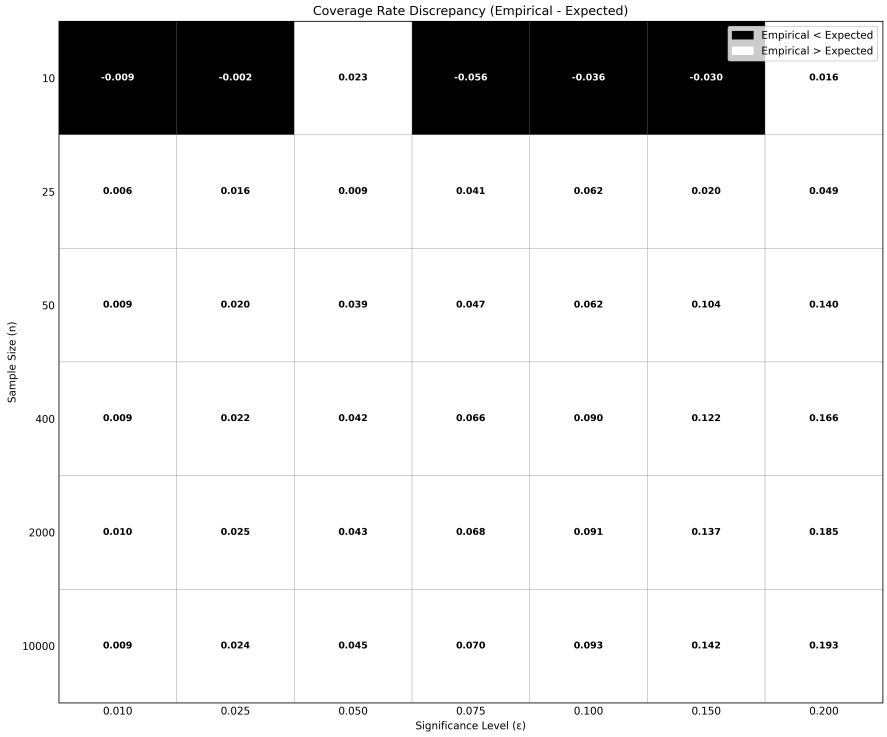

Figure 10 Discrepancy between empirical and expected coverage rates for CBs across different sample sizes and significance levels. Black cells indicate regions where empirical coverage is lower than expected, while white cells show where empirical coverage exceeds the expected rate. The numerical value in each cell represents the exact magnitude of the discrepancy. We observe that in the case of small sample sizes, the Monte Carlo CBs is under-covering, while for large sample sizes, the Monte Carlo CBs is over-covering.

- **Sample size effect:** At the smallest sample size ($n = 10$), the CBs exhibit a mixed behavior with predominantly negative discrepancies (under-coverage) for most significance levels, notably at $\varepsilon \in \{0.01, 0.025, 0.075, 0.1, 0.15\}$, with the exception of positive discrepancies at $\varepsilon = 0.05$ and $\varepsilon = 0.2$. In stark contrast, for all larger sample sizes ($n \geq 25$), the CBs consistently demonstrate over-coverage with positive discrepancies across all significance levels.
- **Monotonic increase with sample size:** For each fixed significance level, the discrepancy generally increases monotonically with sample size. This indicates that the bands become progressively more conservative as sample size increases, contrary to the expectation that larger samples should lead to more precise coverage.
- **Significance level effect:** The largest positive discrepancies occur at higher significance levels ($\varepsilon = 0.15$ and $\varepsilon = 0.2$) and larger sample sizes, reaching values as

high as 0.193 for $n = 10000$ and $\varepsilon = 0.2$. This suggests that the Monte Carlo CBs are particularly conservative when both the sample size and significance level are large.

- **Consistent over-coverage at moderate and large sample sizes:** For practical applications with moderate to large datasets ($n \geq 25$), the CBs consistently provide more coverage than theoretically required, which can be beneficial from a risk-averse perspective but may lead to unnecessarily wide bands.

These findings reveal an interesting property of Monte Carlo CBs: rather than converging exactly to the expected theoretical coverage as sample size increases, they tend to maintain and even increase their conservative nature. This systematic over-coverage appears to be an intrinsic characteristic of the method rather than a statistical anomaly, and may be attributed to the inherent properties of the supremum-based non-conformity score used in Monte Carlo CBs. These results suggest that Monte Carlo CBs provides reliable (albeit conservative) CBs across a wide range of sample sizes and significance levels, with the only case of potential under-coverage occurring at very small sample sizes ($n = 10$).

5.7 Beta-Binomial CB for ROC Curves

After establishing the coverage properties of both Monte Carlo CBs and Beta-Binomial CBs through controlled experiments, we now compare their performance in constructing CBs for ROC curves in a more realistic setting. In this comparison, we investigate:

> **How do the Beta-Binomial CBs compare with Monte Carlo CBs in terms of width, shape, and convergence properties as sample size increases?**

We use the same experimental setup as the previous section 5.3, but with the Beta-Binomial approach instead of bootstrap to construct the CBs.

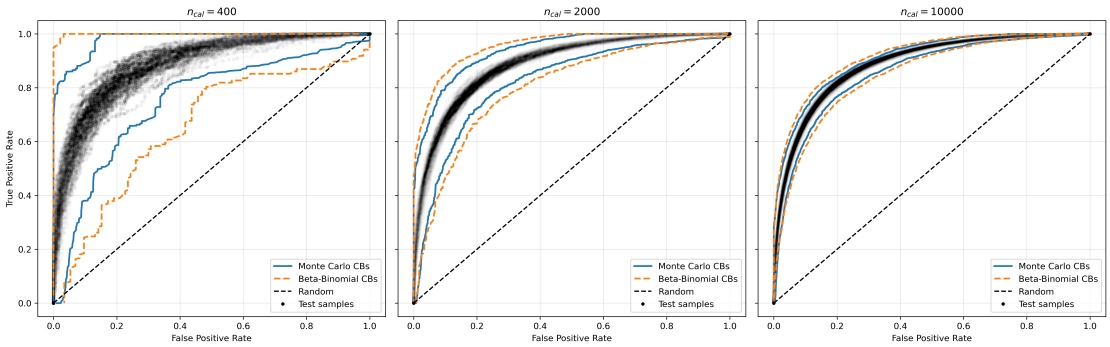

Figure 11 Comparison of Monte Carlo CBs and Beta-Binomial CBs for ROC curves across different calibration set sizes. Black dots represent 100 test set ROC curves.

Figure 11 illustrates the CBs constructed by both methods across three different calibration set sizes. Several key observations emerge:

- **Convergence with increasing sample size:** For small calibration sets ($n_{\text{cal}} = 400$), both methods produce relatively wide CBs, with the Beta-Binomial approach yielding slightly wider bands particularly in the middle sections of the ROC curve. As the calibration set size increases to $n_{\text{cal}} = 2000$ and further to $n_{\text{cal}} = 10000$, the bands from both methods narrow considerably and converge toward similar boundaries.

- **Variance adaptation:** Both methods demonstrate variance-adaptive properties, with wider bands in regions of higher uncertainty (typically in the middle sections of the ROC curve) and narrower bands near the extremes where variance is inherently lower. This illustrates the effectiveness of the variance scaling approach in both methods.

- **Prior influence:** The Beta-Binomial CBs tend to be consistently wider than Monte Carlo CBs across different calibration set sizes, particularly noticeable in the smaller sample size regime. This is likely due to the uniform prior (Beta$(1, 1)$) adding pseudo-counts to the confusion matrix elements, which systematically over-estimates uncertainties compared to the distribution-free Monte Carlo approach. The effect of this prior over-estimation diminishes as sample size increases, which explains the convergence between methods at larger n_{cal} values.

- **Test curve containment:** Visual inspection suggests that both methods successfully contain the vast majority of the 100 test ROC curves (shown as black dots), even at the smallest calibration set size. This aligns with our earlier empirical coverage tests showing that both methods tend toward over-coverage, particularly at larger sample sizes.

- **Practical differences:** For $n_{\text{cal}} = 400$, the Beta-Binomial bands appear more jagged and less smooth compared to Monte Carlo CBs, likely due to the direct influence of the observed confusion matrix counts on the Beta posterior parameters. At larger sample sizes, this difference becomes negligible.

These visual comparisons substantiate our earlier findings from the empirical coverage tests. The similarity in band shape and width between the two methods, especially at larger sample sizes, reinforces our conclusion that the behavior is inherent to the variance-adaptive conformal prediction approach rather than to the specific method of generating samples.

The practical implications are significant: practitioners can choose either method based on computational considerations or conceptual preference without substantial differences in the resulting CBs, particularly for moderate to large calibration sets. The Beta-Binomial approach may be conceptually simpler for those familiar with Bayesian statistics, while the Monte Carlo CBs method connects more directly to the theory of order statistics and empirical CDFs.

Regardless of the chosen method, our experiments demonstrate that both approaches provide robust uncertainty quantification for ROC curves, with CBs that appropriately adapt to regions of varying uncertainty while maintaining the target coverage level.

5.8 Empirical Coverage Test for Beta-Binomial CBs

We investigate whether if the Beta-Binomial CBs is a valid candidate for constructing CBs with desired coverage rate. We aim to answer the following question:

> **How does the empirical coverage of CBs constructed using the Beta-Binomial CBs compare to the expected theoretical coverage across different sample sizes and significance levels?**

The experimental set-up is similar to the previous section, except that we use the Beta-Binomial model to construct the CBs. We then measure the empirical coverage rate of the CBs and compare it to the expected theoretical coverage.

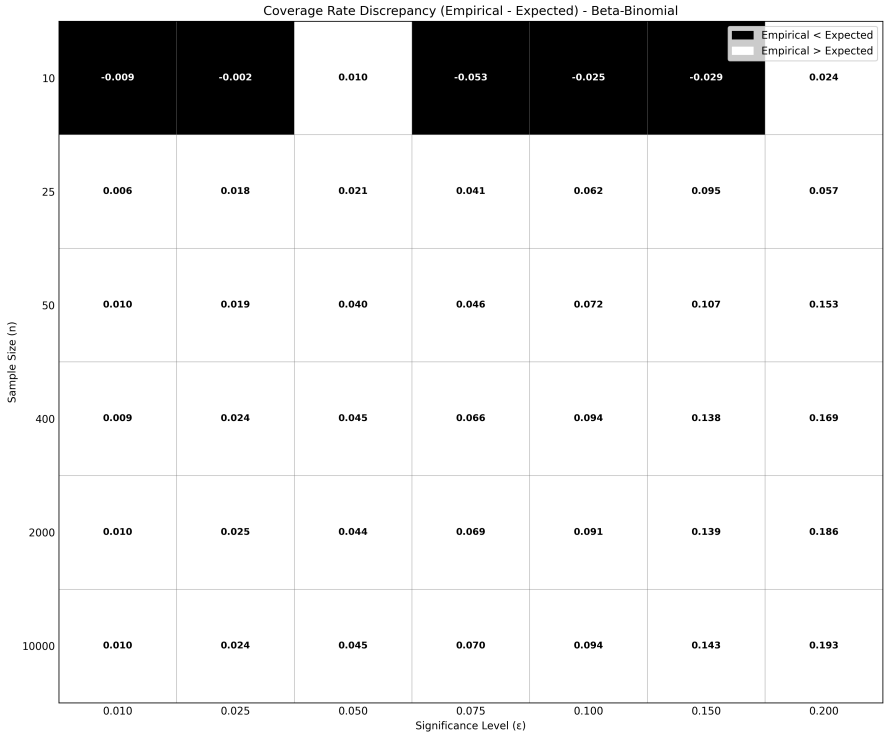

Figure 12 Discrepancy between empirical and expected coverage rates for CBs across different sample sizes and significance levels. Black cells indicate regions where empirical coverage is lower than expected, while white cells show where empirical coverage exceeds the expected rate. The numerical value in each cell represents the exact magnitude of the discrepancy. We observe that in the case of small sample sizes, the Beta-Binomial CBs is under-covering, while for large sample sizes, the Beta-Binomial CBs is over-covering.

Figure 12 displays the discrepancy between empirical and expected coverage for the Beta-Binomial CBs. Overall, we observe a similar result as the Monte Carlo CBs. These parallel findings suggest that the observed pattern is not unique to either method but may be inherent to the approach of constructing CBs for CDFs using the variance-adaptive non-conformity scores. The Beta-Binomial CBs, despite its different mathematical foundation, produces quantitatively similar coverage behavior as the Monte Carlo CBs. However, the CBs are consistently wider due to the uninformative uniform prior.

5.9 Empirical Coverage Test Using the Standardized L2 Norm

To further investigate the source of the over-coverage phenomenon observed with the supremum-norm, we conducted an additional experiment using the Monte Carlo CBs framework, but replaced the supremum-norm in the non-conformity score with the standardized L2 norm, see Equation (11). We aim to answer the following question:

> **Is the over-coverage phenomenon a consequence of using the standardized supremum-norm as the non-conformity score?**

The experiment was implemented using the same Monte Carlo simulation approach as before, with the only exception that the empirical coverage is computed by checking whether if the non-conformity scores for new test instances are below the empirical quantile. This is mathematically equivalent of explicitly checking each point of the CDF [AB22]. The results, shown in Figure 13, reveal that the empirical coverage rates do not exhibit a consistent pattern of over-coverage as sample size increases.

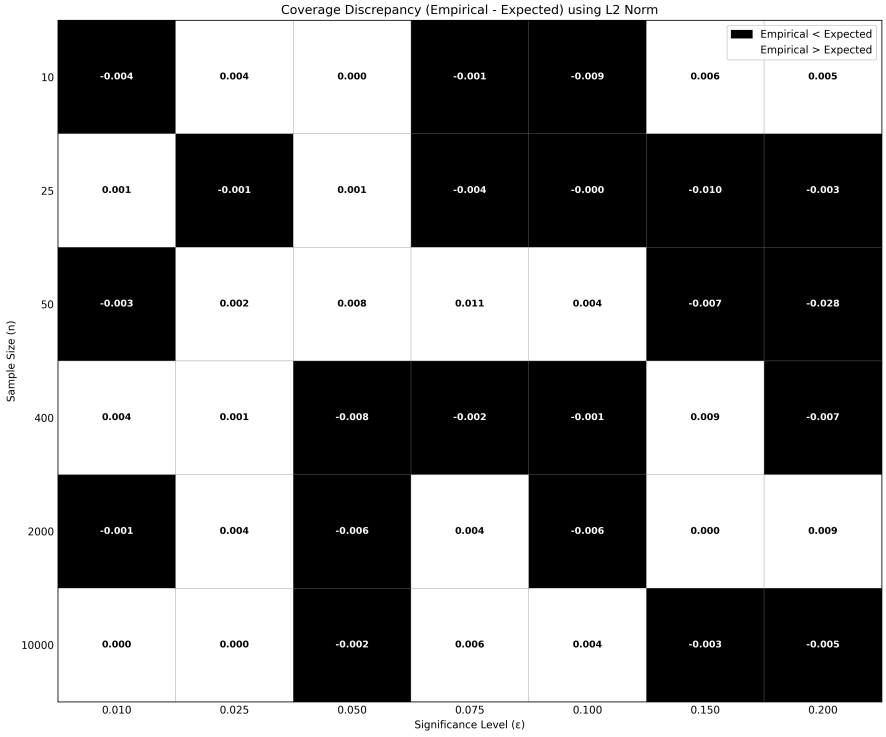

Figure 13 Coverage Discrepancy (Empirical - Expected) using standardized L2 Norm as the non-conformity score for Monte Carlo CBs. Black cells indicate empirical coverage below expected, white cells above. No systematic over-coverage is observed.

This experiment supports the hypothesis that the over-coverage observed with the supremum-norm is a consequence of its scaling properties, rather than an inherent flaw in the variance-adaptive conformal prediction framework. When the numerator part of the non-conformity score is scaled by a non-comparable denominator, the empirical coverage tends to be missaligned.

6 Discussion

This thesis has addressed the challenge of how to reliably quantify uncertainty in the performance evaluation of binary classifiers. The need for robust uncertainty quantification in ROC curves arises from the inherent variability in model performance across different datasets and the critical importance of reliable decision-making in high-stakes applications. Our work has yielded several successful approaches for constructing CBs around ROC curves, with the Monte Carlo CBs, demonstrating particularly strong properties in terms of conditional coverage, finite-sample guarantees, distribution-free, and variance adaptivity.

6.1 Theoretical Connections Between Methods

An interesting theoretical insight emerged from our investigation of the three uncertainty quantification methods: despite their different mathematical foundations, Monte Carlo CBs, Beta-Binomial CBs, and bootstrap CBs exhibit remarkably similar asymptotic behavior. This connection helps explain why our experimental results revealed increasingly similar performance metrics and coverage patterns as sample size increased.

6.1.1 Asymptotic Behavior of Monte Carlo CBs

In the Monte Carlo approach, we directly leverage the distribution of order statistics through the PIT, where the i-th order statistic follows a Beta distribution:

$$Y_{(i)} \sim \text{Beta}(i, n+1-i)$$

The variance-adaptive scaling factor in this method is:

$$\sigma(\hat{p}_{(i)}) = \sqrt{\frac{i(n+1-i)}{(n+1)^2(n+2)}}$$

For large n and when i is proportional to n (i.e., $i/n \approx F(t)$), this expression approaches:

$$\begin{aligned}\sigma(\hat{p}_{(i)}) &= \sqrt{\frac{i(n+1-i)}{(n+1)^2(n+2)}} \\ &\approx \sqrt{\frac{nF(t)(n-nF(t))}{n^2 \cdot n}} \\ &= \sqrt{\frac{F(t)(1-F(t))}{n}}\end{aligned}$$

This is precisely the standard form of the variance for a binomial proportion scaled by sample size, which appears in variance-adaptive concentration inequalities such as the one by Bartl and Mendelson [BM23]. This demonstrates that Monte Carlo CBs intrinsically adapt to the local variance structure of the CDF without requiring asymptotic assumptions, while still converging to the theoretically expected form as n increases.

6.1.2 Asymptotic Behavior of Beta-Binomial CBs

The Beta-Binomial approach models uncertainty through Beta posteriors at each threshold:

$$\begin{aligned}F_{n_1}(t)|\text{data} &\sim \text{Beta}(\text{TP}(t)+1, \text{FN}(t)+1) \\ F_{n_0}(t)|\text{data} &\sim \text{Beta}(\text{FP}(t)+1, \text{TN}(t)+1)\end{aligned}$$

As the sample size increases, these Beta distributions approach normality according to the Bernstein-von Mises theorem (see Section 2.5.2). For large n, if we denote $\text{TPR}(t) = \text{TP}(t)/(\text{TP}(t)+\text{FN}(t))$ and $\text{FPR}(t) = \text{FP}(t)/(\text{FP}(t)+\text{TN}(t))$, then:

$$\begin{aligned}\text{Beta}(\text{TP}(t)+1, \text{FN}(t)+1) &\approx \mathcal{N}\left(\text{TPR}(t), \frac{\text{TPR}(t)(1-\text{TPR}(t))}{n_1}\right) \\ \text{Beta}(\text{FP}(t)+1, \text{TN}(t)+1) &\approx \mathcal{N}\left(\text{FPR}(t), \frac{\text{FPR}(t)(1-\text{FPR}(t))}{n_0}\right)\end{aligned}$$

This shows that for large sample sizes, the Beta-Binomial approach effectively constructs normal confidence intervals with variance scaling according to the same $\frac{p(1-p)}{n}$ formula that emerges in the asymptotic behavior of Monte Carlo CBs.

6.1.3 Asymptotic Behavior of Bootstrap CBs

The bootstrap approach relies on resampling to estimate the sampling distribution of the ECDF. For a specific threshold t, the ECDF $F_n(t)$ represents an average of i.i.d. Bernoulli random variables with parameter $p = F(t)$. By the Central Limit Theorem, as n increases:

$$\sqrt{n}(F_n(t) - F(t)) \xrightarrow{d} \mathcal{N}(0, F(t)(1 - F(t)))$$

This implies that for large n, the variance of $F_n(t)$ is approximately $\frac{F(t)(1-F(t))}{n}$. The bootstrap variance estimator converges to this theoretical variance as both n and the number of bootstrap samples B increase:

$$\frac{1}{B-1}\sum_{b=1}^{B}\left(F_n^{(b)}(t) - \bar{F}_B(t)\right)^2 \xrightarrow{p} \frac{F(t)(1-F(t))}{n}$$

6.1.4 Unified Asymptotic Behavior

What emerges from this analysis is a striking convergence: all three methods, Monte Carlo, Beta-Binomial, and bootstrap, utilize different theoretical foundations but ultimately converge to the same asymptotic variance structure:

$$\sigma^2(t) = \frac{F(t)(1-F(t))}{n}$$

This explains the similar patterns observed in our empirical coverage tests, particularly for moderate to large sample sizes, see Figures 10 and 12 in Section 5. The coverage rate discrepancies for both Monte Carlo CBs and Beta-Binomial CBs exhibit nearly identical patterns as sample size increases, with both methods showing consistent over-coverage of similar magnitude for $n \geq 25$.

While the methods differ in how they quantify uncertainty for small sample sizes, with Monte Carlo CBs providing exact finite-sample guarantees through order statistics, Beta-Binomial CBs leveraging conjugate priors, and bootstrap CBs using empirical resampling, they converge to equivalent behavior as the sample size increases. This unification of seemingly disparate approaches under a common asymptotic framework provides a deeper theoretical understanding of uncertainty quantification for CDFs and ROC curves.

6.2 Limitations and Assumptions

While our methods demonstrate strong theoretical properties and empirical performance, several limitations and assumptions should be acknowledged:

6.2.1 Threshold Dependency

As demonstrated in the experiments, the choice between fixed and adaptive thresholds can impact the performance of the CBs. While adaptive thresholds generally performed better in our experiments, this highlights the sensitivity of the methods to implementation choices.

6.2.2 Conservativeness of Conditional Coverage

To bound any future test curve, we had to apply triangle inequality and union bound to the non-conformity score, which resulted in CBs that was effectively more than twice as wide as the CBs for achieving marginal coverage. Although strong theoretical guarantees are achieved, the CBs became very wide.

6.2.3 Limitation in the Beta-Binomial Approach

A fundamental limitation in the Beta-Binomial approach is the independent modeling of TPR and FPR at each threshold. The posterior distributions are defined separately:

$$\begin{aligned} \text{Posterior:} \quad & F_{n_1}(t)|\text{data} \sim \text{Beta}(\text{TP}(t)+1, \text{FN}(t)+1) \\ \text{Posterior:} \quad & F_{n_0}(t)|\text{data} \sim \text{Beta}(\text{FP}(t)+1, \text{TN}(t)+1) \end{aligned}$$

This modeling choice means the joint uncertainty of TPR and FPR at any threshold is represented as a product of independent distributions rather than a true bivariate distribution. Consequently, the Beta-Binomial model fails to capture the natural correlation between TPR and FPR that exists in practice, particularly for classifiers that produce similar score distributions for positive and negative classes. This independence assumption results in a loss of information about the joint behavior of confusion matrix elements.

6.3 Future Work

As this thesis has laid a robust foundation for uncertainty quantification in binary classification through CBs for ROC curves and CDFs, several promising directions emerge for future research. These directions aim to address current limitations and extend the applicability of the proposed methods, particularly in real-world, high-stakes domains such as healthcare. Below, we outline key areas of exploration that build on the theoretical and empirical insights developed in this work.

6.3.1 Adaptation to Distribution Shifts

A critical assumption in the current framework (e.g., Monte Carlo CB) is dataset-level exchangeability between calibration and test sets, as discussed in Section 4.1.1. However, real-world applications often encounter distribution shifts, where the test data distribution diverges from the training or calibration data due to evolving patient demographics, changing environmental conditions, or equipment variations. Such shifts can undermine the coverage guarantees of our CBs, limiting their reliability in dynamic settings.

Future work will focus on developing robust uncertainty quantification methods that maintain validity under distribution shifts, drawing on advances in distributionally robust learning and covariate shift correction. One approach could involve integrating importance weighting into the conformal prediction framework, estimating density ratios between calibration and test distributions to adjust non-conformity scores or sampling weights in Monte Carlo and bootstrap processes. Additionally, exploring online updating mechanisms, where CBs are dynamically refined as new data arrives could enable real-time adaptation to detected shifts.

6.3.2 Generalization to Multi-Class and Multi-Label Classification

The current thesis focuses on binary classification, where ROC curves are directly represented as transformations of ECDFs conditional on the class label. However, many critical applications, particularly in medical diagnostics, involve multi-class or multi-label classification tasks—such as detecting multiple cardiac conditions from ECG data—where performance evaluation and uncertainty quantification are more complex due to inter-class dependencies and higher-dimensional metrics.

Future research will extend the conformal prediction framework to multi-class settings by constructing CBs for one-vs-rest ROC curves or higher-dimensional metrics

like the Volume Under the ROC Surface. A key challenge is modeling dependencies between class-specific CDFs, which could be addressed through multivariate nonconformity scores or copula-based approaches. For multi-label settings, per-label CBs for precision-recall curves or aggregated metrics like Hamming loss will be explored, incorporating label correlations via graphical models. Theoretical efforts will focus on deriving finite-sample coverage guarantees for these higher-dimensional functionals, building on the principles established in Theorem 4.1.

6.3.3 Adressing the Conservativeness of Conditional Coverage

A key limitation in the current framework for constructing CBs around ROC curves is the conservative assumption of a worst-case scenario, where the calibration curve is positioned at an extreme opposite to the test curve. This assumption, enforced through the triangle inequality for conditional coverage guarantees, results in CBs that are effectively twice as wide as those for marginal coverage. However, with a moderately large and balanced calibration set, such extreme divergence is unlikely to occur in practice. Addressing this conservativeness by evaluating the balance of the calibration set offers a pathway to significantly reduce CB widths, potentially by a factor close to 2, thereby enhancing the practical utility of uncertainty quantification in binary classification.

Future research could explore the use of k-fold cross-validation to measure the degree of balance within the calibration set and adjust the conservativeness of CBs accordingly. By splitting the calibration set into k subsets (folds) and computing ROC curves for each fold, we can examine the aleatory uncertainty across these subsets to infer balance or imbalance. High variability in ROC curves across folds would indicate potential imbalances, suggesting that subsets of the calibration set may not be representative of the overall distribution, even if sampled in an i.i.d. manner. Additionally, identifying folds that significantly deviate from the majority could highlight local subset imbalances. Ideally, we would observe relatively evenly distributed ROC curves across folds with small variance, indicating coherence and balance where each subset behaves roughly the same. This assessment of balance could then inform the adjustment of CB widths, reducing the multiplier used in test set coverage extensions (e.g., from 2 to a value closer to 1) when the calibration set is deemed well-balanced. Theoretical efforts should focus on deriving coverage guarantees under these relaxed assumptions, while practical implementations would require defining metrics for variability (e.g., variance in AUC or pointwise TPR/FPR) and thresholds for balance that trigger width adjustments.

6.3.4 Estimating Variance-Adaptive Constants

In the pursuit of precise CBs for CDFs and ROC curves, understanding the deviation of the ECDF from the true CDF as a function of ε and sample size n is essential. Bartl and Mendelson's variance-adaptive DKW inequality offers a promising framework by scaling deviations with local variance $\sigma(t) = \sqrt{F(t)(1 - F(t))/n}$, potentially yielding tighter bounds than uniform-width approaches. However, the undetermined absolute constants c_0 and c_1 in their formulation limit direct application. Moreover, current methods like Monte Carlo CBs, while distribution-free and effective, require computationally intensive Monte Carlo simulations (B simulations for n data points) to estimate the deviations $\hat{\delta}_{1-\varepsilon}$. Establishing an empirical relationship for delta based on ε and n could enable simulation-free, efficient CB construction, significantly enhancing the practical utility of uncertainty quantification in finite-sample settings.

Future research could focus on empirically estimating the constants c_0 (related to the lower bound of the squared deviation, $\delta^2 \geq c_0 \frac{\log \log n}{n}$) and c_1 (related to the scaling of epsilon, $\delta = \sqrt{\frac{\ln(2/\varepsilon)}{c_1 n}}$) in Bartl and Mendelson's inequality through an extensive simulation setup. This would involve conducting experiments across a wide range of sample sizes n and confidence levels $1 - \varepsilon$ using both synthetic datasets. For each combination, the ECDF would be computed, the supremum deviation scaled by local variance measured, and curve-fitting or regression techniques applied to estimate c_0 and c_1. Leveraging the distribution-free property of delta in Monte Carlo CBs, as demonstrated in this thesis, ensures that results should be consistent across datasets.

6.3.5 Confidence Intervals for Probability Predictions

While this thesis has focused on constructing CBs for global performance metrics like ROC curves, the underlying Monte Carlo framework naturally suggests a pathway for uncertainty quantification at the level of individual probability predictions. Specifically, it is possible to construct confidence intervals for the true probability associated with a given classifier output p by leveraging the distribution over CDFs implied by the Monte Carlo CBs.

The key idea is to use the Monte Carlo CBs methodology to estimate the probability density over the space of CDFs. This can be achieved by gradually varying the significance level ε in the Monte Carlo CB construction, thereby mapping out the density of plausible CDFs that are consistent with the observed data. As the significance level changes, a family of CBs is obtained, and the rate at which CDFs exit or enter these bands provides an empirical estimate of the probability density over the CDF space.

Using this estimated density, one can sample plausible CDFs directly from the function space characterized by the Monte Carlo CBs. Each sample represents a possible realization of the true CDF, consistent with the observed data and the uncertainty quantified by the Monte Carlo approach.

Once a plausible CDF is sampled from this space, it can be interpolated to form a smooth, continuous function. This function then allows for mapping any classifier score p to a corresponding probability value using inverse transform sampling. Repeating this sampling and mapping process generates an entire ensemble of potential true probabilities for the score p. The distribution of this ensemble directly reflects the uncertainty in the prediction, from which a statistically sound, non-biased confidence interval can be constructed by taking its quantiles. This approach directly leverages the Monte Carlo CBs' ability to characterize the uncertainty in the CDF itself and translates it into uncertainty for individual predictions. It avoids the need to estimate PDFs via differentiation or to rely on parametric assumptions, instead using the empirical distribution over CDFs as the foundation for uncertainty quantification. This method provides a principled, simulation-based way to generate confidence intervals for classifier outputs, and represents a promising direction for future research in model calibration and uncertainty quantification.

7 Conclusion

This thesis has addressed the challenge of how to reliably quantify uncertainty in the performance evaluation of binary classifiers. The need for robust uncertainty quantification in ROC curves arises from the inherent variability in model performance across different datasets and the critical importance of reliable decision-making in high-stakes applications. Our work has yielded several successful approaches for constructing CBs around ROC curves, with the Monte Carlo CBs, demonstrating particularly strong properties in terms of conditional coverage, finite-sample guarantees, distribution-free, and variance adaptivity.

The central contribution of this work is the development of the Monte Carlo CBs, a method that leverages the theoretical properties of order statistics and the PIT to construct CBs with strong theoretical guarantees. Unlike existing methods that often rely on asymptotic assumptions or are limited by the sample size of the calibration set, our approach provides distribution-free, finite-sample guarantees that are independent of the classifier and dataset. We have shown that this method can be extended to provide coverage for future test sets, a crucial property for practical applications.

Through extensive empirical evaluation, we have demonstrated the consistency of boot-

strap CBs and compared our proposed method against it. Our results show that Monte Carlo CBs provides a more robust uncertainty quantification by accounting for the epistemic uncertainty that arises from finite sample sizes. We have also explored a Bayesian alternative, the Beta-Binomial CBs, which yielded similar results but was found to be more conservative due to its reliance on priors. Our experiments on both synthetic and real-world ECG data have confirmed the practical applicability and theoretical soundness of our approach.

This work not only provides a practical and reliable tool for uncertainty quantification in ROC analysis but also contributes to a deeper understanding of the theoretical foundations of CB construction. By highlighting the connection between ROC curves and CDF uncertainty quantification, we have opened up new avenues for research in this area. Future work could explore the development of alternative non-conformity scores to address the conservatism of the supremum-norm, investigate the impact of different dataset-level exchangeability assumptions, and explore the application of our framework to other areas of machine learning where uncertainty quantification is critical.

References

[AB22] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," 2022. [Online]. Available: https://arxiv.org/abs/2107.07511

[BH95] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: https://www.jstor.org/stable/2346101

[BK89] P. J. Bickel and A. M. Krieger, "Confidence bands for a distribution function using the bootstrap," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 95–100, 1989.

[BM23] D. Bartl and S. Mendelson, "On a variance dependent dvoretzky-kiefer-wolfowitz inequality," 2023. [Online]. Available: https://arxiv.org/abs/2308.04757

[CB12] R. Chicheportiche and J.-P. Bouchaud, "Weighted kolmogorov-smirnov test: Accounting for the tails," *Physical Review E*, vol. 86, no. 4, Oct. 2012. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.86.041115

[CM04] T. Cai and C. S. Moskowitz, "Semiparametric ROC regression analysis with placement values," *Biostatistics*, vol. 5, no. 1, pp. 45–60, 2004. [Online]. Available: https://academic.oup.com/biostatistics/article/5/1/45/316508

[Dem12] E. Demidenko, "Confidence intervals and bands for the binormal ROC curve revisited," *Journal of Applied Statistics*, vol. 39, no. 1, pp. 67–79, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3329129/

[DFV21] J. Diquigiovanni, M. Fontana, and S. Vantini, "The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data," 2021. [Online]. Available: https://arxiv.org/abs/2102.06746

[DW22] L. Duembgen and J. A. Wellner, "A new approach to tests and confidence bands for distribution functions," 2022. [Online]. Available: https://arxiv.org/abs/1402.2918

[GGL+22] S. Gustafsson, D. Gedon, E. Lampa, A. H. Ribeiro, M. J. Holzmann, T. B. Schön, and J. Sundström, "Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department

patients," *Scientific Reports*, vol. 12, no. 1, p. 19615, Nov. 2022. [Online]. Available: https://doi.org/10.1038/s41598-022-24254-x

[GOD+23] F. Ghilardi, G. Oliveira, J. V. L. Dias, M. Pereira, A. C. d. B. Silva, E. D. Gontijo, E. C. Sabino, and A. L. P. Ribeiro, "Machine learning for predicting Chagas disease infection in rural communities: Development of screening algorithms," *PLOS Neglected Tropical Diseases*, 2023. [Online]. Available: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0012026

[GPSW17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 1321–1330, 2017. [Online]. Available: https://arxiv.org/abs/1706.04599

[Hab22] T. Habineza, "Deep learning-based risk prediction of atrial fibrillation using the 12-lead ECG," Master's thesis, Uppsala University, Uppsala, Sweden, June 2022. [Online]. Available: https://github.com/mygiththh27/af-risk-prediction-by-ecg-dnn

[Hol79] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: https://www.jstor.org/stable/4615733

[LRP+21] E. M. Lima, A. H. Ribeiro, G. M. Paixão, M. H. Ribeiro, M. M. P. Filho, P. R. Gomes, D. M. Oliveira, E. C. Sabino, B. B. Duncan, L. Giatti, S. M. Barreto, W. Meira, T. B. Schön, and A. L. P. Ribeiro, "Deep neural network estimated electrocardiographic-age as a mortality predictor," *Nature Communications*, vol. 12, 2021. [Online]. Available: https://www.nature.com/articles/s41467-021-25351-7

[Mas90] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *The Annals of Probability*, vol. 18, no. 3, pp. 1269–1283, 1990. [Online]. Available: https://www.jstor.org/stable/2244426

[MPR05] S. A. Macskassy, F. Provost, and S. Rosset, "ROC confidence bands: An empirical evaluation," in *Proceedings of the 22nd International Conference on Machine Learning*, ser. ICML '05, 2005, pp. 537–544. [Online]. Available: https://icml.cc/Conferences/2005/proceedings/papers/068_ROC_MacskassyEtAl.pdf

[Run12] C. Rundel, "Lecture 15: Order statistics," Duke University Department of Statistical Science, Durham, NC, Lecture notes, Mar. 2012, open-

access PDF available from Duke Stat. [Online]. Available: https://www2.stat.duke.edu/courses/Spring12/sta104.1/Lectures/Lec15.pdf

[SP16] N. Stepanova and T. Pavlenko, "Goodness-of-fit tests based on sup-functionals of weighted empirical processes," 2016. [Online]. Available: https://arxiv.org/abs/1406.0526

[Was20] L. Wasserman, "Lecture notes 24: Bayesian inference," Lecture notes for course 36-705 Intermediate Statistics (Fall 2020), 2020, accessed on May 23, 2025. The course syllabus indicates Bayesian Inference was covered around this lecture number in Fall 2020. [Online]. Available: https://www.stat.cmu.edu/~larry/=stat705/Lecture24.pdf

[Wor24] World Health Organization, "Cardiovascular diseases (CVDs)," 2024, accessed: 2024-11-05. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[ZYS24] Z. Zheng, B. Yang, and P. Song, "Quantifying uncertainty in classification performance: ROC confidence bands using conformal prediction," *arXiv preprint*, 2024. [Online]. Available: https://arxiv.org/abs/2405.12953