Conformal Inference for Open-Set and Imbalanced Classification

Tianmin Xie[1], Yanfei Zhou[1], Ziyi Liang[2], Stefano Favaro[3], Matteo Sesia[1,4]

[1]Department of Data Sciences and Operations, University of Southern California, USA

[2]Department of Statistics, University of California, Irvine, USA

[3]Dipartimento di Scienze Economico-Sociali e Matematico-Statistiche,
Università di Torino and Collegio Carlo Alberto, Italy

[4]Department of Computer Science, University of Southern California, USA

tianminx@marshall.usc.edu, yanfeizh@marshall.usc.edu, liangz25@uci.edu,
stefano.favaro@unito.it, sesia@marshall.usc.edu

Abstract

This paper presents a conformal prediction method for classification in highly imbalanced and open-set settings, where there are many possible classes and not all may be represented in the data. Existing approaches require a finite, known label space and typically involve random sample splitting, which works well when there is a sufficient number of observations from each class. Consequently, they have two limitations: (i) they fail to provide adequate coverage when encountering new labels at test time, and (ii) they may become overly conservative when predicting previously seen labels. To obtain valid prediction sets in the presence of unseen labels, we compute and integrate into our predictions a new family of conformal p-values that can test whether a new data point belongs to a previously unseen class. We study these p-values theoretically, establishing their optimality, and uncover an intriguing connection with the classical Good–Turing estimator for the probability of observing a new species. To make more efficient use of imbalanced data, we also develop a selective sample splitting algorithm that partitions training and calibration data based on label frequency, leading to more informative predictions. Despite breaking exchangeability, this allows maintaining finite-sample guarantees through suitable re-weighting. With both simulated and real data, we demonstrate our method leads to prediction sets with valid coverage even in challenging open-set scenarios with infinite numbers of possible labels, and produces more informative predictions under extreme class imbalance.

Keywords: Classification; Conformal Prediction; Good-Turing Estimator; Imbalanced Data; Machine Learning; Species Problem.

1 Introduction

1.1 Background and Motivation

Real-world classification problems often involve not only many classes, but also the possibility that a test point may belong to a new, previously unseen category. In machine learning, this task is known as *open-set* classification or *recognition* [1]. Open-set classification is more difficult than standard classification, which operates under the *closed-set* assumption that all relevant labels are included in the observed data. A well-known example is face recognition, where the task is to identify individuals from a gallery of known identities while also being able to recognize that some test images may correspond to people never seen before

[2]. Similar challenges arise in many other domains, including voice recognition [3], biometric identification [4], medicine [5], forensic statistics [6], ecology [7], cybersecurity [8], and anomaly detection [9].

Many modern applications rely on sophisticated classification models such as support vector machines [10], random forests [11], and deep neural networks [12], including transformer-based architectures [13]. While such models are often necessary to achieve state-of-the-art predictive accuracy, they inevitably make mistakes and are generally not designed to provide rigorous quantification of statistical uncertainty, leading to sometimes overconfident and potentially unsafe usage.

To address this issue, we build on *conformal prediction* [14, 15], a flexible statistical framework for uncertainty quantification that can construct principled *prediction sets* based on the output of any model. Conformal prediction methods have found significant traction in recent years because they are able to treat the underlying model as a "black box", making them compatible with rapidly evolving machine learning algorithms, and do not rely on potentially unrealistic parametric assumptions about the data distribution, which broadens their applicability to a wide range of real-world problems. Instead, the finite-sample validity of conformal prediction guarantees derives from a fundamental assumption of data exchangeability, a condition than is even weaker than requiring independent and identically distributed samples from an unknown population. In this paper, we extend conformal prediction for classification by developing a new method and theoretical results that relax the standard closed-set assumption.

1.2 Preview of Contributions and Paper Outline

Developing a practical method for open-set conformal classification requires addressing several technical challenges. First, we analyze the behavior of existing conformal classification methods in settings where the population distribution has support over many possible labels, but the observed data may not include all of them. We show that even under exchangeability, standard conformal methods can suffer from systematic under-coverage because of their implicit, potentially misspecified assumption that all relevant labels are represented in the observed data.

Second, we introduce and study a family of hypothesis testing problems in which the null hypothesis concerns how often the label of a new data point, drawn from the population, has been observed in an exchangeable data set. A key special case is the hypothesis that the test point has a new, previously unseen label. For these problems, we construct conformal p-values, establish their finite-sample super-uniformity under the null, investigate their optimality properties, and uncover an intriguing connection to the classical Good–Turing estimator [16].

Third, we incorporate these *conformal Good–Turing p-values* into a conformal classification algorithm, enabling the construction of prediction sets with guaranteed coverage even in open-set scenarios. Intuitively, this algorithm consists of two-steps: testing and prediction. First we test whether the new data point has a previously seen label, then we output a subset of likely labels possibly (if necessary) including a place-holder "joker" symbol representing a new label.

Fourth, we further extend this methodology to improve the informativeness of our prediction sets in situations where data are limited. Specifically, we propose a non-random, label-frequency–based data splitting scheme that makes more efficient use of data with severe class imbalance, extending the simple random sample splitting approach typically used in conformal prediction, and a cross-validation–based hyper-parameter tuning strategy to optimally allocate the significance level between the hypothesis-testing and prediction components of our method.

Finally, we demonstrate the practical effectiveness of our approach on both synthetic and real datasets, showing that it delivers valid coverage in challenging open-set regimes, where existing methods fail, while often producing substantially more informative prediction sets compared to approaches based on simple random sample splitting.

1.3 Related Work

There is an extensive literature on open-set recognition (see [1] for a survey), but most prior efforts have emphasized algorithmic advances aimed at improving classification accuracy rather than providing finite-sample statistical guarantees. One line of work developed classifiers capable of "rejecting" to classify when presented with novel inputs [17], for example by leveraging extreme value theory to model tail probabilities associated with rare labels [18]. Other approaches extended deep neural networks with additional layers designed to estimate the probability that an input belongs to an unknown class [19], with subsequent extensions incorporating generative modeling techniques [20–22]. Some studies have attempted to quantify uncertainty; for example, [23] proposed a Bayesian approach, but their method is model-specific and does not offer finite-sample guarantees. By contrast, our framework is model-agnostic and provides distribution-free, finite-sample, guarantees.

Within the conformal inference literature, most prior work has focused on constructing prediction sets in the classical closed-set setting, where the label space is finite and known. Methods differ primarily in their choice of nonconformity score—for example, some aim to minimize the expected size of prediction sets [24], while others focus on maximizing conditional coverage [25]. Our framework can leverage any of these approaches while extending their applicability to the open-set regime. Although marginal coverage is the standard guarantee in conformal inference, stronger notions of validity have also been explored, such as label-conditional coverage [26]. These stronger guarantees, however, typically require many labeled samples per class. To address settings with large numbers of classes and scarce data, [27] proposed a relaxed form of label-conditional coverage by clustering similar labels and assigning rare classes to a null cluster. While their method remains restricted to the closed-set setting, it is complementary to ours and could, in principle, be combined with the open-set methodology developed in this paper.

Our work is also related to conformal classification with abstention [24, 28, 29] and to conformal out-of-distribution detection [30–32]. These approaches, however, tackle a fundamentally different problem: they typically assume the labeled data are exchangeable draws from a fixed "inlier" distribution while some test points may come from a distinct "outlier" distribution. In contrast, our setting assumes all data are exchangeable samples from the same distribution, but with potentially infinitely many possible labels. Here, a novel label is not treated as an "outlier" but simply as previously unseen. That said, there is a methodological connection: both lines of work compute feature-based conformal p-values to assess similarity between new and observed data points. We draw inspiration from existing approaches but use the resulting p-values for a different purpose: constructing prediction sets that provide finite-sample guarantees in the open-label regime. Looking ahead, these two perspectives could be combined to design open-set classification methods that not only provide distribution-free guarantees but also incorporate abstention when encountering inputs that are likely to be genuine outliers.

2 Preliminaries

2.1 Setup and Problem Statement

We consider $n+1$ paired data points $Z_i = (X_i, Y_i)$, for $i \in [n+1] := \{1, \ldots, n+1\}$, where X_i denotes a feature vector in a (possibly high-dimensional) space $\mathcal{X}$ and Y_i is a categorical label taking values in a countable dictionary $\mathcal{Y}$. Thus, each pair satisfies $Z_i \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.

We assume the sequence $Z = (Z_1, \ldots, Z_{n+1})$ is *exchangeable*, meaning that its joint distribution is invariant under arbitrary permutations. Below, we illustrate a concrete example of a possible model generating such an exchangeable sequence. Importantly, however, our methodology does not rely on any particular generative model.

Example 1 (Dirichlet Process Model)**.** A classical example of an exchangeable sequence with infinitely many possible labels is given by random sampling the Dirichlet process [33, 34]. Suppose the label space is $\mathcal{Y} = \mathbb{R}$. The Dirichlet process depends on a (non-atomic) base distribution P_0 on $\mathbb{R}$ (e.g., a standard normal) that governs how new labels are drawn, and a concentration parameter $\theta > 0$ that controls how often new labels appear. Given n previous labels $Y_1, \ldots, Y_n$, the next label Y_{n+1} is drawn from P_0 with probability $\theta/(\theta + n)$, or set equal to an existing label $y \in \{Y_1, \ldots, Y_n\}$ with probability $n_y/(\theta + n)$, where n_y is the number of times y has appeared. Conditional on the labels, features are sampled independently from some distribution $P_{X|Y}$; for example, one may take $P_{X|Y=y} = \text{Normal}(y, \sigma^2)$, for some $\sigma^2 > 0$ [35]. This construction naturally models the open-set case, since the probability of seeing a new label, $\theta/(\theta + n)$, is always positive.

Throughout the paper, we assume $Z_{1:n} = ((X_1, Y_1), \ldots, (X_n, Y_n))$ are observed and the goal is to predict the label Y_{n+1} of a new sample with features X_{n+1}. To account for uncertainty, we seek a prediction set $\hat{C}_\alpha(X_{n+1})$, implicitly depending on $Z_{1:n}$, with guaranteed *marginal coverage*:

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1})\right] \geq 1 - \alpha. \tag{1}$$

Above, the probability is taken with respect to the randomness in all variables, $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$, while $\alpha \in (0, 1)$ is a pre-defined significance level; e.g., $\alpha = 0.1$.

Marginal coverage is a standard goal in conformal inference, aiming to ensure that prediction sets contain the true label with the desired frequency on average across the population. A stronger guarantee, called *conditional coverage*, would target $\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}) \mid X_{n+1} = x\right]$ for every possible value of $x \in \mathcal{X}$, but that is theoretically unattainable without stronger distributional assumptions or infinite data [36]. Consequently, we focus on marginal coverage, while aiming for good conditional performance in practice by leveraging accurate predictive models and carefully designed nonconformity scores, as explained in more detail below.

2.2 From Closed-Set Conformal Classification to Open-Set Scenarios

Existing conformal prediction methods operate as follows, assuming the label dictionary $\mathcal{Y}$ is finite and known. For each possible label $y \in \mathcal{Y}$, they compute a *conformal p-value*, denoted as $p(y)$, that measures how well the augmented dataset $\mathcal{D}(y) := \{(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\}$ conforms to the assumption of exchangeability. These p-values are obtained by computing relative ranks of *nonconformity scores*, which are scalar statistics quantifying the degree to which an observation looks unusual relative to the rest of the data. Formally:

$$p(y) := \frac{1 + \sum_{i=1}^{n} \mathbb{I}\left[S_i^{(y)} \geq S_{n+1}^{(y)}\right]}{1 + n}, \tag{2}$$

where $S_1^{(y)}, \ldots, S_{n+1}^{(y)}$ are the non-conformity scores, computed by applying a suitable *symmetric score function* to the augmented dataset $\mathcal{D}(y)$; i.e., $S_i^{(y)} = s((X_i, Y_i); \mathcal{D}(y))$ for $i \in [n]$ and $S_{n+1}^{(y)} = s((X_{n+1}, y); \mathcal{D}(y))$. Here, s is said to be "symmetric" because it does not look at the order of the data in $\mathcal{D}(y)$. This is designed to ensure the p-values are marginally super-uniform when evaluated at $y = Y_{n+1}$; that is, $\mathbb{P}\left[p(Y_{n+1}) \leq u\right] \leq u$ for any $u \in (0, 1)$, as explained in more detail below. In turn, this implies that marginal coverage (1) is achieved by the prediction set comprising all labels $y \in \mathcal{Y}$ whose conformal p-value exceeds the target level α:

$$\hat{C}_\alpha(X_{n+1}; \mathcal{Y}) := \{y \in \mathcal{Y} : p(y) > \alpha\}. \tag{3}$$

There are two key methodological choices involving the score function that are worth recalling here. The first choice concerns how the output of a black-box classification model is mapped into scalar nonconformity scores. A classical and simple option is to use the negative estimated conditional probability of y given x;

i.e., $S_i^{(y)} = \hat{\mathbb{P}}[Y = Y_i \mid X = X_i]$ for $i \in [n]$ and $S_{n+1}^{(y)} = \hat{\mathbb{P}}[Y = y \mid X = X_{n+1}]$, where $\hat{\mathbb{P}}$ denotes an estimate of the conditional probability; e.g., as provided by the final soft-max layer of a deep neural network. This choice of score function is designed to approximately minimize the expected size of the prediction sets [26]. An alternative, designed to improve conditional coverage, is to apply rank-based transformations to the estimated class probabilities, yielding generalized inverse quantile scores [37]. If the classification model can accurately estimate the true conditional class probabilities in the population, then inverse quantile scores lead to prediction sets with not only marginal but also conditional coverage. We refer to Appendix A1 for further details on this score function.

The second choice concerns which data points are used for training the model and which for computing nonconformity scores. In the *full conformal* approach, the augmented dataset $\mathcal{D}(y)$ is used for both tasks. This means the model must be re-trained for each candidate label, which can be computationally demanding. In the more practical *split conformal* approach, the model is pre-trained once using an independent training data set, so that the scores $s((X_i, Y_i); \mathcal{D}(y))$ and $s((X_i, Y_i); \mathcal{D}(y))$ do not actually depend on the second argument $\mathcal{D}(y)$. For completeness, a detailed implementation outline of split-conformal classification with generalized inverse quantile scores in the closed-set setting is provided by Algorithm A1 in Appendix A1.

For any of these existing conformal prediction methods, the data exchangeability translates into exchangeability of the nonconformity scores, resulting in marginally valid prediction sets satisfying (1). Moreover, as long as the scores are almost-surely distinct, one can also prove that conformal prediction sets are not overly conservative, in the sense that they satisfy an almost-matching coverage upper bound: $\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n)\right] \leq 1 - \alpha + \min\{1/(n+1), \alpha\}$. The methods and theory developed in this paper apply to any of these frameworks, but in our empirical demonstrations we will focus on split conformal prediction with generalized inverse quantile scores, which offers computational efficiency while often producing more informative sets with relatively good conditional coverage.

The first question we consider is how existing conformal prediction methods (like Algorithm A1), originally developed for the closed-set setting, behave when applied in an open-set scenario. In particular, we ask: what coverage can be expected if the unknown full label space $\mathcal{Y}$ is replaced by the subset of observed labels, $\mathcal{Y}_n := \text{unique}(Y_1, \ldots, Y_n)$? Concretely, Algorithm A2 in Appendix A1 describes the corresponding plug-in implementation of Algorithm A1, using $\mathcal{Y}_n$ instead of $\mathcal{Y}$. Intuitively, the performance of this approach must depend on the probability that the next label has not been seen before—that is, on the probability of the event

$$N_{n+1} := \{ Y_{n+1} \notin \mathcal{Y}_n \}. \tag{4}$$

Theorem 1. *Assume $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable and the scores $\{S_1^{(y)}, \ldots, S_{n+1}^{(y)}\}$ are almost-surely distinct for any $y \in \mathcal{Y}$, with $\mathcal{Y}$ being unknown and potentially infinite. Then,*

$$1 - \alpha - \mathbb{P}[N_{n+1}] \leq \mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n)\right] \leq 1 - \alpha + \min\left\{\frac{1}{n+1}, \alpha - \mathbb{P}[N_{n+1}]\right\}.$$

Moreover, $\mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \in \mathcal{Y}_n\right] \leq \alpha$.

In the special case where $\mathbb{P}[N_{n+1}] = 0$, Theorem 1 reduces to the familiar result $1 - \alpha \leq \mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n)\right] \leq 1 - \alpha + \min\{1/(n+1), \alpha\}$, telling us that standard conformal prediction has valid coverage (slightly) above $1 - \alpha$ in the closed-set setting. In general, however, this shows that standard conformal prediction sets may significantly under-cover if $\mathbb{P}[N_{n+1}] > 0$, especially if new labels are relatively common in the sense that $\mathbb{P}[N_{n+1}] > \alpha$.

As a concrete illustration, under the same Dirichlet process model previously introduced in Example 1, we can compute exactly $\mathbb{P}[N_{n+1}] = \theta/(\theta + n)$. This probability decreases as n grows but remains strictly positive for any finite n. By Theorem 1, standard conformal prediction coverage is upper bounded by $1 - \theta/(\theta + n)$, which can be substantially below the target $1 - \alpha$ when θ is large or n is small. For instance, with $\theta = 100$, $n = 100$, and $\alpha = 0.1$ the coverage upper bound becomes 0.5, far below the nominal 0.9 level.

2.3 Preview of Proposed Method

To address the problem that standard conformal methods may under-cover in the open-set setting, we must allow prediction sets to include labels beyond those observed in $\mathcal{Y}_n$. Since the remainder of the label space $\mathcal{Y} \setminus \mathcal{Y}_n$ is entirely obscure, it is necessary to introduce a special "joker" symbol, denoted as $\circledast$, which acts as a placeholder for all unseen labels. The joker provides a principled and clear way to signal that X_{n+1} may be associated with a novel label Y_{n+1}, even if its identity cannot be specified. This is a natural approach in purely categorical problems, where labels carry no intrinsic meaning beyond whether they are the same or different, and leads to informative prediction sets as long as jokers are included parsimoniously.

The central challenge is to determine when to include the joker in a prediction set, or equivalently, how to test whether Y_{n+1} is novel. We cast this as a special case of a more general hypothesis testing problem: for each $k \geq 0$, we can ask whether the test label has appeared exactly k times in the data, with $k = 0$ indicating a new label. In Section 3, we formalize this problem, develop tests with finite-sample type-I error control, and analyze their optimality. We also draw a connection to the classical work of Good and Turing on estimating the probability of unseen species [16], which motivates the terminology *conformal Good–Turing p-values* for our proposed tests. Then, in Section 4, we show how to leverage these p-values within a conformal prediction algorithm that yields open-set prediction sets with valid marginal coverage.

Section 4 also makes two additional methodological contributions. First, because our procedure integrates two distinct components (hypothesis testing and prediction), we propose an effective data-driven strategy for allocating the miscoverage budget α across the two parts as efficiently as possible. Second, in the split conformal setting, we develop a non-random, frequency-based train–calibration sample split with proper reweighting. This can sometimes greatly improve predictive efficiency while preserving coverage, and is particularly advantageous in the presence of strong class imbalance, where standard random sample splitting would be sub-optimal.

3 Testing Hypotheses about Label Frequencies

3.1 Setup

We introduce a family of statistical hypotheses about the empirical frequency of the label Y_{n+1} within the first n observations. Let $Y_{1:n} = (Y_1, \ldots, Y_n)$ denote the sequence of observed labels, define the *empirical frequency* of a label $y \in \mathcal{Y}$ as the number of times it occurs in $Y_{1:n}$:

$$N(y; Y_{1:n}) = \sum_{i=1}^{n} \mathbf{1}\{Y_i = y\}.$$

For each non-negative integer $k \in \{0\} \cup \mathbb{N}$, we define the hypothesis H_k as:

$$H_k : N(Y_{n+1}; Y_{1:n}) = k. \tag{5}$$

Note that this is a *random* hypothesis because its truth value depends on the outcome of random variables $Y_1, \ldots, Y_{n+1}$. A special case of particular interest is H_0, which is true if and only if Y_{n+1} is a new label. We will now study how to construct a powerful p-value ψ_k for testing H_k, using the data in $(X_1, Y_1), \ldots, (X_n, Y_n)$ and possibly also X_{n+1}. Concretely, we seek a ψ_k satisfying

$$\mathbb{P}\left[\psi_k \leq u, H_k\right] \leq u, \qquad \forall u \in (0, 1), \tag{6}$$

which is a sufficient notion of super-uniformity for our purposes.

3.2 Feature-Blind Conformal Good–Turing p-Values

We begin by developing p-values for testing H_k using only the observed labels $\{Y_1, \ldots, Y_n\}$, postponing to the next section the incorporation of feature information. Throughout, we treat the label space $\mathcal{Y}$ as purely categorical and unstructured: the labels have no intrinsic meaning beyond equality or inequality. It is therefore natural to focus on statistics of $\{Y_1, \ldots, Y_n\}$ that depend only on the sample *frequency profile*: $\mathcal{F}_n = \{M_0, M_1, \ldots, M_n\}$, where M_k is the number of *distinct* labels that appear exactly k times in $\{Y_1, \ldots, Y_n\}$, so that $\sum_{k=0}^{n} k M_k = n$.

For each $k \in \{0, 1, \ldots, n\}$, we define the feature-blind *Good–Turing* conformal p-value ψ_k^{GT}:

$$\psi_k^{\mathrm{GT}} = \frac{(k+1)M_{k+1} + k + 1}{n+1}. \tag{7}$$

We can prove ψ_k^{GT} is a valid conformal p-value for testing H_k, and is optimal among all possible conformal p-values that are (deterministic) functions of the sample frequency profile.

Theorem 2. *Assume $\{(Y_i)\}_{i=1}^{n+1}$ are exchangeable. For any $k \in \{0, 1, \ldots, n\}$ and $u \in (0, 1)$,*

$$\mathbb{P}\left[\psi_k^{GT} \leq u, H_k\right] \leq u. \tag{8}$$

Moreover, if ψ'_k is any deterministic function of $\mathcal{F}_n$ satisfying $\mathbb{P}\left[\psi'_k \leq u, H_k\right] \leq u$ for all $u \in (0, 1)$, then $\psi'_k \geq \psi_k^{GT}$ almost surely.

The term "Good–Turing" emphasizes the connection between (7) and the classical Good–Turing estimator for the probability of unseen species. [16] proposed estimating the total probability mass of all species observed exactly k times in a sample of size n as $(k+1)M_{k+1}/n$, where M_{k+1} is the number of species appearing $k+1$ times. In particular, for $k = 0$, the Good–Turing estimator of the probability of observing a new species is M_1/n, the fraction of observations corresponding to species seen only once. The conformal p-value in (7), though derived from different principles (hypothesis testing via conformal inference rather than estimation), mirrors this form and is asymptotically equivalent as $n \to \infty$ with k fixed. The role of the additional $(k+1)/(n+1)$ term in (7) is to ensure the finite-sample conservativeness of conformal p-values, in contrast to the Good–Turing estimator's focus on unbiasedness. Moreover, under the null hypothesis H_k, the p-value in (7) recovers exactly the classical Good–Turing estimator evaluated on the augmented data set $Y_1, \ldots, Y_n, Y_{n+1}$.

The power of the conformal p-values in (7) can be further increased through randomization. For each $k \in \{0, 1, \ldots, n\}$, we define the randomized feature-blind *Good–Turing* p-value ψ_k^{GT} as:

$$\psi_k^{\mathrm{RGT}} = \frac{\mathrm{Uniform}\left(\{0, 1, \ldots, (k+1)M_{k+1} + k\}\right) + 1}{n+1}, \tag{9}$$

where $\mathrm{Uniform}(A)$ denotes a uniform random variable taking on a discrete set $A \subset \mathbb{N}$, independent of everything else. It is clear that ψ_k^{RGT} is more powerful than ψ_k^{GT} because $\mathbb{P}\left[\psi_k^{\mathrm{RGT}} < \psi_k^{\mathrm{GT}} \mid \mathcal{F}_n\right] = 1 - 1/((k+1)M_{k+1} + k + 1)$. Moreover, it is still a valid conformal p-value.

Proposition 1. *Assume $\{(Y_i)\}_{i=1}^{n+1}$ are exchangeable. Then, $\mathbb{P}\left[\psi_k^{RGT} \leq u, H_k\right] \leq u$ for all $u \in (0, 1)$ and any $k \in \{0, 1, \ldots, n\}$.*

3.3 Feature-Based Conformal Good–Turing p-Values

We now extend the conformal Good–Turing p-values to incorporate potentially valuable feature information contained in $\{X_1, \ldots, X_n, X_{n+1}\}$. To that end, we introduce new notation.

For each $k \in \{0, 1, \ldots, n\}$, let $\mathcal{S}_k$ be the set of indices of points whose label appears exactly k times among $\{Y_1, \ldots, Y_n\}$:

$$\mathcal{S}_k := \Big\{ i \in [n] : N(Y_i; Y_{1:n}) = k \Big\}.$$

Next, let $\hat{s}_k : \mathcal{X} \to \mathbb{R}_+$ be an arbitrary score function, treated as fixed for now (we discuss data-driven choices at the end of this section). The function $\hat{s}_k$ measures how atypical a feature vector x is for a label with frequency k: larger values of $\hat{s}_k(X_{n+1})$ indicate that X_{n+1} is less consistent with Y_{n+1} having frequency k, thus providing stronger evidence against H_k.

For any $k \in \{0, 1, \ldots, n\}$, we define the feature-based conformal Good–Turing p-value as:

$$\psi_k^{\mathrm{XGT}} := \frac{1}{n+1} \left(1 + \sum_{i \in \mathcal{S}_{k+1}} \mathbb{I}\left[\hat{s}_k(X_i) \geq \hat{s}_k(X_{n+1}) \right] + \max_{i \in \mathcal{S}_k} \sum_{j \in [n] : Y_j = Y_i} \mathbb{I}\left[\hat{s}_k(X_j) \geq \hat{s}_k(X_{n+1}) \right] \right). \tag{10}$$

Before formally stating the super-uniformity of (10) under H_k, we discuss an interesting special case to gain more intuition. If $k = 0$, the expression in (10) simplifies to:

$$\psi_0^{\mathrm{XGT}} = \frac{1 + \sum_{i \in \mathcal{S}_1} \mathbb{I}\left[\hat{s}_0(X_i) \geq \hat{s}_0(X_{n+1}) \right]}{n+1} \in \left\{ \frac{1}{n+1}, \ldots, \frac{M_1 + 1}{n+1} \right\}.$$

Let us now consider two extreme situations. On the one hand, suppose $\hat{s}_0(X_{n+1})$ is greater than $\hat{s}_0(X_i)$ for all $i \in \mathcal{S}_1$. Intuitively, this means test point appears *more unusual* as a rare species, based on its features, than all observed data points with frequency one. This suggests the features X_{n+1} carry strong evidence that Y_{n+1} is not a new label, and indeed in this case the conformal p-value ψ_0^{XGT} attains its smallest possible value at $1/(n+1)$. On the other hand, suppose $\hat{s}_0(X_{n+1})$ is smaller than $\hat{s}_0(X_i)$ for all $i \in \mathcal{S}_1$. Intuitively, this means test point appears *less unusual* as a rare species, based on its features, than all observed data points with frequency one. This suggests the features X_{n+1} do not carry any evidence that Y_{n+1} is not a new label, and indeed in this case the conformal p-value ψ_0^{XGT} attains its largest possible value at $(1 + M_1)/(n+1)$, which coincides with the deterministic feature-blind conformal Good–Turing p-value introduced earlier in (7).

Overall, (10) behaves similarly to the randomized p-value in (9), but ties are now broken based on the observed feature-based scores (which are likely to be informative) rather than on independent random noise. For $k > 0$, the expression in (10) becomes a little more complicated, as it requires comparing the score $\hat{s}_k(X_{n+1})$ not only to the scores $\hat{s}_k(X_i)$ for $i \in \mathcal{S}_{k+1}$ but also to the scores $\hat{s}_k(X_i)$ for $i \in [n]$ such that $Y_i = y$, for all $y \in \mathcal{S}_k$. A similar intuition, however, still applies. Therefore, feature-based conformal Good–Turing p-values are typically more useful than (randomized) feature-blind p-values in practice, as we will demonstrate empirically later.

We are now ready to state the validity of ψ_k^{XGT}, defined in (10), as a conformal p-value for H_k.

Theorem 3. *Assume $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, and fix any $k \in \{0, 1, \ldots, n\}$. Suppose also the score function $\hat{s}_k$ in (10) is either fixed or invariant to permutations of $\{(X_i, Y_i)\}_{i=1}^{n+1}$ under H_k. Then, $\mathbb{P}\left[\psi_k^{XGT} \leq u, H_k\right] \leq u$ for all $u \in (0, 1)$.*

Theorem 3 allows the score function $\hat{s}_k$ to be data-dependent, provided it is invariant to permutations of $\{(X_i, Y_i)\}_{i=1}^{n+1}$ under H_k. We now describe a practical, though not unique, way to construct such a function. Although the label Y_{n+1} is latent, we know that under H_k it must be one of those observed exactly k times among $\{Y_1, \ldots, Y_n\}$. Therefore, the value of Y_{n+1} is irrelevant if $\hat{s}_k$ is fitted using only data with label frequency different from both k and $k+1$.

Concretely, a natural strategy is to train a *one-class classifier* [38, 39] on the features of the observed data *excluding* all samples with label frequency k or $k+1$. For example, for $k = 0$, we would fit the scoring function on the features from samples with labels appearing at least twice in the available data. In general, making its

dependence on the training data explicit, this score function can be written as $\hat{s}_k(\cdot\,; \{X_i\}_{i\in[n]\setminus(\mathcal{S}_k\cup\mathcal{S}_{k+1})})$. It is easy to verify the data subset $\{X_i\}_{i\in[n]\setminus(\mathcal{S}_k\cup\mathcal{S}_{k+1})}$ is invariant to permutation of $\{(X_i, Y_i)\}_{i=1}^{n+1}$ under H_k, which leads to an invariant score function, as required by Theorem 3, as long as the learning algorithm is insensitive to the order of the input data points (if that is not the case, the training data can also simply be randomly shuffled). Moreover, this approach is computationally efficient because X_{n+1} is always excluded from this training set, ensuring the same scoring function can be re-used for different test points.

In this paper, we compute feature-based conformal Good–Turing p-values using a scoring function obtained from a local outlier factor one-class classification model, as implemented by the `Scikit-learn` Python package [40]. By convention, this model assigns larger scores to features conforming to the training distribution, and therefore larger values of $\hat{s}_k(X_{n+1}\,; \{X_i\}_{i\in[n]\setminus(\mathcal{S}_k\cup\mathcal{S}_{k+1})})$ indicate Y_{n+1} is less likely to have frequency k or $k+1$, thereby providing evidence against H_k. While alternative constructions are certainly possible, this approach is intuitive, computationally straightforward, and—as we show in our experiments—often yields substantially higher power than feature-blind p-values. The underlying assumption is, of course, that the features are informative as to the true label frequencies. Otherwise, ψ_k^{XGT} would not bring any advantages compared to ψ_k^{RGT}.

3.4 Testing the Composite Hypothesis that Y_{n+1} is Previously Seen Label

When developing our conformal prediction method in Section 4, it will be useful not only to test H_0 but also to test the complementary null hypothesis that Y_{n+1} is *not* a new label. The latter is a composite null hypothesis, which can be written as:

$$H_{\mathrm{seen}} : \bigcup_{k\in K_n} H_k, \tag{11}$$

where $K_n := \{k \in [n] : M_k > 0\}$ is the set of observed label frequencies.

We propose to test H_{seen} using a *combination p-value* of the form:

$$\psi_{\mathrm{seen}} = \max_{k\in K_n} \frac{\psi_k}{c_k}, \tag{12}$$

where each ψ_k is a valid p-value for H_k, as in the previous section, and the c_k's are suitable constants. We now provide sufficient conditions under which ψ_{seen} is super-uniform under H_{seen}.

Theorem 4. *Assume $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, $\mathbb{P}[\psi_k \le u, H_k] \le u$ for all $u \in (0,1)$ and $k \in [n]$. Suppose also $c_k \ge 0$ and $\sum_{k=1}^{n} c_k \le 1$. Then, with ψ_{seen} defined as in* (12)*, we have $\mathbb{P}[\psi_{seen} \le u, H_{seen}] \le u$ for all $u \in (0,1)$.*

We now discuss the choice of the c_k constants. The *empirically* optimal weights, minimizing ψ_{seen} conditional on $\{\psi_k\}_{k\in K_n}$, are obtained by solving

$$\min_{\{c_k\ge 0\}} \max_{k\in K_n} \left\{\frac{\psi_k}{c_k}\right\} \quad \text{subject to} \quad \sum_{k\in K_n} c_k \le 1.$$

This solution sets $c_k \propto \psi_k$ for $k \in K_n$. However, such weights are infeasible in practice since they depend on the data, while Theorem 4 requires fixed constants.

Still, the ideal target $c_k \propto \psi_k$ suggests a practical approximation. From (7), we know that $\psi_k \propto M_{k+1}$, and it is well documented that in many real-world categorical data sets M_{k+1} exhibits approximate power-law tail behavior [41, 42]. This motivates choosing a fixed value of c_k according to a power law as well:

$$c_k = \frac{k^{-\beta}}{\sum_{j=1}^{n} j^{-\beta}}, \quad k \in [n], \tag{13}$$

where $\beta > 0$ is a fixed scaling parameter. Based on extensive numerical experiments, we recommend $\beta = 1.6$ as a robust default, though β may be tuned (e.g., via cross-validation) for specific applications. All experiments in this paper use $\beta = 1.6$ unless otherwise stated.

4 Open-Set Conformal Classification

4.1 Conformal Good–Turing Classification

We can now present our method for open-set conformal classification, which we call *conformal Good–Turing classification*. Let ψ_{unseen} denote a conformal Good–Turing p-value for testing the null hypothesis $H_{\text{unseen}} := H_0$, corresponding to $k = 0$ in (11). Let ψ_{seen} denote a conformal Good–Turing p-value for testing the null hypothesis H_{seen}, defined in (11). Fix any $\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}} \in [0, 1]$ such that $\alpha_{\text{class}} + \alpha_{\text{unseen}} + \alpha_{\text{seen}} = \alpha$, where $1 - \alpha$ is the desired coverage level. We postpone to later the discussion of how to best allocate a given α budget to $\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}}$.

As anticipated earlier, our prediction sets will sometimes include a "joker" symbol $\circledast$, whose interpretation is that of a "catch-all" for every label not observed among the first n data points; i.e., $\{\circledast\} := \mathcal{Y} \setminus \mathcal{Y}_n$. Intuitively, the prediction set for Y_{n+1} should include the joker unless there is sufficient evidence that Y_{n+1} is not a new label. Conversely, it is safe for the prediction set to include only the joker, and no previously seen labels, if there is sufficient evidence that Y_{n+1} is a new label. This leads to the following open-set prediction set:

$$\hat{C}^*_\alpha(X_{n+1}) := \begin{cases} \hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n), & \text{if } \psi_{\text{unseen}} \le \alpha_{\text{unseen}} \text{ and } \psi_{\text{seen}} > \alpha_{\text{seen}}, \\ \{\circledast\}, & \text{if } \psi_{\text{unseen}} > \alpha_{\text{unseen}} \text{ and } \psi_{\text{seen}} \le \alpha_{\text{seen}}, \\ \hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n) \cup \{\circledast\}, & \text{otherwise.} \end{cases} \tag{14}$$

Above, $\hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n)$ denotes a standard conformal prediction set constructed under the closed-set setting, assuming $\mathcal{Y} = \mathcal{Y}_n$, at level α_{class}. Any existing conformal classification method can be applied for this purpose, as reviewed in Section 2. In particular, this includes both split- and full-conformal prediction approaches, based on any choice of non-conformity scores.

The following result establishes that $\hat{C}^*_\alpha(X_{n+1})$ has marginal coverage at level $1 - \alpha$ as long as $\hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n)$ is valid under the closed-set setting and $\psi_{\text{unseen}}, \psi_{\text{seen}}$ are super-uniform under their respective null hypotheses.

Theorem 5. *Assume* $\{(X_i, Y_i)\}_{i=1}^{n+1}$ *are exchangeable,* $\mathbb{P}[\psi_{unseen} \le \alpha_{unseen}, H_{unseen}] \le \alpha_{unseen}$ *and* $\mathbb{P}[\psi_{seen} \le \alpha_{seen}, H_{seen}] \le \alpha_{seen}$. *Assume also* $\hat{C}_{\alpha_{class}}(X_{n+1}; \mathcal{Y})$ *is a valid* α_{class}*-level prediction set under the closed-set setting, for any label dictionary* $\mathcal{Y}$*; i.e., it suffices that* $\mathbb{P}\left[Y_{n+1} \notin \hat{C}_{\alpha_{class}}(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \in \mathcal{Y}_n\right] \le \alpha_{class}$. *Then,* $\mathbb{P}\left[Y_{n+1} \in \hat{C}^*_\alpha(X_{n+1})\right] \ge 1 - \alpha$, *where* $\alpha = \alpha_{class} + \alpha_{unseen} + \alpha_{seen}$.

Moreover, the next result demonstrates that our method is a natural generalization of existing approaches. In particular, if the label dictionary $\mathcal{Y}$ has finite cardinality $K = |\mathcal{Y}|$, there exists an allocation of the α-budget (depending on K) that allows our method to recover the standard conformal prediction sets, up to a small conservative α-adjustment of order $O(K/n)$. This asymptotically vanishing extra conservativeness is the price to pay for having guaranteed coverage at level $1 - \alpha$ despite not knowing in advance the identity of all possible labels.

Proposition 2. *Assume* $\{(X_i, Y_i)\}_{i=1}^{n+1}$ *are exchangeable and the label space satisfies* $|\mathcal{Y}| = K$ *for some finite* K, *with* $(K + 1)/(n + 1) < \alpha$. *If we set* $(\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}}) = (\alpha - (K + 1)/(n + 1), (K + 1)/(n + 1), 0)$, *the prediction set in* (14) *obtained using any of the p-values* ψ_0^{GT}, ψ_0^{RGT}, *or* ψ_0^{XGT} *in* (7), (9), *or* (10) *respectively, satisfies* $\hat{C}^*_\alpha(X_{n+1}) = \hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n)$ *almost surely.*

4.2 Data-Driven Allocation of Significance Budget

While Theorem 5 guarantees that (14) achieves marginal coverage at level $1-\alpha$ for any allocation $(\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}})$ with $\alpha = \alpha_{\text{class}} + \alpha_{\text{unseen}} + \alpha_{\text{seen}}$, the usefulness of the resulting prediction sets depends on how this α budget is distributed. Increasing α_{class} produces smaller sets of seen labels, increasing α_{unseen} reduces reliance on the $\circledast$ symbol, and increasing α_{seen} allows the method to reject all seen labels in cases where new labels are very common. These trade-offs depend on the problem setting: for example, in a closed-set regime where $\Pr(N_{n+1}) = 0$, the optimal allocation is $\alpha_{\text{class}} = \alpha$, while in a fully open-set regime with $\Pr(N_{n+1}) = 1$, the optimal allocation is $\alpha_{\text{seen}} = \alpha$. To adapt across these and all possible intermediate cases, we therefore develop a practical data-driven strategy for tuning $(\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}})$.

It is worth emphasizing the data-driven approach presented below is heuristic, in the sense it may theoretically invalidate the coverage guarantee provided by Theorem 5, which assumes $(\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}})$ are fixed. Nonetheless, it is useful and will be validated empirically in Section 5, where we show the coverage of our method is robust this tuning in practice.

Our tuning strategy is intended to (approximately) maximize the informativeness of the prediction sets, which in this setting depends on two factors: their size and how often they include the joker symbol. Formally, we define the size of a prediction set as the number of seen labels it contains, plus one if it also includes the joker:

$$|\hat{C}^*_\alpha(X_{n+1})| := \begin{cases} |\hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n)|, & \text{if } \psi_{\text{unseen}} \le \alpha_{\text{unseen}} \text{ and } \psi_{\text{seen}} > \alpha_{\text{seen}}, \\ 1, & \text{if } \psi_{\text{unseen}} > \alpha_{\text{unseen}} \text{ and } \psi_{\text{seen}} \le \alpha_{\text{seen}}, \\ |\hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n)| + 1, & \text{otherwise.} \end{cases} \tag{15}$$

Minimizing the prediction set size alone is not entirely satisfactory, however, because a set containing $\circledast$ intuitively seems less informative than one with an extra seen label but no joker. To capture this trade-off between seen labels and the joker, we define a composite loss function designed to penalize both larger prediction sets and, separately, sets including the joker symbol:

$$L\big(\hat{C}^*_\alpha(X_{n+1}), Y_{n+1}; \lambda\big) = \lambda \frac{\big|\hat{C}^*_\alpha(X_{n+1})\big|}{\big|\hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n)\big|} \; + \; (1-\lambda)\, \mathbb{I}\big\{\circledast \in \hat{C}^*_\alpha(X_{n+1}),\, Y_{n+1} \in \mathcal{Y}_n\big\}. \tag{16}$$

Above, $\lambda \in [0,1]$ is a user-specified parameter balancing the desire for small prediction sets (first term) against the cost of wasting a joker when the true label is seen (second term).

The optimal allocation of $(\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}})$ is then defined as

$$\arg\min_{\substack{\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}} \ge 0 \\ \alpha_{\text{class}} + \alpha_{\text{unseen}} + \alpha_{\text{seen}} = \alpha}} \; \mathbb{E}\Big[L\big(\hat{C}^*_\alpha(X_{n+1}), Y_{n+1}; \lambda\big) \Big]$$

In practice, we approximate this expectation empirically using a cross-validation approach: the data are split into K folds, and for each candidate allocation we compute the empirical loss on the held-out fold while applying our method on the data from the remaining folds. This procedure is summarized by Algorithm A3 in Appendix A2.1.

4.3 Frequency-Based Selective Sample Splitting

So far we have not specified how the closed-set conformal predictor $\hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n)$ is constructed, and our results apply equally to split and full conformal methods. Since full conformal prediction is often computationally prohibitive, we expect that in practice our approach will typically rely on a split-conformal scheme. This raises the question of how best to divide the data into training and calibration sets when the label

space is very large. The standard practice in conformal prediction is to split samples uniformly at random, but this can be highly suboptimal in open-set settings, where class distributions are extremely imbalanced.

The core issue is that single observations from rare classes are far more valuable than single observations from common ones. If such rare examples are placed in the calibration set rather than training, the model cannot learn how to classify them. Therefore, rare observations should be prioritized for training rather than calibration. Such a "selective" splitting strategy however has two costs: (1) it breaks exchangeability between calibration and test samples, though this can be corrected through appropriate reweighting [43]; and (2) it may reduce conditional coverage for rare classes [44], although marginal validity is still guaranteed. We will show empirically that these trade-offs are worthwhile in practice.

It is important to clarify that, in this section, we assume $\hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n)$ is built following the standard procedure reviewed in Section 2.2, but evaluating the nonconformity scores only on a calibration subset $\mathcal{D}^{\text{cal}} \subset [n]$, while the remaining data in $\mathcal{D}^{\text{train}} = [n] \setminus \mathcal{D}^{\text{cal}}$ are used to fit the model defining the score function. Crucially, sample-splitting will apply only to the construction of $\hat{C}_{\alpha_{\text{class}}}$; all other components of our method continue to utilize the full data sample. In particular, the observed label subspace is still defined as $\mathcal{Y}_n := \text{unique}(Y_1, \ldots, Y_n)$, and the conformal p-values ψ_{seen} and ψ_{unseen} in (14) are likewise computed from all available samples, as in Section 3.

The selective sample-splitting scheme we propose is as follows. Recall that, for any $y \in \mathcal{Y}$, $N(y; Y_{1:n})$ counts how many times y appears in $\{Y_1, \ldots, Y_n\}$. For each $y \in \mathcal{Y}_n$ and $i \in [n] : Y_i = y$, we randomly sample an independent binary variable $I_i \in \{0, 1\}$ conditional on the label sequence $Y_{1:n}$ from a Bernoulli distribution whose success probability depends on $N(Y_i; Y_{1:n})$:

$$I_i \overset{\text{ind.}}{\sim} \text{Bernoulli}\big(\pi(N(Y_i; Y_{1:n}))\big),$$

where $\pi : \mathbb{N} \to [0, 1]$ is a monotone increasing function, whose choice is discussed at the end of this section. Then, we assign the samples with $I_i = 1$ to the calibration set:

$$\mathcal{D}^{\text{cal}} = \{i \in [n] : I_i = 1\}, \qquad \mathcal{D}^{\text{train}} = [n] \setminus \mathcal{D}^{\text{cal}}.$$

In general, selective sample splitting breaks the exchangeability between the data indexed by $\mathcal{D}^{\text{cal}}$ and the test point (X_{n+1}, Y_{n+1}), which could potentially invalidate the assumption $\mathbb{P}\left[Y_{n+1} \notin \hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \in \mathcal{Y}_n\right] \leq \alpha_{\text{class}}$ required in Theorem 5 to ensure the validity of our method. Fortunately, the closed-set validity of $\hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n)$ can be recovered by applying *weighted* conformal prediction techniques [43] to account for the distribution shift when computing the conformal p-values $p(y)$ in (2), as described next.

For each candidate test label $y \in \mathcal{Y}_n$, consider swapping the test point (with assumed label $Y_{n+1} = y$) with a calibration point $j \in \mathcal{D}^{\text{cal}}$ (with label Y_j). Define the corresponding swapped label sequences $Y_{1:n}^{(j,y)} = (Y_1^{(j,y)}, \ldots, Y_n^{(j,y)})$ and $Y_{1:(n+1)}^{(j,y)} = (Y_1^{(j,y)}, \ldots, Y_{n+1}^{(j,y)})$, where:

$$Y_i^{(j,y)} = \begin{cases} y & \text{if } i = j, \\ Y_j & \text{if } i = n+1, \\ Y_i & \text{otherwise.} \end{cases}$$

After such swapping, the empirical frequency of each label $y' \in \mathcal{Y}_n$ becomes:

$$N(y'; Y_{1:n}^{(j,y)}) = \begin{cases} N(y'; Y_{1:n}) + 1 & \text{if } y' = y \text{ and } y \neq Y_j, \\ N(y'; Y_{1:n}) - 1 & \text{if } y' = Y_j \text{ and } y \neq Y_j, \\ N(y'; Y_{1:n}) & \text{otherwise.} \end{cases}$$

With this notation, for all $j \in \mathcal{D}^{\text{cal}} \cup \{n+1\}$, we define the *conformalization weights*

$$w_j(y) = \frac{p^{(j)}(y)}{p^{(n+1)}(y) + \sum_{k \in \mathcal{D}^{\text{cal}}} p^{(k)}(y)}, \tag{17}$$

where

$$p^{(j)}(y) := \pi\big(N(y; Y_{1:n}^{(j,y)})\big) \prod_{i \in \mathcal{D}^{\text{cal}}, i \neq j} \pi\big(N(Y_i^{(j,y)}; Y_{1:n}^{(j,y)})\big) \prod_{i \in \mathcal{D}^{\text{train}}} \big(1 - \pi(N(Y_i^{(j,y)}; Y_{1:n}^{(j,y)}))\big).$$

At first glance, computing each of these weights directly appears to require evaluating a product over all n samples; consequently, computing all weights naively incurs a total cost of $O(n^2)$ per test point, which is computationally expensive. However, with a more careful implementation, this cost can be reduced from $O(n^2)$ to $O(n)$. Additional implementation details, including a fast computational shortcut for computing these weights in practice, are provided in Appendix A2.2.

Finally, the weighted version of the closed-set prediction set $\hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n)$ is given by

$$\hat{C}_\alpha^\pi(X_{n+1}; \mathcal{Y}_n) = \left\{ y \in \mathcal{Y}_n : w_{n+1}(y) + \sum_{j \in \mathcal{D}^{\text{cal}}} w_j(y) \mathbb{I}\left[S_j \geq S_{n+1}\right] > \alpha \right\}, \tag{18}$$

where S_j is the non-conformity score for calibration point j and S_{n+1} is the score for the test point assuming $Y_{n+1} = y$. In the split conformal approach, the model is pre-trained once using an independent training data set, so that the scores $s((X_i, Y_i); \mathcal{D}(y))$ do not depend on $\mathcal{D}(y)$. Consequently, we can write S_j and S_{n+1} instead of $S_j^{(y)}$ and $S_{n+1}^{(y)}$.

Intuitively, this is equivalent to $\hat{C}_\alpha^\pi(X_{n+1}; \mathcal{Y}_n) := \{y \in \mathcal{Y}_n : p(y) > \alpha\}$, as in (3), after replacing the unweighted p-value $p(y)$ defined in (2) with its weighted counterpart

$$p^\pi(y) := w_{n+1}(y) + \sum_{j \in \mathcal{D}^{\text{cal}}} w_j(y) \mathbb{I}\left[S_j \geq S_{n+1}\right].$$

In fact, if the probability function π is constant, it is easy to verify that $w_j(y) = 1/(1 + |\mathcal{D}^{\text{cal}}|)$.

The following result establishes that this weighted conformal prediction set satisfies the closed-set validity requirement needed by Theorem 5 to guaranteed the validity of our open-set method, for any choice of the function π.

Theorem 6. *Assume $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable. For any $\alpha \in (0, 1)$, the conformal prediction set $\hat{C}_\alpha^\pi(X_{n+1}; \mathcal{Y}_n)$ constructed as in* (18) *using a fixed probability function π satisfies:*

$$\mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha^\pi(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \in \mathcal{Y}_n\right] \leq \alpha.$$

Returning to the choice of the probability function π, recall that its purpose is to favor assigning rarer labels to the training set, where they are most useful. We propose setting $\pi(1) = 0$, ensuring that singleton labels are always used for training. For labels with frequency $k \geq 2$, we use $\pi(k) = \min\{n^{\text{cal}}/n(1 - p_1),\ 1\}$, where $n^{\text{cal}} \leq n$ is the target calibration set size (chosen by the user) and p_1 denotes the proportion of singleton labels in the population. In practice, we recommend the plug-in estimate $p_1 = M_1/n$. Strictly speaking, this substitution is not fully valid, since M_1 is random and Theorem 6 assumes π is fixed. Nonetheless, our simulations demonstrate that this heuristic works well in practice. Other choices of π are also possible.

5 Empirical Demonstrations

5.1 Setup

We empirically evaluate the validity and effectiveness of the conformal Good–Turing classification method in open-set classification scenarios, using both simulated and real datasets.

We compare our method against the standard closed-set conformal approach reviewed in Section 2.2, which naively assumes $\mathcal{Y} = \mathcal{Y}_n$. Both methods are implemented with two sample-splitting strategies: (i) standard random splitting and (ii) the selective splitting approach introduced in Section 4.3. In the standard random splitting case, exactly 90% of the samples are assigned to the training set and the remaining 10% to the calibration set. In the selective splitting approach, the split is randomized according to the probability function π described in Section 4.3, resulting in approximately 90% of the data being used for training on average.

This setup yields four approaches under comparison, all using the same k-nearest neighbor classifier with $k = 5$, and the same inverse quantile nonconformity scores [37]. For the conformal Good–Turing method, feature-based conformal Good–Turing p-values (implemented as detailed in Section 3.3) are used by default when testing H_{unseen}, while randomized feature-blind conformal Good–Turing p-values (Section 3.2) are used when testing H_{seen}. The constants c_k used to compute ψ_{seen} in (12) are set according to the power-law weighting scheme defined in (13), with $\beta = 1.6$. The significance level α is allocated using the cross-validation tuning strategy introduced in Section 4.2. Additional implementation details are in Appendix A4.

Prediction sets are evaluated based on two criteria: coverage and informativeness. Coverage is measured by whether the true label Y_{n+1} is contained in the prediction set; when Y_{n+1} is a new label, coverage equals one if and only if the prediction set includes the joker. Informativeness is assessed in two ways: first, by the cardinality of the prediction set, defined as in (15) with the joker counted as one element; and second, by the frequency with which the joker appears in the prediction sets. Smaller sets and less frequent use of the joker indicate higher informativeness.

All reported metrics are averaged over multiple independent repetitions using different training, calibration, and test data sets.

5.2 Numerical Experiments with Synthetic Data

We begin by evaluating the proposed method on synthetic data generated under a Dirichlet process model, as described in Example 1. The labels $Y_1, \ldots, Y_{n+1} \in \mathbb{R}$ are drawn from a Dirichlet process with (nonatomic) base distribution $P_0 = \text{Uniform}(0, 1)$ and concentration parameter $\theta > 0$, which governs the probability of new labels. Conditional on each label Y_i, the corresponding feature vector $X_i \in \mathbb{R}^3$ is independently sampled from a multivariate Gaussian distribution with 3 independent coordinates, mean Y_i, and variance 5×10^{-6}. The total sample size is $n = 2000$.

Figure 1 summarizes the results, comparing the performance of different methods across a range of θ values (from 0 to 1000). Each point represents an average over 1000 test samples and 50 independent repetitions of the experiment.

The first panel empirically verifies Theorem 1, showing that standard conformal classification methods fail to maintain the nominal 90% marginal coverage as the proportion of unseen labels increases. In contrast, consistent with Theorem 5, conformal Good–Turing classification maintains valid marginal coverage by incorporating the joker symbol to account for new labels. The second panel illustrates the efficiency advantage of the selective sample splitting strategy, which yields substantially smaller (and thus more informative) prediction sets than random splitting, while maintaining valid coverage. The two rightmost panels further demonstrate that the proposed method produces informative prediction sets by using the joker symbol parsimoniously, including it only when necessary to maintain coverage due to high prevalence of new labels.

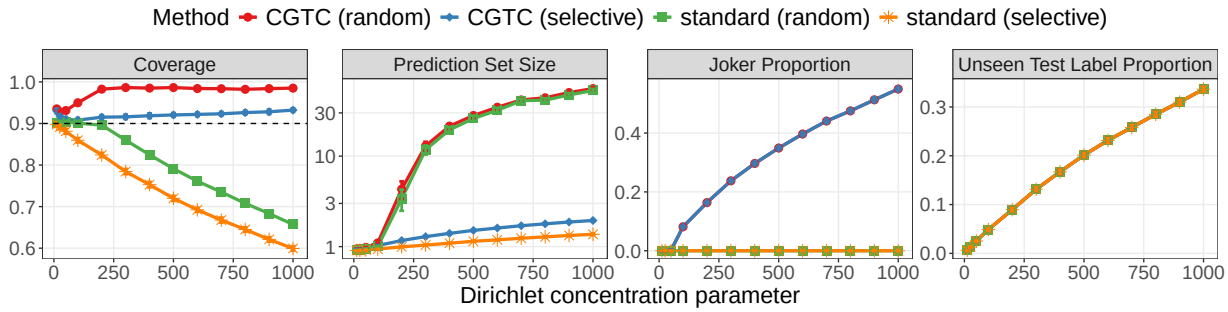

Figure 1: Performance of conformal prediction sets constructed using different methods on synthetic data generated from a Dirichlet process model, as a function of the concentration parameter θ. Larger θ values correspond to a higher probability of unseen labels at test time. The nominal coverage level is 90%. The conformal Good–Turing classification (CGTC) method maintains valid marginal coverage even when new labels are common by incorporating a "catch-all" joker symbol into the prediction sets. When combined with the selective sample-splitting strategy, it also produces smaller and more informative prediction sets. Error bars represent ±1.96 standard errors.

Figure 2 provides a more detailed view of the results from Figure 1 by stratifying coverage according to the frequency of the test label. Specifically, coverage is computed by grouping test points based on the true frequency of their labels, divided into four bins. The "very rare" bin includes labels that were either unseen or appeared only once in the combined training and calibration sets, while the remaining bins correspond to labels of increasing frequency. These results demonstrate that the standard conformal method performs poorly on very rare labels, whereas the conformal Good–Turing approach achieves higher coverage not only marginally but also separately across all frequency levels.

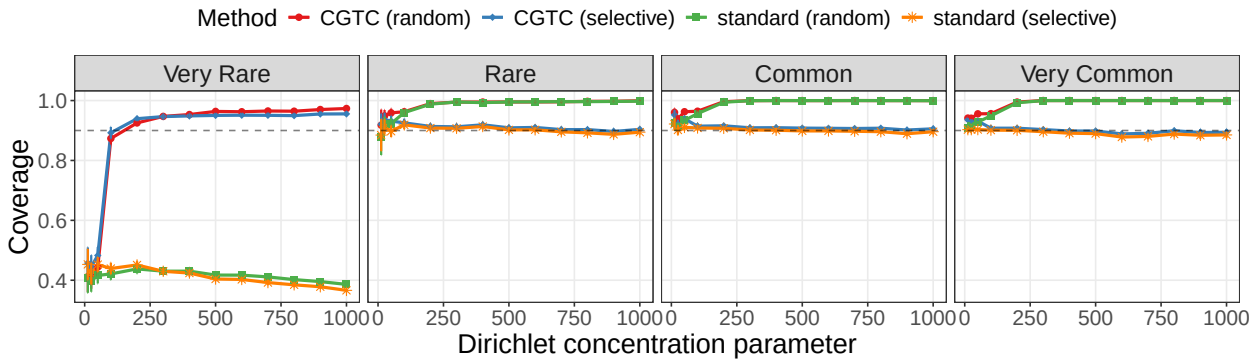

Figure 2: Coverage of conformal prediction sets constructed using different methods on synthetic data from a Dirichlet process model, as in Figure 1, stratified by test-label frequency. The standard approach fails to achieve coverage for very rare labels, whereas the conformal Good–Turing method is valid for both rare and common labels.

Finally, Figure 3 illustrates the advantages of incorporating feature information into the conformal Good–Turing p-values (Section 3.3), compared with the feature-blind deterministic and randomized versions introduced in Section 3.2. The feature-based statistic ψ_0^{XGT} achieves the highest power, most frequently rejecting H_{unseen} and thereby minimizing the inclusion of the joker symbol in the prediction sets shown in Figure 1. In contrast, the feature-blind variants ψ_0^{GT} and ψ_0^{RGT} are substantially more conservative; if used to

construct the prediction sets in Figure 1, they would yield less informative results with more frequent joker inclusion. The randomized version ψ_0^{RGT} offers only a modest gain over its deterministic counterpart.

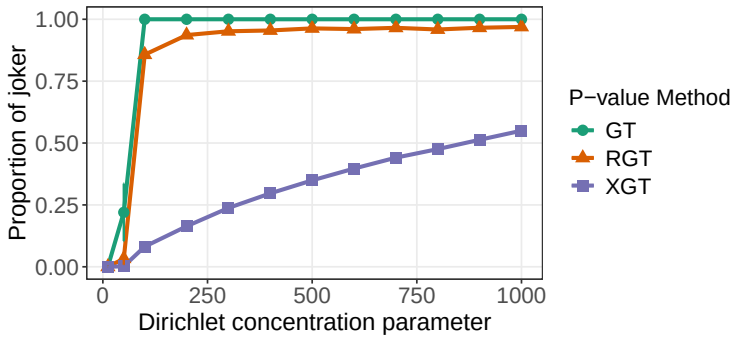

Figure 3: Percentage of conformal Good–Turing prediction sets including the joker symbol under different p-value constructions for testing the null hypothesis H_{unseen}, which states that the test point corresponds to a new label, on synthetic data. The feature-based method (XGT)—corresponding to the results shown in Figure 1—makes the most parsimonious use of the joker by attaining the highest power to reject H_{unseen}. In contrast, the deterministic feature-blind approach (GT) is overly conservative, and its randomized variant (RGT) reduces this conservatism only slightly.

In conclusion, these results show that conformal Good–Turing classification, when combined with selective sample splitting, outperforms standard benchmarks by producing more informative prediction sets while maintaining valid marginal and empirically higher conditional coverage. Additional experimental results are presented in Appendix A4: Figure A1 illustrates the allocation of significance levels, Figure A2 provides a detailed comparison of the three conformal Good–Turing p-value constructions, and Figures A3–A5 evaluate performance under an alternative setting where the concentration parameter θ is fixed and the sample size varies.

5.3 Numerical Experiments with Real Data

We next evaluate the methods from the previous section on real data using the CelebFaces Attributes (CelebA) dataset, which contains 202,599 facial images of size 178×218 pixels from 10,177 celebrity identities. Each image is preprocessed using a pre-trained multitask cascaded convolutional network for face detection and alignment [45] and FaceNet [46] for feature extraction, resulting in a robust embedding matrix with corresponding identity labels. To construct a more manageable (for repeated experiments) yet representative subset, we randomly sample 2,000 identities and retain all associated images. Each experiment then partitions this subset into training, calibration, and test sets, as detailed in the previous section. All methods are applied to target 80% marginal coverage.

Figure 4 reports on the performances of the prediction sets as a function of reference sample size, defined as the total number of training plus calibration samples, which is varied between 2,000 and 6,000. Results are averaged across 1,000 test points and 20 independent experiments. Across all sizes, our conformal Good–Turing classification method achieves valid marginal coverage, whereas the standard conformal approach under-covers when the the reference sample size is small and the test points are thus likely to contain previously unseen identities. Moreover, these results confirm that the selective sample splitting strategy substantially reduces the average size of the prediction sets. Therefore, the combination of the conformal Good–Turing classification with selective splitting again leads to the most informative prediction sets.

Additional experimental results on the CelebA dataset are presented in Appendix A4, yielding qualitatively consistent findings. Figure A6 reports coverage stratified by test-label frequency. Figure A7 illustrates the

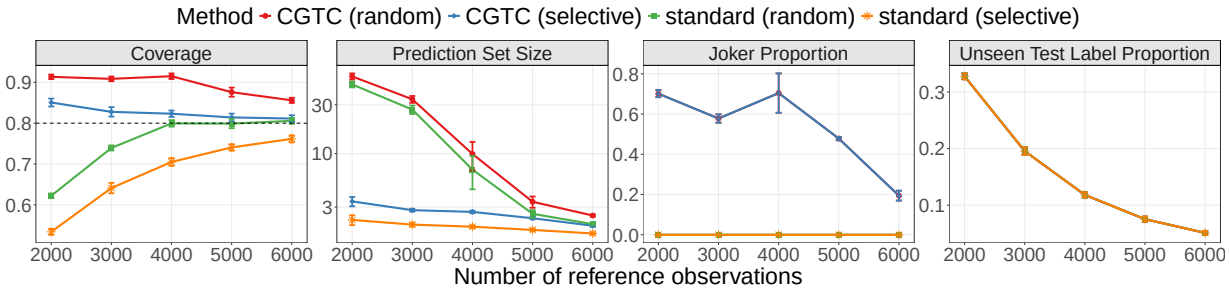

Figure 4: Performance of conformal prediction sets constructed using different methods on face recognition data from the CelebA resource, as a function of the total number of labeled data points. Conformal Good–Turing classification (CGTC) maintains valid marginal coverage across all settings and achieves smaller prediction sets when combined with the selective sample splitting strategy. Error bars indicate 1.96 standard errors.

adaptive allocation of significance levels obtained through cross-validation tuning, while Figure A8 compares the three conformal Good–Turing p-value constructions. Finally, Figures A9–A11 analyze the performance of conformal Good–Turing classification as the total number of possible identities varies.

6 Discussion

This paper extends the applicability of conformal prediction to open-set classification tasks, revealing an intriguing connection between this modern topic and classical statistical work on species sampling models [16]. In addition, we introduced a selective sample-splitting strategy that substantially improves the performance of the proposed Good–Turing classification method in open-set settings. Beyond this context, the same strategy may also prove broadly useful for closed-set conformal classification, particularly in the presence of imbalanced data.

A limitation of conformal Good–Turing classification lies in its potential conservativeness. Because the significance levels for its classification and testing components are combined via a union bound to ensure a theoretical lower bound on coverage (Theorem 5), no corresponding upper bound is guaranteed. In particular, since the total error budget α is divided into α_{class}, α_{unseen}, and α_{seen}—each strictly less than α—our method tends to be more conservative than standard conformal prediction in closed-set settings, where the joker symbol is unnecessary. Although some conservativeness is inevitable, as our approach operates under weaker assumptions that permit new labels to arise at test time, this effect is largely mitigated by data-driven tuning (Section 4.2), which performs well empirically despite lacking formal theoretical justification.

Several directions offer promising opportunities for future work. First, relaxing the assumption of exchangeability [47] would enable the method to address distributional shifts, such as covariate or class-conditional shift, potentially through additional re-weighting [43]. Second, our method could be adapted to online settings [14, 48], where observations arrive sequentially and the label set $\mathcal{Y}_n$ expands over time; in practice, such an extension would be most useful if it could be implemented efficiently without retraining from scratch. Third, this method could be extended to handle structured label spaces, such as ordinal or hierarchical labels. Finally, while this work tests separately the complementary null hypotheses H_{unseen} and H_{seen}, splitting the error budget across them, alternative joint testing formulations may yield tighter coverage guarantees.

Software Availability

A software package implementing the methods described in this paper and containing code needed to reproduce all numerical experiments is available online at `https://github.com/TianminX/open-set-conformal-classification`.

Acknowledgments

T. X., Y. Z., and M. S were partly supported by a USC-Capital One CREDIF award. M. S. was also partly supported by a Google Research Scholar Award.

References

[1] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.

[2] Fayin Li and Harry Wechsler. Open set face recognition using transduction. *IEEE transactions on pattern analysis and machine intelligence*, 27(11):1686–1697, 2005.

[3] Hitham Jleed and Martin Bouchard. Open set audio recognition for multi-class classification with rejection. *IEEE Access*, 8:146523–146534, 2020.

[4] Jie Sun, Shipeng Zhao, Sheng Miao, Xuan Wang, and Yanan Yu. Open-set iris recognition based on deep learning. *IET image processing*, 16(9):2361–2372, 2022.

[5] Alexander Cao, Diego Klabjan, and Yuan Luo. Open-set recognition of breast cancer treatments. *Artificial intelligence in medicine*, 135:102451, 2023.

[6] Charles H Brenner. Fundamental problem of forensic mathematics - the evidential value of a rare haplotype. *Forensic Science International: Genetics*, 4(5):281–191, 2010.

[7] Yefeng Shen, Md Zakir Hossain, Khandaker Asif Ahmed, and Shafin Rahman. An open set model for pest identification. *Computational Biology and Chemistry*, 108:108002, 2024.

[8] Steve Cruz, Cora Coleman, Ethan M Rudd, and Terrance E Boult. Open set intrusion recognition for fine-grained attack categorization. In *2017 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–6. IEEE, 2017.

[9] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*, 2020.

[10] Pedro Ribeiro Mendes Júnior, Terrance E Boult, Jacques Wainer, and Anderson Rocha. Open-set support vector machines. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(6):3785–3798, 2021.

[11] Guanchao Feng, Dhruv Desai, Stefano Pasquali, and Dhagash Mehta. Open set recognition for random forest. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 45–53, 2024.

[12] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[13] Emily R Bartusiak and Edward J Delp. Transformer-based speech synthesizer attribution in an open set scenario. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 329–336. IEEE, 2022.

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

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

[16] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[17] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[18] Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult. Probability models for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2317–2324, 2014. doi: 10.1109/TPAMI.2014.2321392.

[19] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1563–1572, 2016. doi: 10.1109/CVPR.2016.173.

[20] ZongYuan Ge, Sergey Demyanov, Zetao Chen, and Rahil Garnavi. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*, 2017.

[21] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. Classification-reconstruction learning for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4016–4025, 2019.

[22] Poojan Oza and Vishal M. Patel. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2307–2316, 2019.

[23] Leonid Erlygin and Alexey Zaytsev. Holistic uncertainty estimation for open-set recognition. *arXiv preprint arXiv:2408.14229*, 2024.

[24] Yotam Hechtlinger, Barnabás Póczos, and Larry Wasserman. Cautious deep learning, 2018. arXiv:1805.09460.

[25] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33, 2020.

[26] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

[27] Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 64555–64576. Curran Associates, Inc., 2023.

[28] Leying Guan and Robert Tibshirani. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society Series B*, 84(2):524–546, 2022.

[29] Zhou Wang and Xingye Qiao. Set-valued classification with out-of-distribution detection for many classes. *Journal of Machine Learning Research*, 24(375):1–39, 2023.

[30] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

[31] Ziyi Liang, Matteo Sesia, and Wenguang Sun. Integrative conformal p-values for out-of-distribution testing with labelled outliers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):671–693, 2024.

[32] Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183, 2024.

[33] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[34] David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[35] Albert Y Lo. On a class of Bayesian nonparametric estimates:i. density estimate. *The Annals of Statistics*, 12(1):351–357, 1984.

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

[37] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Curran Associates Inc., 2020. ISBN 9781713829546.

[38] Mary M Moya, Mark W Koch, and Larry D Hostetler. One-class classifier networks for target recognition applications. Technical report, Sandia National Labs., Albuquerque, NM (United States), 1993.

[39] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal processing*, 99:215–249, 2014.

[40] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[41] Ramon Ferrer i Cancho and Ricard V Solé. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8(3):165–173, 2001.

[42] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books, 2016.

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

[44] Matteo Sesia, Stefano Favaro, and Edgar Dobriban. Conformal frequency estimation using discrete sketched data with coverage for distinct queries. *Journal of Machine Learning Research*, 24(348):1–80, 2023.

[45] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yunde Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016. ISSN 1070-9908. doi: 10.1109/LSP.2016.2603342.

[46] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. doi: 10.1109/CVPR.2015.7298682.

[47] Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

[48] Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. *arXiv preprint arXiv:2402.01139*, 2024.

[49] Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024.

A1 Review of Closed-Set Conformal Classification

A1.1 Non-Conformity Scores Based on Generalized Inverse Quantiles

We briefly review the construction of standard conformal prediction sets based on the *generalized inverse quantile* non-conformity scores proposed by [25], which we adopt for the numerical experiments presented in this paper. Compared with the classical probability scores introduced in Section 2.2, these scores are more sophisticated but offer the advantage of improved conditional coverage, as they explicitly account for potential heteroscedasticity in the conditional distribution of $Y \mid X$.

Assume that the label space $\mathcal{Y}$ is finite with cardinality $K = |\mathcal{Y}|$. Without loss of generality, let $\mathcal{Y} = [K] := \{1, \ldots, K\}$. For any $x \in \mathcal{X}$ and $k \in [K]$, let $\hat{\pi}(x, k)$ denote an estimate of $\mathbb{P}[Y = k \mid X = x]$ produced by a classification model (e.g., the output of the final softmax layer in a neural network).

For any $x \in \mathcal{X}$ and $t \in [0, 1]$, let $\hat{\pi}_{(1)}(x) \geq \ldots \geq \hat{\pi}_{(K)}(x)$ denote the descending order statistics of $\hat{\pi}(x, 1), \ldots, \hat{\pi}(x, K)$. Similarly, let $\hat{r}(x, \hat{\pi}, k)$ denote the rank of $\hat{\pi}(x, k)$ among the values $\hat{\pi}(x, 1), \ldots, \hat{\pi}(x, K)$. Using this notation, we can define the generalized cumulative distribution function:

$$\hat{\Pi}(x, \hat{\pi}, k) = \hat{\pi}_{(1)}(x) + \hat{\pi}_{(2)}(x) + \ldots + \hat{\pi}_{(\hat{r}(x, \hat{\pi}, k))}(x).$$

Following the notation from Section 2.2, and with sufficient generality to cover both split- and full-conformal approaches, the non-conformity scores of [25] can be written as:

$$S_i^{(y)} := \hat{\Pi}(X_i, \hat{\pi}^{(y)}, Y_i), \quad i \in [n], \qquad\qquad S_{n+1}^{(y)} := \hat{\Pi}(X_{n+1}, \hat{\pi}^{(y)}, y), \tag{A19}$$

where $\hat{\pi}^{(y)}$ represents an estimate of $\mathbb{P}[Y = k \mid X = x]$ obtained from a model trained on the augmented dataset $\mathcal{D}(y) := \{(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y)\}$.

We now recall a convenient and easily computable characterization of the conformal prediction set $\hat{C}_\alpha(X_{n+1}; \mathcal{Y})$, as defined in (3), corresponding to this choice of non-conformity scores.

A1.2 Implementation of Closed-Set Split-Conformal Classification

In the split-conformal implementation of closed-set classification, where the model $\hat{\pi}$ is trained on an independent dataset that does not depend on the placeholder test label y, the non-conformity scores can be written more simply as $S_i = \hat{\Pi}(X_i, \hat{\pi}, Y_i)$, for $i \in [n]$. Then, with the choice of scores reviewed in Section A1.1, the prediction set $\hat{C}_\alpha(X_{n+1}; \mathcal{Y})$ in (3) can be conveniently expressed as follows.

For any $x \in \mathbb{R}^d$ and $t \in [0, 1]$, define the generalized quantile function:

$$\hat{Q}(x, \hat{\pi}, t) := \min\{k \in \{1, \ldots, K\} : \hat{\pi}_{(1)}(x) + \hat{\pi}_{(2)}(x) + \ldots + \hat{\pi}_{(k)}(x) \geq t\}. \tag{A20}$$

This makes it possible to write $\hat{C}_\alpha(X_{n+1}; \mathcal{Y})$ in (3) as:

$$\hat{C}_\alpha(X_{n+1}; \mathcal{Y}) := \{k \in \mathcal{Y} : \hat{r}(X_{n+1}, \hat{\pi}, k) \leq \hat{Q}(X_{n+1}, \hat{\pi}, \hat{\tau}_{n,\alpha})\}, \tag{A21}$$

where

$$\hat{\tau}_{n,\alpha} = \lceil (1 + n) \cdot (1 - \alpha) \rceil\text{-th smallest value in } \{S_1, \ldots, S_n\}. \tag{A22}$$

This procedure is summarized by Algorithm A1, under the closed-set assumption that the label space $\mathcal{Y}$ is known. If the full label space is not known in advance, but is instead replaced by the plug-in estimate $\mathcal{Y}_n := \text{unique}(Y_1, \ldots, Y_n)$, then Algorithm A1 becomes Algorithm A2.

Algorithm A1. Standard Split-Conformal Classification with Known Label Space

Input : Calibration data set $\mathcal{D}^{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^{n}$;
Training data set $\mathcal{D}^{\text{train}} = \{(X_i^{\text{train}}, Y_i^{\text{train}})\}_{i=1}^{n_{\text{train}}}$;
Known label space $\mathcal{Y} = \{1, \ldots, K\}$;
Unlabeled test point with features X_{n+1};
Machine-learning algorithm $\mathcal{A}$ for training a K-class classification model;
Desired miscoverage probability $\alpha \in (0, 1)$.
Train model $\hat{\pi}$ by fitting classifier $\mathcal{A}$ on $\mathcal{D}^{\text{train}}$;
Compute non-conformity scores $S_i = \hat{s}(X_i, \tilde{Y}_i)$ for all $i \in [n]$; e.g., by using (A19);
Compute threshold $\hat{\tau}_{n,\alpha}$ as in (A22);
Evaluate prediction set $\hat{C}_\alpha(X_{n+1}; \mathcal{Y})$ as in (A21).
return $\hat{C}_\alpha(X_{n+1}; \mathcal{Y}) \subseteq \mathcal{Y}$.

We refer to [25] for additional details on closed-set conformal classification, including a rigorous justification of why the non-conformity scores reviewed in Section A1.1 tend to achieve relatively high conditional coverage. We also refer to [25] for details on a randomized variant of these non-conformity scores that retains the same theoretical guarantees while producing even more informative prediction sets. Our framework can seamlessly incorporate this randomized extension; indeed, it is the approach used in the empirical results presented in Section 5. However, to avoid unnecessary notational complexity, we omit the explicit description of this extension here.

Algorithm A2. Standard Split-Conformal Classification with Plug-In Label Space

Input : Data set $\{(X_i, Y_i)\}_{i=1}^n$;
Unlabeled test point with features X_{n+1};
Machine-learning algorithm $\mathcal{A}$ for training a K-class classification model;
Desired miscoverage probability $\alpha \in (0, 1)$.

Define observed label space $\mathcal{Y}_n := \text{unique}(Y_1, \ldots, Y_n)$;
Randomly split $[n]$ into two disjoint subsets, $\mathcal{I}^{\text{train}}$ and $\mathcal{I}^{\text{cal}}$;
Apply Algorithm A1 with $\mathcal{Y} = \mathcal{Y}_n$, $\mathcal{D}^{\text{cal}} = \{(X_i, Y_i)\}_{i \in \mathcal{I}^{\text{cal}}}$, and $\mathcal{D}^{\text{train}} = \{(X_i, Y_i)\}_{i \in \mathcal{I}^{\text{train}}}$;
return $\hat{C}_\alpha(X_{n+1}; \mathcal{Y}) \subseteq \mathcal{Y}_n$.

A2 Additional Methodological Details

A2.1 Data-Driven Allocation of Significance Budget

Algorithm A3. Data-Driven Allocation of Significance Budget

Input : tuning data $\{Z_1, \ldots, Z_n\}$; total significance budget α; number of folds K; grid size G for α_{class}; preference parameter $\lambda \in [0, 1]$.

// Initialization

best_loss $\leftarrow \infty$; $\quad (\alpha^*_{\text{seen}}, \alpha^*_{\text{class}}) \leftarrow (0, 0)$

$\alpha^{\text{candidates}}_{\text{seen}} \leftarrow \{0, 0.01, 0.02, 0.05, 0.1\} \cap [0, \alpha)$

foreach $\alpha_{\text{seen}} \in \alpha^{\text{candidates}}_{\text{seen}}$ **do**

  $\alpha_{\text{remaining}} \leftarrow \alpha - \alpha_{\text{seen}}$

  $\alpha^{\text{grid}}_{\text{class}} \leftarrow \{0, \frac{\alpha_{\text{remaining}}}{G}, \frac{2\alpha_{\text{remaining}}}{G}, \ldots, \alpha_{\text{remaining}}\}$

  foreach $\alpha_{\text{class}} \in \alpha^{\text{grid}}_{\text{class}}$ **do**

    $\alpha_{\text{unseen}} \leftarrow \alpha_{\text{remaining}} - \alpha_{\text{class}}$

    Partition labeled data into K folds $\mathcal{D}_1, \ldots, \mathcal{D}_K$

    for $k \leftarrow 1$ **to** K **do**

      $\mathcal{D}^{(k)}_{\text{train}} \leftarrow \bigcup_{j \neq k} \mathcal{D}_j$; $\quad \mathcal{D}^{(k)}_{\text{val}} \leftarrow \mathcal{D}_k$

      foreach $(X_i, Y_i) \in \mathcal{D}^{(k)}_{\text{val}}$ **do**

        Apply conformal Good–Turing classification for X_i using $\mathcal{D}^{(k)}_{\text{train}}$ as reference data at $(\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}})$

        Compute point loss $\ell_i \leftarrow L(\hat{C}^*_\alpha(X_i), Y_i; \lambda)$ for $i \in \mathcal{D}^{(k)}_{\text{val}}$ as in in (16)

      Compute fold loss $\hat{L}_k \leftarrow \frac{1}{|\mathcal{D}^{(k)}_{\text{val}}|} \sum_{(X_i, Y_i) \in \mathcal{D}^{(k)}_{\text{val}}} \ell_i$

    Compute average loss $\bar{L}(\alpha_{\text{seen}}, \alpha_{\text{class}}) \leftarrow \frac{1}{K} \sum_{k=1}^{K} \hat{L}_k$

    if $\bar{L}(\alpha_{\text{seen}}, \alpha_{\text{class}}) <$ *best_loss* **then**

      best_loss $\leftarrow \bar{L}(\alpha_{\text{seen}}, \alpha_{\text{class}})$

      $(\alpha^*_{\text{seen}}, \alpha^*_{\text{class}}) \leftarrow (\alpha_{\text{seen}}, \alpha_{\text{class}})$

$\alpha^*_{\text{unseen}} \leftarrow \alpha - \alpha^*_{\text{seen}} - \alpha^*_{\text{class}}$

return $(\alpha^*_{\text{class}}, \alpha^*_{\text{unseen}}, \alpha^*_{\text{seen}})$

A2.2 Fast Computation of Conformalization Weights for Selective Sample Splitting

In Section 4.3, we introduced a selective sample-splitting procedure for constructing the conformal prediction set $\hat{C}^{\pi}_{\alpha}(X_{n+1}; \mathcal{Y}_n)$ as defined in (18). This construction requires computing the "conformalization" weights $w_j(y)$ for each calibration point $j \in \mathcal{D}^{\text{cal}}$ and each candidate label $y \in \mathcal{Y}_n$. In this section, we describe how these weights can be computed efficiently.

Recall that the original definition of the weights in (18) is given by

$$w_j(y) = \frac{p^{(j)}(y)}{p^{(n+1)}(y) + \sum_{k \in \mathcal{D}^{\text{cal}}} p^{(k)}(y)},$$

where $p^{(j)}(y)$ involve products over all n data points:

$$p^{(j)}(y) = \pi\big(N(y; Y^{(j,y)}_{1:n})\big) \prod_{\substack{i \in \mathcal{D}^{\text{cal}} \\ i \neq j}} \pi\big(N(Y^{(j,y)}_i; Y^{(j,y)}_{1:n})\big) \prod_{i \in \mathcal{D}^{\text{train}}} \big(1 - \pi(N(Y^{(j,y)}_i; Y^{(j,y)}_{1:n}))\big).$$

Computing each of these weights directly would require evaluating a product over all n samples; consequently, computing all weights naively incurs a total cost of $O(n^2)$ per test point, which is computationally prohibitive. However, with a more careful implementation this cost can be reduced from $O(n^2)$ to $O(n)$.

The key insight is that swapping the test point with calibration point j affects only the frequencies of two labels: y (the test label) and Y_j (the label at position j). Therefore, instead of calculating the entire product, we can focus only on the factors that change.

To make this intuition precise, define the set of calibration labels as $\mathcal{Y}_{\text{cal}} := \{Y_i : i \in \mathcal{D}^{\text{cal}}\}$. The complete set of labels involved in the weight calculations is then $\mathcal{T} := \mathcal{Y}_{\text{cal}} \cup \{y\}$. For any label $\ell \notin \{y, Y_j\}$, the frequency $N(\ell; Y_{1:n}^{(j,y)}) = N(\ell; Y_{1:n})$ remains unchanged after swapping. Consequently, the corresponding factors in $p^{(j)}(y)$ are identical across different values of j and will cancel in the normalization.

Let us define the product of unchanged factors when swapping the test label and the label at position j:

$$A_j := \prod_{\substack{i=1 \\ Y_i \notin \{y, Y_j\}}}^{n} \begin{cases} \pi\big(N(Y_i; Y_{1:n})\big), & \text{if } i \in \mathcal{D}^{\text{cal}}, \\ 1 - \pi\big(N(Y_i; Y_{1:n})\big), & \text{if } i \in \mathcal{D}^{\text{train}}. \end{cases}$$

To simplify the notation further, introduce the frequency counts within each split:

$$f_{\text{cal}}(\ell) := \sum_{i \in \mathcal{D}^{\text{cal}}} \mathbb{I}\{Y_i = \ell\}, \quad f_{\text{train}}(\ell) := \sum_{i \in \mathcal{D}^{\text{train}}} \mathbb{I}\{Y_i = \ell\}.$$

Note that $f_{\text{cal}}(\ell) + f_{\text{train}}(\ell) = N(\ell; Y_{1:n})$ for any label ℓ.

When $y \neq Y_j$, the weight $p^{(j)}(y)$ can be expressed as:

$$\begin{aligned} p^{(j)}(y) = &\big[\pi\big(N(Y_j; Y_{1:n}) - 1\big)\big]^{f_{\text{cal}}(Y_j)-1} \big[1 - \pi\big(N(Y_j; Y_{1:n}) - 1\big)\big]^{f_{\text{train}}(Y_j)} \\ &\times \big[\pi\big(N(y; Y_{1:n}) + 1\big)\big]^{f_{\text{cal}}(y)+1} \big[1 - \pi\big(N(y; Y_{1:n}) + 1\big)\big]^{f_{\text{train}}(y)} \cdot A_j. \end{aligned}$$

For the baseline case without swapping, we have:

$$\begin{aligned} p^{(n+1)}(y) = &\big[\pi\big(N(Y_j; Y_{1:n})\big)\big]^{f_{\text{cal}}(Y_j)} \big[1 - \pi\big(N(Y_j; Y_{1:n})\big)\big]^{f_{\text{train}}(Y_j)} \\ &\times \big[\pi\big(N(y; Y_{1:n})\big)\big]^{f_{\text{cal}}(y)} \big[1 - \pi\big(N(y; Y_{1:n})\big)\big]^{f_{\text{train}}(y)} \cdot A_j. \end{aligned}$$

The key computational insight is to work with the ratios:

$$\tilde{p}^{(j)}(y) := \frac{p^{(j)}(y)}{p^{(n+1)}(y)}.$$

Since the common factor A_j cancels in this ratio, we need only compute the changes in the factors corresponding to labels y and Y_j.

The normalized weights can then be equivalently expressed as:

$$w_j(y) = \frac{p^{(j)}(y)}{p^{(n+1)}(y) + \sum_{k \in \mathcal{D}^{\text{cal}}} p^{(k)}(y)} = \frac{\tilde{p}^{(j)}(y)}{1 + \sum_{k \in \mathcal{D}^{\text{cal}}} \tilde{p}^{(k)}(y)}. \tag{A23}$$

This computational shortcut reduces the complexity from $O(n)$ per weight to $O(1)$ per weight after an initial $O(n)$ preprocessing step to compute the frequency counts $f_{\text{cal}}(\cdot)$ and $f_{\text{train}}(\cdot)$.

A3 Mathematical Proofs

A3.1 From Closed-Set Conformal Classification to Open-Set Scenarios

Lemma A1. *Assume* $\{(X_i, Y_i)\}_{i=1}^{n+1}$ *are exchangeable. Let* $\hat{C}_\alpha(X_{n+1}; \mathcal{Y}_{n+1})$ *denote the conformal prediction set computed as in* (3)*, using a symmetric score function* s *and replacing the full label dictionary* $\mathcal{Y}$ *with* $\mathcal{Y}_{n+1} := unique(Y_1, \ldots, Y_{n+1})$*. Assume also that the nonconformity scores* $\{S_1^{(y)}, \ldots, S_{n+1}^{(y)}\}$ *are almost-surely distinct for any* $y \in \mathcal{Y}$*, Then,*

$$1 - \alpha \leq \mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_{n+1})\right] \leq 1 - \alpha + \frac{1}{n+1}.$$

of Lemma A1. With $\hat{C}_\alpha(X_{n+1}; \mathcal{Y})$ instead of $\hat{C}_\alpha(X_{n+1}; \mathcal{Y}_{n+1})$, this would be a standard result in conformal prediction; e.g., see [49]. A similar proof also works with $\mathcal{Y}_{n+1}$.

We begin by recalling that the nonconformity scores $\{S_i^{(Y_{n+1})}\}_{i=1}^n$ are exchangeable because they are obtained by applying a symmetric score function s to exchangeable data. This exchangeability is unaffected by conditioning on $\mathcal{Y}_{n+1}$ because the latter is a function of the unordered set $\{Y_1, \ldots, Y_{n+1}\}$, which is invariant to any permutations. Combined with the almost-sure distinctiveness of the scores, this implies that $p(Y_{n+1})$ has a uniform distribution on $\{1/(n+1), 2/(n+2), \ldots, 1\}$ conditional on $\mathcal{Y}_{n+1}$. Therefore,

$$\alpha - \frac{1}{n+1} \leq \mathbb{P}\left[p(Y_{n+1}) \leq \alpha \mid \mathcal{Y}_{n+1}\right] \leq \alpha,$$

and hence:

$$1 - \alpha \leq \mathbb{E}\left[\mathbb{P}\left[p(Y_{n+1}) > \alpha \mid \mathcal{Y}_{n+1}\right]\right] \leq 1 - \alpha + \frac{1}{n+1}.$$

Finally, the proof is completed by noting that $Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_{n+1})$ if and only if $p(Y_{n+1}) > \alpha$. □

of Theorem 1. By definition of $\hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n)$,

$$\begin{aligned}
&\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n)\right] \\
&\quad = \mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \in \mathcal{Y}_n\right] + \mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \notin \mathcal{Y}_n\right] \\
&\quad = \mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \in \mathcal{Y}_n\right] \\
&\quad = \mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_{n+1}), Y_{n+1} \in \mathcal{Y}_n\right] \\
&\quad \leq \min\left\{\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_{n+1})\right], \mathbb{P}\left[Y_{n+1} \in \mathcal{Y}_n\right]\right\} \\
&\quad \leq \min\left\{1 - \alpha + \frac{1}{n+1}, 1 - \mathbb{P}\left[N_{n+1}\right]\right\} \\
&\quad = 1 - \alpha - \mathbb{P}\left[N_{n+1}\right] + \min\left\{\mathbb{P}\left[N_{n+1}\right] + \frac{1}{n+1}, \alpha\right\},
\end{aligned}$$

where the last inequality above follows from Lemma A1. Moreover,

$$\begin{aligned}
&\mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n)\right] \\
&\quad = \mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \in \mathcal{Y}_n\right] + \mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \notin \mathcal{Y}_n\right] \\
&\quad = \mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_{n+1}), Y_{n+1} \in \mathcal{Y}_n\right] + \mathbb{P}\left[Y_{n+1} \notin \mathcal{Y}_n\right] \\
&\quad \leq \mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_{n+1})\right] + \mathbb{P}\left[Y_{n+1} \notin \mathcal{Y}_n\right] \\
&\quad \leq \alpha + \mathbb{P}\left[N_{n+1}\right],
\end{aligned}$$

where the last inequality above follows from Lemma A1. Combining these two results gives:

$$1-\alpha-\mathbb{P}\left[N_{n+1}\right] \leq \mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha(X_{n+1}; \mathcal{Y}_n)\right] \leq 1-\alpha+\min\left\{\frac{1}{n+1}, \alpha-\mathbb{P}\left[N_{n+1}\right]\right\}.$$

□

A3.2 Main Auxiliary Lemma for Conformal Good–Turing p-Values

Lemma A2. *Let $\{Z_i = (X_i, Y_i)\}_{i=1}^{n+1}$ be exchangeable random variables, where $X_i \in \mathcal{X}$ are features and $Y_i \in \mathcal{Y}$ are labels. Let also $\{\eta_i\}_{i=1}^{n+1} \in \mathbb{R}^{n+1}$ represent an independent collection of $n+1$ exchangeable random variables. Consider a symmetric score function $f^{(k)}$ that may take one of the following forms:*

(i) Label-only: $f^{(k)} : \mathcal{Y} \times \mathcal{Y}^{n+1} \times \mathbb{R} \to \mathbb{R}$, whose first argument depends only on labels;

(ii) Feature-based: $f^{(k)} : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})^{n+1} \times \mathbb{R} \to \mathbb{R}$, whose first argument depends also on features.

For any candidate value $y \in \mathcal{Y}$ for Y_{n+1}, define the scores:

$$F_i^{(k,y)} := \begin{cases} f^{(k)}(Y_i; \{Y_1, \ldots, Y_n, y\}, \eta_i), & \text{if label-only,} \\ f^{(k)}(Z_i; \{Z_1, \ldots, Z_n, (X_{n+1}, y)\}, \eta_i), & \text{if feature-based,} \end{cases}$$

for $i \in [n]$, and

$$F_{n+1}^{(k,y)} := \begin{cases} f^{(k)}(y; \{Y_1, \ldots, Y_n, y\}, \eta_{n+1}), & \text{if label-only,} \\ f^{(k)}((X_{n+1}, y); \{Z_1, \ldots, Z_n, (X_{n+1}, y)\}, \eta_{n+1}), & \text{if feature-based.} \end{cases}$$

Define also the imaginary conformal p-value:

$$\widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, y) := \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{I}[F_i^{(k,y)} \geq F_{n+1}^{(k,y)}],$$

leaving its dependence on $\{\eta_i\}_{i=1}^{n+}$ implicit. Then, for any $u \in (0,1)$:

$$\mathbb{P}\{\widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, Y_{n+1}) \leq u\} \leq u.$$

Furthermore, assume that for any two values $y, y' \in \mathcal{Y}$ both appearing exactly k times in $\{Y_1, \ldots, Y_n\}$, the following invariance property holds:

$$\widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, y) = \widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, y').$$

Then, under this additional assumption, we can define a computable p-value ψ_k that depends only on $(Z_1, \ldots, Z_n, X_{n+1})$ such that:

$$\mathbb{P}\left[\psi_k(Z_1, \ldots, Z_n, X_{n+1}) \leq u, H_k\right] \leq u,$$

where $H_k : \{\sum_{i=1}^n \mathbb{I}\left[Y_i = Y_{n+1}\right] = k\}$. In particular, $\psi_k(Z_1, \ldots, Z_n, X_{n+1})$ is defined as the common value of $\widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, y)$ for any y satisfying H_k; that is, $\psi_k(Z_1, \ldots, Z_n, X_{n+1}) := \widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, y)$ for y such that $\sum_{i=1}^n \mathbb{I}\left[Y_i = y\right] = k$. This is uniquely well-defined under the above invariance assumption.

of Lemma A2. We first prove $\widetilde{\psi}_k = \widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, Y_{n+1})$ is super-uniform. Note that $\widetilde{\psi}_k$ is the relative rank of $F_{n+1}^{(k,Y_{n+1})}$ among $\{F_1^{(k,Y_{n+1})}, \ldots, F_{n+1}^{(k,Y_{n+1})}\}$. Therefore, to establish super-uniformity, it suffices to show $(F_1^{(k,Y_{n+1})}, \ldots, F_{n+1}^{(k,Y_{n+1})})$ are exchangeable. We prove this under the more general feature-based case; the label-only case follows similarly.

Let σ be any permutation of $[n+1]$, and define $\widetilde{Z}_j = Z_{\sigma(j)}$, $\widetilde{Y}_j = Y_{\sigma(j)}$, and $\widetilde{\eta}_j = \eta_{\sigma(j)}$ for $j = 1, \ldots, n+1$. For each $i \in [n]$, define:

$$\widetilde{F}_i^{(k,\widetilde{Y}_{n+1})} := f^{(k)}(\widetilde{Z}_i; \{\widetilde{Z}_1, \ldots, \widetilde{Z}_n, \widetilde{Z}_{n+1}\}, \widetilde{\eta}_i),$$
$$\widetilde{F}_{n+1}^{(k,\widetilde{Y}_{n+1})} := f^{(k)}(\widetilde{Z}_{n+1}; \{\widetilde{Z}_1, \ldots, \widetilde{Z}_n, \widetilde{Z}_{n+1}\}, \widetilde{\eta}_{n+1}).$$

By the symmetry of the function $f^{(k)}$ (it only depends on the multiset structure, not the ordering),

$$(F_{\sigma(1)}^{(k,Y_{n+1})}, \ldots, F_{\sigma(n+1)}^{(k,Y_{n+1})}) = (\widetilde{F}_1^{(k,\widetilde{Y}_{n+1})}, \ldots, \widetilde{F}_{n+1}^{(k,\widetilde{Y}_{n+1})}).$$

Since $((Z_1, \eta_1), \ldots, (Z_{n+1}, \eta_{n+1}))$ are jointly exchangeable, it follows that:

$$(\widetilde{F}_1^{(k,\widetilde{Y}_{n+1})}, \ldots, \widetilde{F}_{n+1}^{(k,\widetilde{Y}_{n+1})}) \overset{d}{=} (F_1^{(k,Y_{n+1})}, \ldots, F_{n+1}^{(k,Y_{n+1})}).$$

Therefore, the non-conformity scores are exchangeable, and the relative rank of $F_{n+1}^{(k,Y_{n+1})}$ among $\{F_1^{(k,Y_{n+1})}, \ldots, F_{n+1}^{(k,Y_{n+1})}\}$ follows a discrete uniform distribution on $\{1/(n+1), 2/(n+1), \ldots, 1\}$, yielding:

$$\mathbb{P}\left[\widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, Y_{n+1}) \le u\right] \le u \quad \text{for all } u \in (0, 1).$$

This completes the first part of the proof.

We now discuss the construction of a computable p-value that does not depend on the unknown label Y_{n+1}. Suppose that, for any two values $y, y' \in \mathcal{Y}$ both appearing exactly k times in $\{Y_1, \ldots, Y_n\}$:

$$\widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, y) = \widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, y').$$

Under this invariance property, we can define $\psi_k(Z_1, \ldots, Z_n, X_{n+1})$ as the common value of $\widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, y)$ for any y satisfying H_k; that is, $\psi_k(Z_1, \ldots, Z_n, X_{n+1}) := \widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, y)$ for y such that $\sum_{i=1}^n \mathbb{I}[Y_i = y] = k$. This is uniquely well-defined under the above invariance assumption. Then, for any $u \in (0, 1)$:

$$\begin{aligned}\mathbb{P}[\psi_k(Z_1, \ldots, Z_n, X_{n+1}) \le u, H_k] &= \mathbb{P}\left[\widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, Y_{n+1}) \le u, H_k\right] \\ &\le \mathbb{P}\left[\widetilde{\psi}_k(Z_1, \ldots, Z_n, X_{n+1}, Y_{n+1}) \le u\right] \\ &\le u,\end{aligned}$$

where the last inequality follows from the first part of this result. □

A3.3 Feature-Blind Conformal Good–Turing p-Values

Lemma A3. *Let $\psi'_k : \mathcal{Y}^n \to \mathbb{R}$ be a function such that $\psi'_k(Y_1, \ldots, Y_n)$ depends only on the frequency profile $\mathcal{F}_n$ of $(Y_1, \ldots, Y_n)$ without regard to their order or actual face values. If ψ'_k is a valid p-value for all exchangeable distributions over $(Y_1, \ldots, Y_n, Y_{n+1})$, that is, for all $u \in (0, 1)$:*

$$\mathbb{P}\left[\psi'_k(Y_1, \ldots, Y_n) \le u, H_k\right] \le u,$$

then for any sequence $(Y_1, \ldots, Y_n)$ with frequency profile $\mathcal{F}_n$:

$$\psi'_k(Y_1, \ldots, Y_n) \ge \frac{(k+1)M_{k+1} + k + 1}{n+1},$$

where M_{k+1} is the number of distinct labels appearing exactly $k+1$ times in $(Y_1, \ldots, Y_n)$.

of Lemma A3. Throughout this proof, we use $\psi'_k(\mathcal{F}_n)$ and $\psi'_k(Y_1, \ldots, Y_n)$ interchangeably, since $\psi'_k(Y_1, \ldots, Y_n)$ depends on $(Y_1, \ldots, Y_n)$ only through $\mathcal{F}_n$. We prove this result by contradiction. Assume there exists a frequency profile $\mathcal{F}_n$ with M_{k+1} distinct labels appearing exactly $k+1$ times such that:

$$\psi'_k(\mathcal{F}_n) < \frac{(k+1)M_{k+1} + k + 1}{n+1}.$$

We construct a specific sequence of labels $(y_1, \ldots, y_n, y_{n+1}) \in \mathcal{Y}^{n+1}$ as follows. Specifically, we include M_{k+1} distinct labels, each appearing exactly $k+1$ times among positions 1 through n. We also include one additional label y^* appearing exactly k times among positions 1 through n, and set $y_{n+1} = y^*$. Any remaining positions are filled with labels of frequencies different from k and $k+1$.

Consider the uniform distribution over all permutations of this sequence. Let π denote a uniformly random permutation, so:

$$(Y_1, \ldots, Y_n, Y_{n+1}) = (y_{\pi(1)}, \ldots, y_{\pi(n)}, y_{\pi(n+1)}).$$

Set $u = \frac{1}{2}\left(\psi'_k(\mathcal{F}_n) + \frac{(k+1)M_{k+1}+k+1}{n+1}\right)$, so that:

$$\psi'_k(\mathcal{F}_n) < u < \frac{(k+1)M_{k+1} + k + 1}{n+1}.$$

Now we count the permutations where H_k occurs; i.e., where Y_{n+1} appears exactly k times in $Y_1, \ldots, Y_n$. The event H_k occurs if and only if position $n+1$ contains one of the $(k+1)M_{k+1}$ copies of labels with frequency $k+1$, one of the k copies of y^* from positions 1 through n, or the original $y_{n+1} = y^*$ itself.

Let $\mathcal{P}$ denote this set of permutations. We have $|\mathcal{P}| = ((k+1)M_{k+1} + k + 1) \cdot n!$ since there are $(k+1)M_{k+1} + k + 1$ choices for which element goes to position $n+1$, and the remaining n positions can be arranged in $n!$ ways. For any $\pi \in \mathcal{P}$, the sequence $(Y_1, \ldots, Y_n) = (y_{\pi(1)}, \ldots, y_{\pi(n)})$ has frequency profile $\mathcal{F}_n$. Since ψ'_k depends only on the frequency profile, we have:

$$\psi'_k(Y_1, \ldots, Y_n) = \psi'_k(\mathcal{F}_n) < u \quad \text{whenever } \pi \in \mathcal{P}.$$

Therefore:

$$\begin{aligned}
\mathbb{P}\left[\psi'_k(Y_1, \ldots, Y_n) \le u, H_k\right] &\ge \mathbb{P}\left[\pi \in \mathcal{P}\right] \\
&= \frac{|\mathcal{P}|}{(n+1)!} \\
&= \frac{((k+1)M_{k+1} + k + 1) \cdot n!}{(n+1)!} \\
&= \frac{(k+1)M_{k+1} + k + 1}{n+1} \\
&> u,
\end{aligned}$$

contradicting the validity requirement $\mathbb{P}\left[\psi'_k(Y_1, \ldots, Y_n) \le u, H_k\right] \le u$,. □

of Theorem 2. We prove both the validity and optimality of ψ_k^{GT}. We first demonstrate the validity. By Lemma A2 (applied in the label-only setting, ignoring the η noise variables), it suffices to specify an appropriate score function $f^{(k)}$ and show that under H_k, the augmented conformal p-value $\widetilde{\psi}_k$ is invariant to the specific value of Y_{n+1}. Define the following score function, for $k \in \{0, 1, 2, \ldots\}$:

$$f^{(k)}(y; A) := \phi_{k+1}(y; A),$$

where for any multiset $A \subseteq \mathcal{Y}$ and label $y \in \mathcal{Y}$,

$$\phi_{k+1}(y; A) := \mathbb{I}\left[\sum_{a \in A} \mathbb{I}[a = y] = k + 1\right].$$

In words, $\phi_{k+1}(y; A)$ is an indicator that the label y appears exactly $k + 1$ times in the multiset A.

Under H_k (that Y_{n+1} appears exactly k times in $\{Y_1, \ldots, Y_n\}$), we have $\phi_{k+1}(Y_{n+1}; \{Y_1, \ldots, Y_n, Y_{n+1}\}) = 1$, since Y_{n+1} appears $k + 1$ times in the augmented set. Following the construction in Lemma A2, for $i \in [n]$:

$$F_i^{(k,Y_{n+1})} = \phi_{k+1}(Y_i; \{Y_1, \ldots, Y_n, Y_{n+1}\}),$$
$$F_{n+1}^{(k,Y_{n+1})} = \phi_{k+1}(Y_{n+1}; \{Y_1, \ldots, Y_n, Y_{n+1}\}) = 1.$$

The indicator $\mathbb{I}[F_i^{(k,Y_{n+1})} \geq F_{n+1}^{(k,Y_{n+1})}] = \mathbb{I}[F_i^{(k,Y_{n+1})} \geq 1] = \mathbb{I}[F_i^{(k,Y_{n+1})} = 1]$ equals 1 if and only if Y_i appears exactly $k + 1$ times in the augmented set $\{Y_1, \ldots, Y_n, Y_{n+1}\}$. This occurs when:

- $Y_i = Y_{n+1}$: Since Y_{n+1} appears k times in $\{Y_1, \ldots, Y_n\}$, there are exactly k such indices.
- $Y_i \neq Y_{n+1}$: Then Y_i must appear exactly $k + 1$ times in $\{Y_1, \ldots, Y_n\}$. The total number of such indices is $(k + 1)M_{k+1}$, where M_{k+1} is the number of distinct labels with frequency $k + 1$.

Therefore, under H_k, we can compute:

$$\begin{aligned}
\widetilde{\psi}_k &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{I}[F_i^{(k,Y_{n+1})} \geq F_{n+1}^{(k,Y_{n+1})}] \\
&= \frac{1}{n+1} \left(\mathbb{I}[F_{n+1}^{(k,Y_{n+1})} = 1] + \sum_{i=1}^{n} \mathbb{I}[F_i^{(k,Y_{n+1})} = 1] \right) \\
&= \frac{1}{n+1} \left(1 + \sum_{i \in [n]: Y_i = Y_{n+1}} 1 + \sum_{i \in [n]: Y_i \neq Y_{n+1}, \phi_{k+1}(Y_i; \{Y_1, \ldots, Y_n\}) = 1} 1 \right) \\
&= \frac{1}{n+1} (1 + k + (k+1)M_{k+1}) \\
&= \frac{(k+1)M_{k+1} + k + 1}{n+1},
\end{aligned}$$

which depends only on the frequency profile and not on the specific value of Y_{n+1}. Thus we can set:

$$\psi_k^{\mathrm{GT}} := \frac{(k+1)M_{k+1} + k + 1}{n+1}. \tag{A24}$$

By Lemma A2, $\mathbb{P}\left[\psi_k^{\mathrm{GT}} \leq u, H_k\right] \leq u$ for all $u \in (0, 1)$.

We then prove the optimality. Let ψ'_k be any deterministic function of $\mathcal{F}_n$ satisfying $\mathbb{P}\left[\psi'_k \leq u, H_k\right] \leq u$ for all $u \in (0, 1)$. By Lemma A3, for any frequency profile $\mathcal{F}_n$ with M_{k+1} distinct labels appearing exactly $k + 1$ times:

$$\psi'_k(\mathcal{F}_n) \geq \frac{(k+1)M_{k+1} + k + 1}{n+1}.$$

Since $\psi_k^{\mathrm{GT}} = \frac{(k+1)M_{k+1}+k+1}{n+1}$ by definition, we have:

$$\psi'_k(\mathcal{F}_n) \geq \psi_k^{\mathrm{GT}}(\mathcal{F}_n)$$

for every possible frequency profile $\mathcal{F}_n$. Therefore, $\psi'_k \geq \psi_k^{\mathrm{GT}}$ almost surely, establishing that ψ_k^{GT} is most powerful among all deterministic p-values that depend only on the frequency profile. □

of Proposition 1. Define a randomized label-only score function as in Lemma A2:

$$f^{(k)}(y; A, \eta) = \phi_{k+1}(y; A) \cdot \eta.$$

Let $\{\eta_i\}_{i=1}^{n+1}$ be i.i.d. Uniform$[0,1]$ random variables, independent of $\{Y_i\}_{i=1}^{n+1}$. Under H_k, following the construction in Lemma A2 (applied in the label-only setting, including the η noise variables),

$$\begin{aligned} F_i^{(k,Y_{n+1})} &= \phi_{k+1}(Y_i; \{Y_1, \ldots, Y_n, Y_{n+1}\}) \cdot \eta_i, \\ F_{n+1}^{(k,Y_{n+1})} &= \phi_{k+1}(Y_{n+1}; \{Y_1, \ldots, Y_n, Y_{n+1}\}) \cdot \eta_{n+1} = \eta_{n+1}, \end{aligned}$$

where the second equation follows from $\phi_{k+1}(Y_{n+1}; \{Y_1, \ldots, Y_n, Y_{n+1}\}) = 1$. Therefore, the indicator $\mathbb{I}[F_i^{(k,Y_{n+1})} \geq F_{n+1}^{(k,Y_{n+1})}]$ equals 1 if and only if $\phi_{k+1}(Y_i; \{Y_1, \ldots, Y_n, Y_{n+1}\}) = 1$ and $\eta_i \geq \eta_{n+1}$.

Under H_k, the indices satisfying $\phi_{k+1}(Y_i; \{Y_1, \ldots, Y_n, Y_{n+1}\}) = 1$ are indices in $\mathcal{S}_{k+1}$ (labels with frequency $k+1$ in the original data) and indices where $Y_i = Y_{n+1}$ (exactly k such indices, forming a subset of $\mathcal{S}_k$). Thus there are $(k+1)M_{k+1} + k$ such indices in total. Since $\{\eta_i\}_{i=1}^{n+1}$ are i.i.d. continuous random variables, they are almost surely distinct. By exchangeability, η_{n+1} has equal probability of ranking in any position among these $(k+1)M_{k+1} + k + 1$ relevant η values (those with $\phi_{k+1} = 1$ plus η_{n+1} itself). Therefore, the number of indices with both $\phi_{k+1} = 1$ and $\eta_i \geq \eta_{n+1}$ follows a discrete uniform distribution on $\{0, 1, \ldots, (k+1)M_{k+1} + k\}$. The valid conformal p-value under H_k becomes:

$$\widetilde{\psi}_k = \frac{1}{n+1}\left(1 + \sum_{i \in \mathcal{S}_{k+1}} \mathbb{I}[\eta_i \geq \eta_{n+1}] + \sum_{j: Y_j = Y_{n+1}} \mathbb{I}[\eta_j \geq \eta_{n+1}]\right) = \frac{U+1}{n+1},$$

where $U \sim \text{Uniform}\{0, 1, \ldots, (k+1)M_{k+1} + k\}$. Since this depends only on $\mathcal{F}_n$ and not on the specific value of Y_{n+1} under H_k, we can define:

$$\psi_k^{\text{RGT}} := \frac{\text{Uniform}\{0, 1, \ldots, (k+1)M_{k+1} + k\} + 1}{n+1}.$$

By Lemma A2, we have $\mathbb{P}\left[\psi_k^{\text{RGT}} \leq u, H_k\right] \leq u$ for all $u \in (0,1)$. □

A3.4 Feature-Based Conformal Good–Turing p-Values

of Theorem 3. We apply Lemma A2 with a feature-based score function:

$$f^{(k)}((x,y); \mathcal{D}) = \frac{\phi_{k+1}(y; \{Y_j : Z_j \in \mathcal{D}\})}{\hat{s}_k(x)},$$

where $\hat{s}_k : \mathcal{X} \to \mathbb{R}_+$ is a feature score function quantifying how atypical feature x is for a label with frequency k.

Under H_k, since Y_{n+1} appears exactly k times in $\{Y_1, \ldots, Y_n\}$, we have $\phi_{k+1}(Y_{n+1}; \{Y_1, \ldots, Y_n, Y_{n+1}\}) = 1$. Therefore:

$$F_{n+1}^{(k,Y_{n+1})} = \frac{1}{\hat{s}_k(X_{n+1})}.$$

For $i \in [n]$, we have:

$$F_i^{(k,Y_{n+1})} = \frac{\phi_{k+1}(Y_i; \{Y_1, \ldots, Y_n, Y_{n+1}\})}{\hat{s}_k(X_i)}.$$

The indicator $\mathbb{I}[F_i^{(k,Y_{n+1})} \geq F_{n+1}^{(k,Y_{n+1})}]$ equals 1 if and only if:

1. $\phi_{k+1}(Y_i; \{Y_1, \ldots, Y_n, Y_{n+1}\}) = 1$ (i.e., Y_i appears exactly $k+1$ times in the augmented set), and
2. $\hat{s}_k(X_{n+1}) \geq \hat{s}_k(X_i)$.

Under H_k, the indices satisfying $\phi_{k+1} = 1$ are precisely:

1. Indices in $\mathcal{S}_{k+1} = \{i \in [n] : \sum_{j=1}^{n} \mathbb{I}[Y_i = Y_j] = k+1\}$ (labels with frequency $k+1$ in the original data)
2. Indices where $Y_i = Y_{n+1}$ (exactly k such indices under H_k)

Therefore, the augmented conformal p-value is:

$$\widetilde{\psi}_k = \frac{1}{n+1}\left(1 + \sum_{i \in \mathcal{S}_{k+1}} \mathbb{I}\{\hat{s}_k(X_{n+1}) \geq \hat{s}_k(X_i)\} + \sum_{j \in [n]: Y_j = Y_{n+1}} \mathbb{I}\{\hat{s}_k(X_{n+1}) \geq \hat{s}_k(X_j)\}\right).$$

Case $k = 0$: Under H_0, Y_{n+1} is a new label not in $\{Y_1, \ldots, Y_n\}$. Thus the second sum vanishes, making $\widetilde{\psi}_0$ invariant to the specific value of Y_{n+1}. We can therefore define:

$$\psi_0^{\mathrm{XGT}} := \frac{1}{n+1}\left(1 + \sum_{i \in \mathcal{S}_1} \mathbb{I}\{\hat{s}_0(X_{n+1}) \geq \hat{s}_0(X_i)\}\right).$$

Case $k > 0$: The second sum depends on which specific label (among those appearing k times) equals Y_{n+1}. Under H_k, there are M_k distinct labels with frequency k, and Y_{n+1} must be one of them. To ensure validity without knowing Y_{n+1}, we take the conservative approach of maximizing over all M_k possible values:

$$\psi_k^{\mathrm{XGT}} := \frac{1}{n+1}\left(1 + \sum_{i \in \mathcal{S}_{k+1}} \mathbb{I}\{\hat{s}_k(X_{n+1}) \geq \hat{s}_k(X_i)\} + \max_{y \in \{Y_i : i \in \mathcal{S}_k\}} \sum_{j \in [n]: Y_j = y} \mathbb{I}\{\hat{s}_k(X_{n+1}) \geq \hat{s}_k(X_j)\}\right).$$

Therefore, under H_k, we have:

$$\psi_k^{\mathrm{XGT}} \geq \widetilde{\psi}_k \quad \text{almost surely,}$$

since the maximum is at least as large as the value for the true (but unknown) Y_{n+1}.

Therefore, for any $u \in (0, 1)$:

$$\mathbb{P}\left[\psi_k^{\mathrm{XGT}} \leq u, H_k\right] \leq \mathbb{P}\left[\widetilde{\psi}_k \leq u, H_k\right] \leq u,$$

where the last inequality follows from Lemma A2, provided $\hat{s}_k$ is either fixed or invariant to permutations of $\{(X_i, Y_i)\}_{i=1}^{n+1}$ under H_k. □

A3.5 Testing the Composite Hypothesis that Y_{n+1} is Previously Seen Label

of Theorem 4. Recall the definition in (11): $H_{\mathrm{seen}} : \bigcup_{k \in K_n} H_k$. Now consider the probability

$$\mathbb{P}\left[\psi_{\mathrm{seen}} \leq u, H_{\mathrm{seen}}\right] = \sum_{k \in K_n} \mathbb{P}\left[\psi_{\mathrm{seen}} \leq u, H_k\right].$$

By the definition of ψ_{seen} in (12),

$$\begin{aligned}
\sum_{k \in K_n} \mathbb{P}\left[\psi_{\text{seen}} \le u, H_k\right] &= \sum_{k \in K_n} \mathbb{P}\left[\max_{j \in K_n} \frac{\psi_j}{c_j} \le u, H_k\right] \\
&\le \sum_{k \in K_n} \mathbb{P}\left[\frac{\psi_k}{c_k} \le u, H_k\right] \\
&= \sum_{k \in K_n} \mathbb{P}\left[\psi_k \le c_k u, H_k\right] \\
&\le \sum_{k \in K_n} c_k u && \text{(since } \mathbb{P}\left[\psi_k \le v, H_k\right] \le v\text{)} \\
&\le u. && \text{(since } \sum_{k=1}^{n} c_k \le 1\text{)}
\end{aligned}$$

This establishes the validity of ψ_{seen} as a conformal p-value for H_{seen}. □

A3.6 Conformal Good–Turing Classification

of Theorem 5. Let $N_{n+1} := \{Y_{n+1} \notin \mathcal{Y}_n\}$ denote the event that Y_{n+1} is a new label (i.e., not observed in the first n data points), and let N_{n+1}^c denote its complement. Note that under N_{n+1}, the hypothesis H_{unseen} is true, while under N_{n+1}^c, the hypothesis H_{seen} is true. The miscoverage probability can be decomposed as:

$$\begin{aligned}
&\mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha^*(X_{n+1})\right] \\
&\quad = \mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha^*(X_{n+1}),\, N_{n+1}^c\right] + \mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha^*(X_{n+1}),\, N_{n+1}\right] \\
&\quad = \mathbb{P}\left[Y_{n+1} \notin \hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n),\, \psi_{\text{seen}} > \alpha_{\text{seen}},\, N_{n+1}^c\right] + \mathbb{P}\left[\psi_{\text{seen}} \le \alpha_{\text{seen}},\, N_{n+1}^c\right] \\
&\qquad + \mathbb{P}\left[\circledast \notin \hat{C}_\alpha^*(X_{n+1}), N_{n+1}\right] \\
&\quad = \mathbb{P}\left[Y_{n+1} \notin \hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n),\, \psi_{\text{seen}} > \alpha_{\text{seen}},\, Y_{n+1} \in \mathcal{Y}_n\right] + \mathbb{P}\left[\psi_{\text{seen}} \le \alpha_{\text{seen}},\, H_{\text{seen}}\right] \\
&\qquad + \mathbb{P}\left[\psi_{\text{unseen}} \le \alpha_{\text{unseen}},\, H_{\text{unseen}}\right] \\
&\quad \le \alpha_{\text{class}} + \alpha_{\text{seen}} + \alpha_{\text{unseen}} \\
&\quad = \alpha.
\end{aligned}$$

The second equality follows from the definition in (14). The inequality follows from: (i) the validity assumption of the closed-set conformal prediction set, (ii) the validity assumption of ψ_{seen} under H_{seen}, and (iii) the validity assumption of ψ_{unseen} under H_{unseen}. □

of Proposition 2. We first show that $\psi_{\text{unseen}} \le (K+1)/(n+1)$ almost surely for all three p-value variants.

For the feature-blind Good–Turing p-value, recall that M_1 denotes the number of distinct labels that appear exactly once in $Y_1, \ldots, Y_n$. Since $|\mathcal{Y}| = K$, we have $M_1 \le K$. Therefore:

$$\psi_0^{\text{GT}} = \frac{M_1 + 1}{n+1} \le \frac{K+1}{n+1} \quad \text{almost surely.}$$

By construction of the randomized feature-blind Good–Turing p-value and the feature-based conformal Good–Turing p-value in Section 3, we have

$$\psi_0^{\text{RGT}} \le \psi_0^{\text{GT}} \quad \text{and} \quad \psi_0^{\text{XGT}} \le \psi_0^{\text{GT}} \quad \text{almost surely.}$$

Thus, $\psi_{\text{unseen}} \le (K+1)/(n+1)$ almost surely for any of the three p-value choices.

With our allocation $\alpha_{\text{unseen}} = (K+1)/(n+1)$, we have $\psi_{\text{unseen}} \leq \alpha_{\text{unseen}}$ almost surely. Since $\alpha_{\text{seen}} = 0$ and ψ_{seen} is strictly positive, we have $\psi_{\text{seen}} > \alpha_{\text{seen}}$ almost surely.

Therefore, the conditions in (14) output the first case:

$$\hat{C}^*_\alpha(X_{n+1}) = \hat{C}_{\alpha_{\text{class}}}(X_{n+1}; \mathcal{Y}_n) \quad \text{almost surely.}$$

This completes the proof. □

A3.7 Selective Sample Splitting

of Theorem 6. Let $\hat{C}^\pi_\alpha(X_{n+1}; \mathcal{Y}_{n+1})$ denote the conformal prediction set computed as in (18), using a symmetric score function s and replacing the full label dictionary $\mathcal{Y}$ with $\mathcal{Y}_{n+1} := \text{unique}(Y_1, \ldots, Y_{n+1})$. Assume also that the nonconformity scores $\{S_1, \ldots, S_{n+1}\}$ are almost-surely distinct for any $y \in \mathcal{Y}$.

A key observation is that $Y_{n+1} \in \mathcal{Y}_n$ implies $\hat{C}^\pi_\alpha(X_{n+1}; \mathcal{Y}_n) = \hat{C}^\pi_\alpha(X_{n+1}; \mathcal{Y}_{n+1})$ almost surely:

$$\mathbb{I}\{Y_{n+1} \notin \hat{C}^\pi_\alpha(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \in \mathcal{Y}_n\} = \mathbb{I}\{Y_{n+1} \notin \hat{C}^\pi_\alpha(X_{n+1}; \mathcal{Y}_{n+1}), Y_{n+1} \in \mathcal{Y}_n\}.$$

To leverage this observation, we now analyze the coverage of $\hat{C}^\pi_\alpha(X_{n+1}; \mathcal{Y}_{n+1})$. By construction of the weights, we have the normalization property:

$$w_{n+1}(y) + \sum_{j \in \mathcal{D}^{\text{cal}}} w_j(y) = 1.$$

This allows us to express the coverage condition through a quantile representation. Specifically, we have:

$$\begin{aligned}
& Y_{n+1} \in \hat{C}^\pi_\alpha(X_{n+1}; \mathcal{Y}_{n+1}) \\
\iff & w_{n+1}(Y_{n+1}) + \sum_{j \in \mathcal{D}^{\text{cal}}} w_j(Y_{n+1}) \mathbb{I}\left[S_j \geq S_{n+1}\right] > \alpha \\
\iff & S_{n+1} \;\leq\; \text{Quantile}\Big(1-\alpha;\; \sum_{j \in \mathcal{D}^{\text{cal}}} w_j(Y_{n+1})\, \delta_{S_j} \;+\; w_{n+1}(Y_{n+1})\, \delta_\infty\Big) \\
\iff & S_{n+1} \;\leq\; \text{Quantile}\Big(1-\alpha;\; \sum_{j \in \mathcal{D}^{\text{cal}}} w_j(Y_{n+1})\, \delta_{S_j} \;+\; w_{n+1}(Y_{n+1})\, \delta_{S_{n+1}}\Big)
\end{aligned}$$

Next, we need to find the distribution of S_{n+1} under the selective sampling scheme. Let E_{n+1} denote the multiset event that $\{Z_1, \ldots, Z_{n+1}\} = \{z_1, \ldots, z_{n+1}\}$, where $Z_i = (X_i, Y_i)$ and $z_i = (x_i, y_i)$. To derive the conditional distribution, we consider the probability that the test score equals any particular calibration score. For some fixed $D^{\text{train}} \in \mathcal{P}([n])$ and $j \in D^{\text{cal}} = [n] \setminus D^{\text{train}}$, we have:

$$\begin{aligned}
& \mathbb{P}\Big\{S_{n+1} = S_j \mid E_{n+1},\, \mathcal{D}^{\text{train}} = D^{\text{train}},\, \{Z_i\}_{i \in \mathcal{D}^{\text{train}}} = \{z_i\}_{i \in D^{\text{train}}}\Big\} \\
& \quad = \mathbb{P}\Big\{(X_{n+1}, Y_{n+1}) = (x_j, y_j) \mid E_{n+1},\, \mathcal{D}^{\text{train}} = D^{\text{train}},\, \{Z_i\}_{i \in \mathcal{D}^{\text{train}}} = \{z_i\}_{i \in D^{\text{train}}}\Big\} \\
& \quad = \frac{\mathbb{P}\Big\{(X_{n+1}, Y_{n+1}) = (x_j, y_j),\, \mathcal{D}^{\text{train}} = D^{\text{train}},\, \{Z_i\}_{i \in \mathcal{D}^{\text{train}}} = \{z_i\}_{i \in D^{\text{train}}} \mid E_{n+1}\Big\}}{\mathbb{P}\Big\{\mathcal{D}^{\text{train}} = D^{\text{train}},\, \{Z_i\}_{i \in \mathcal{D}^{\text{train}}} = \{z_i\}_{i \in D^{\text{train}}} \mid E_{n+1}\Big\}}.
\end{aligned} \tag{A25}$$

The numerator can be rewritten by considering the equivalent event where we specify the calibration set instead of the training set.

$$\begin{aligned}
&\mathbb{P}\Big\{Z_{n+1} = z_j,\, \mathcal{D}^{\text{train}} = D^{\text{train}},\, \{Z_i\}_{i\in\mathcal{D}^{\text{train}}} = \{z_i\}_{i\in D^{\text{train}}} \mid E_{n+1}\Big\} \\
&\quad = \mathbb{P}\Big\{Z_{n+1} = z_j,\, \mathcal{D}^{\text{cal}} = D^{\text{cal}},\, \{Z_i\}_{i\in\mathcal{D}_{\text{cal}}} = \{z_{n+1}\} \cup \{z_i\}_{i\in D^{\text{cal}}\setminus\{j\}} \mid E_{n+1}\Big\} \\
&\quad = \mathbb{P}\left\{\mathcal{D}^{\text{cal}} = D^{\text{cal}},\, \{Z_i\}_{i\in\mathcal{D}_{\text{cal}}} = \{z_{n+1}\} \cup \{z_i\}_{i\in D^{\text{cal}}\setminus\{j\}} \mid Z_{n+1} = z_j,\, E_{n+1}\right\} \mathbb{P}\Big\{Z_{n+1} = z_j \mid E_{n+1}\Big\}.
\end{aligned} \tag{A26}$$

To compute this conditional probability, recall that under our selective sampling model, the inclusion indicators $\{I_i\}_{i=1}^n$ are conditionally independent given the label sequence $Y_{1:n}$. Specifically, for each $i \in [n]$:

$$I_i \mid Y_{1:n} \sim \text{Bernoulli}\big(\pi(N(Y_i; Y_{1:n}))\big),$$

and these indicators are mutually independent conditional on $Y_{1:n}$.

Consider first the probability of observing the unswapped calibration set configuration when $Z_{n+1} = z_{n+1}$. For any $D^{\text{cal}} \subseteq [n]$, we have:

$$\begin{aligned}
&\mathbb{P}\left\{\mathcal{D}^{\text{cal}} = D^{\text{cal}},\, \{Z_i\}_{i\in\mathcal{D}_{\text{cal}}} = \{z_i\}_{i\in D^{\text{cal}}} \,\middle|\, Z_{n+1} = z_{n+1},\, E_{n+1}\right\} \\
=&\mathbb{P}\left\{\bigcap_{i\in D^{\text{cal}}} \{I_i = 1\} \cap \bigcap_{i\in[n]\setminus D^{\text{cal}}} \{I_i = 0\} \,\middle|\, Z_{n+1} = z_{n+1},\, E_{n+1}\right\} \\
=&\prod_{i\in D^{\text{cal}}} \pi\big(N(Y_i; Y_{1:n})\big) \prod_{i\in[n]\setminus D^{\text{cal}}} \big(1 - \pi(N(Y_i; Y_{1:n}))\big)
\end{aligned} \tag{A27}$$

This product form arises directly from the independence structure: each point's inclusion probability depends only on its label frequency in $Y_{1:n}$. Note that the ordering of labels within the calibration and training sets does not affect this probability, as it depends only on the label frequencies.

Now consider the case where we swap the test point with a calibration point $j \in D^{\text{cal}}$. We want to compute:

$$\mathbb{P}\left\{\mathcal{D}^{\text{cal}} = D^{\text{cal}},\, \{Z_i\}_{i\in\mathcal{D}_{\text{cal}}} = \{z_{n+1}\} \cup \{z_i\}_{i\in D^{\text{cal}}\setminus\{j\}} \,\middle|\, Z_{n+1} = z_j,\, E_{n+1}\right\}$$

As introduced in Section 4.3, this swapping operation transforms the label sequence to $Y_{1:n}^{(j,y)}$, where:

$$Y_i^{(j,y)} = \begin{cases} y & \text{if } i = j \\ Y_j & \text{if } i = n+1 \\ Y_i & \text{otherwise} \end{cases}$$

The key insight is that conditional on $Z_{n+1} = z_j$ and E_{n+1}, the swapped configuration corresponds to a different label sequence but preserves the conditional independence structure. Following the same logic as in (A27), but with the swapped label sequence:

$$\begin{aligned}
&\mathbb{P}\left\{\mathcal{D}^{\text{cal}} = D^{\text{cal}},\, \{Z_i\}_{i\in\mathcal{D}_{\text{cal}}} = \{z_{n+1}\} \cup \{z_i\}_{i\in D^{\text{cal}}\setminus\{j\}} \,\middle|\, Z_{n+1} = z_j,\, E_{n+1}\right\} \\
=&\prod_{i\in D^{\text{cal}}} \pi\big(N(Y_i^{(j,y)}; Y_{1:n}^{(j,y)})\big) \prod_{i\in[n]\setminus D^{\text{cal}}} \big(1 - \pi(N(Y_i^{(j,y)}; Y_{1:n}^{(j,y)}))\big) \\
=&\pi\big(N(y; Y_{1:n}^{(j,y)})\big) \prod_{i\in D^{\text{cal}}, i\neq j} \pi\big(N(Y_i^{(j,y)}; Y_{1:n}^{(j,y)})\big) \prod_{i\in[n]\setminus D^{\text{cal}}} \big(1 - \pi(N(Y_i^{(j,y)}; Y_{1:n}^{(j,y)}))\big)
\end{aligned} \tag{A28}$$

where we have separated out the term for index j (which now has label y after swapping) in the first product.

The product form arises from the conditional independence of the Bernoulli inclusion indicators given the label sequence. This independence is preserved under the swapping operation. This motivates our definition of the probability weights:

$$p^{(j)}(y) = \pi\big(N(y; Y_{1:n}^{(j,y)})\big) \prod_{i \in \mathcal{D}^{\text{cal}}, i \neq j} \pi\big(N(Y_i^{(j,y)}; Y_{1:n}^{(j,y)})\big) \prod_{i \in \mathcal{D}^{\text{train}}} \big(1 - \pi(N(Y_i^{(j,y)}; Y_{1:n}^{(j,y)}))\big).$$

For completeness, when $j = n + 1$ (representing no swapping), this formula reduces to:

$$p^{(n+1)}(y) = \prod_{i \in \mathcal{D}^{\text{cal}}} \pi\big(N(Y_i; Y_{1:n})\big) \prod_{i \in \mathcal{D}^{\text{train}}} \big(1 - \pi(N(Y_i; Y_{1:n}))\big)$$

These probabilities $p^{(j)}(y)$ capture the probability of the specific sample splitting configuration under our selective sampling model after swapping the test point (X_{n+1}, Y_{n+1}) with the calibration point (X_j, Y_j).

The second term in (A26) is computed using the exchangeability of the data, which ensures that

$$\mathbb{P}\Big\{Z_{n+1} = z_j \,\Big|\, E_{n+1}\Big\} = \frac{1}{n+1}.$$

Observe that conditional on the event $E_{n+1} \wedge \{\mathcal{D}^{\text{train}} = D^{\text{train}}\} \wedge \{\{Z_i\}_{i \in \mathcal{D}^{\text{train}}} = \{z_i\}_{i \in D^{\text{train}}}\}$, the score function is completely determined by the training set. Consequently, S_{n+1} can only take values in the finite set $\{s((x_i, y_i) : i \in D^{\text{cal}} \cup \{n+1\}\}$. Moreover, since the denominator in Equation (A25) remains constant across all possible values of S_{n+1} (conditional on the same event), we can normalize the probabilities to obtain the conformalization weights:

$$w_j(y) = \frac{p^{(j)}(y)}{p^{(n+1)}(y) + \sum_{k \in \mathcal{D}^{\text{cal}}} p^{(k)}(y)}, \quad \text{for all } j \in \mathcal{D}^{\text{cal}} \cup \{n+1\}.$$

These weights account for the non-exchangeability introduced by our frequency-dependent sampling. By weighting each score according to its probability under the selective model, we restore the validity guarantee.

Combining the above calculations, the conditional distribution of S_{n+1} takes the form:

$$S_{n+1} \,\Big|\, \Big(E_{n+1} \wedge \{\mathcal{D}^{\text{train}} = D^{\text{train}}\} \wedge \{\{Z_i\}_{i \in \mathcal{D}^{\text{train}}} = \{z_i\}_{i \in D^{\text{train}}}\}\Big)$$
$$\sim \sum_{j \in \mathcal{D}_{\text{cal}}} w_j(y_{n+1})\, \delta_{s(x_j, y_j)} \;+\; w_{n+1}(y_{n+1})\, \delta_{s(x_{n+1}, y_{n+1})}.$$

This conditional distribution is exactly the weighted empirical distribution that appears in our quantile condition. After marginalizing, we obtain:

$$\mathbb{P}\left[S_{n+1} \;\leq\; \text{Quantile}\Big(1 - \alpha; \sum_{j \in \mathcal{D}^{\text{cal}}} w_j(Y_{n+1})\, \delta_{S_j} \;+\; w_{n+1}(Y_{n+1})\, \delta_{S_{n+1}}\Big)\right] \geq 1 - \alpha.$$

By our earlier quantile characterization, this is equivalent to

$$\mathbb{P}\left[Y_{n+1} \in \hat{C}_\alpha^\pi(X_{n+1}; \mathcal{Y}_{n+1})\right] \geq 1 - \alpha$$

and,

$$\mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha^\pi(X_{n+1}; \mathcal{Y}_{n+1})\right] \leq \alpha.$$

Finally, we can complete the proof by returning to our initial observation. Since $\hat{C}_\alpha^\pi(X_{n+1}; \mathcal{Y}_n) = \hat{C}_\alpha^\pi(X_{n+1}; \mathcal{Y}_{n+1})$ when $Y_{n+1} \in \mathcal{Y}_n$, we conclude:

$$\mathbb{P}\left[Y_{n+1} \notin \hat{C}_\alpha^\pi(X_{n+1}; \mathcal{Y}_n), Y_{n+1} \in \mathcal{Y}_n\right] \leq \alpha.$$

This completes the proof. □

Remark 1. The probability weights $p^{(j)}(y)$ can potentially equal zero when the probability function π takes boundary values of 0 or 1. This occurs when certain swapping configurations are infeasible.

Consider a concrete example: suppose $\pi(1) = 0$ and $\pi(2) = 1$. This means that any label appearing exactly once is deterministically assigned to the training set (since $I_i = 0$ with probability 1), while any label appearing exactly twice is deterministically assigned to the calibration set (since $I_i = 1$ with probability 1).

Now suppose we have a label y^* that appears exactly twice in $Y_{1:n}$, and both occurrences are in the calibration set (as required by $\pi(2) = 1$). If we attempt to swap the test point (with label $y' \neq y^*$) with one of these calibration points, y^* would appear only once in $Y_{1:n}^{(j,y')}$, requiring it to be in the training set with probability 1 due to $\pi(1) = 0$. This creates a contradiction, making $p^{(j)}(y') = 0$ for this configuration.

A4 Additional Numerical Results

Here we present additional implementation details related to Section 5, and further numerical results.

A4.1 Choice of Classification Model and Adjustments

Across all experiments, the base classifier is a k-nearest neighbors model with $k = 5$. We choose this comparatively simple model to reduce the risk of overfitting in our setting with many class labels and potentially extremely low per-class frequencies, where more complex models are easily prone to poor generalization. In our implementation, neighbor contributions to classification are weighted proportionally to the inverse of their Euclidean distance from the query point. For synthetic datasets, we employ the Minkowski distance metric, whereas for real image datasets, we embed image pixels into numerical representations and apply the cosine similarity metric, as it better reflects visual similarity in the embedding space.

Moreover, when implementing split conformal prediction in the open-set setting, where new classes may appear in the calibration dataset, it is necessary to assign predicted probabilities to calibration labels that are absent from the training set. In our experiments, we assign small and randomized probabilities to such unseen calibration labels according to

$$p_{\text{unseen}} = \frac{1 + n_{\text{singleton}}}{(1 + n_{\text{train}}) \cdot |C_{\text{unseen}}|} + \text{noise},$$

where $n_{\text{singleton}}$ denotes the number of labels that occur only once in the training set, and $|C_{\text{unseen}}|$ denotes the cardinality of the new calibration labels. This probability corresponds to the Good–Turing estimate of encountering an unseen label, evenly distributed across all possible unseen labels. This adjustment ensures that prediction sets can be constructed efficiently when applying the standard split conformal prediction with adaptive nonconformity score[37], both for establishing the baseline benchmark and for constructing $\hat{C}_{\alpha_{\text{class}}}(X_{n+1})$ in our conformal Good–Turing classification.

A4.2 Hyperparameter Settings

When implementing the conformal Good–Turing classification method, we use the feature-based conformal Good–Turing p-value ψ_0^{XGT} to test H_{unseen}, where the score function is derived by training a *local outlier factor* model with the `scikit-learn` package. For testing H_{seen}, we employ the randomized conformal Good–Turing p-value ψ_k^{RGT} for each hypothesis H_k, fixing the multiple testing parameter c_k at a constant value computed with $\beta = 1.6$ to obtain ψ_{seen}. We adopt the data-driven adaptive allocation strategy to distribute the total significance budget α across $(\alpha_{\text{class}}, \alpha_{\text{new}}, \alpha_{\text{old}})$. Specifically, we set $\alpha = 0.1$ for synthetic experiments and $\alpha = 0.2$ for CelebA experiments. We then perform a grid search over $\alpha_{\text{old}} \in \{0, 0.01, 0.02, 0.05\}$ and α_{class} starting from 0.01 with increments of 0.005. The allocation $(\alpha_{\text{class}}, \alpha_{\text{new}}, \alpha_{\text{old}})$ is chosen to minimize

the loss function in 16. In evaluating this loss, we fix $\lambda = 0.5$ and estimate the average size of conformal prediction sets using 10-fold cross-validation, where in each fold the training and calibration sets are randomly partitioned for fast computation.

A4.3 Additional Numerical Results

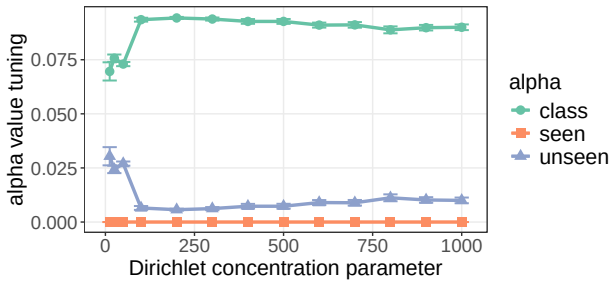

Figure A1: Allocation of significance levels $(\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}})$ when applying the proposed conformal Good–Turing classification method with a total budget $\alpha = 0.1$ on synthetic data generated from a Dirichlet process model, as in Figure 1. Error bars indicate 1.96 standard errors.

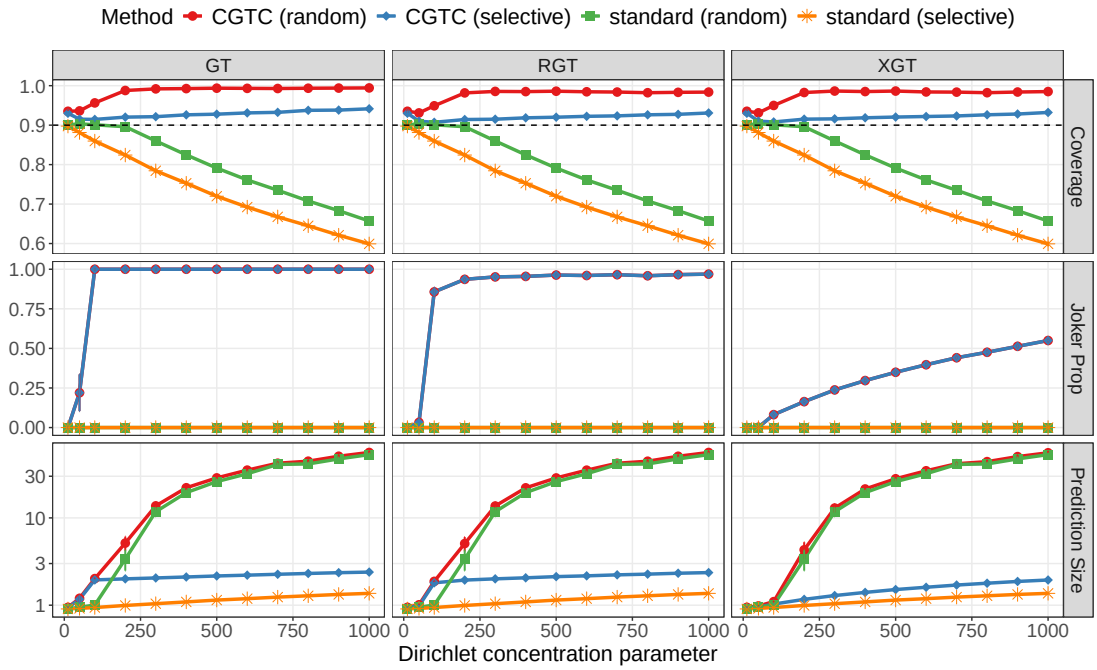

Figure A2: Performance of conformal prediction sets constructed using different methods on synthetic data generated from a Dirichlet process model, as in Figure 1. Each column corresponds to a distinct conformal Good–Turing p-value construction: deterministic feature blind (GT), randomized feature blind (RGT), and feature-based (XGT). All p-value constructions consistently achieve marginal coverage and therefore outperform the standard benchmark, while the feature-based approach minimizes usage of the joker.

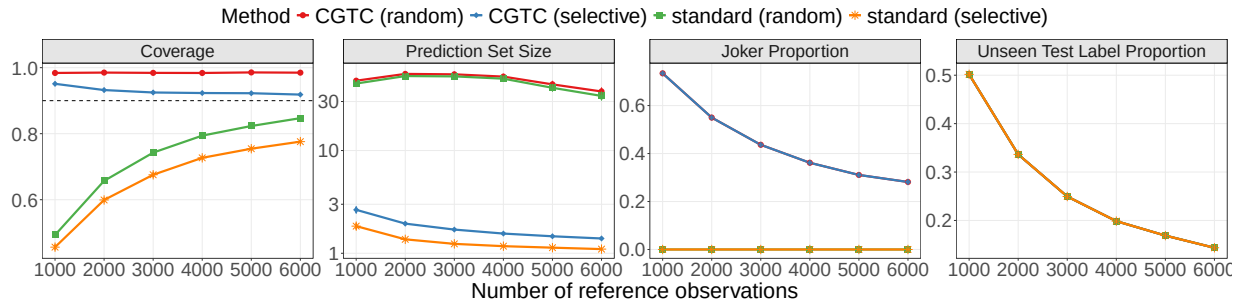

Figure A3: Performance of conformal prediction sets constructed using various methods on synthetic data generated from a Dirichlet process model, as in Figure 1. The results are shown as a function of the total number of samples in the training and calibration data, and the concentration parameter is fixed at $\theta = 1000$. Other details are as in Figure 1.

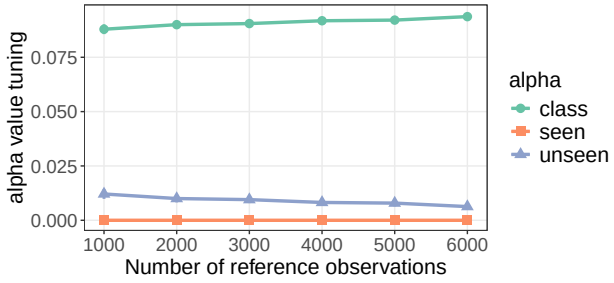

Figure A4: Allocation of significance levels $(\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}})$ for the proposed conformal Good–Turing classifier under a total budget $\alpha = 0.1$ on synthetic data generated from a Dirichlet process model with concentration $\theta = 1000$, shown as a function of the total number of samples in the training and calibration data. Other details are as in Figure 1.

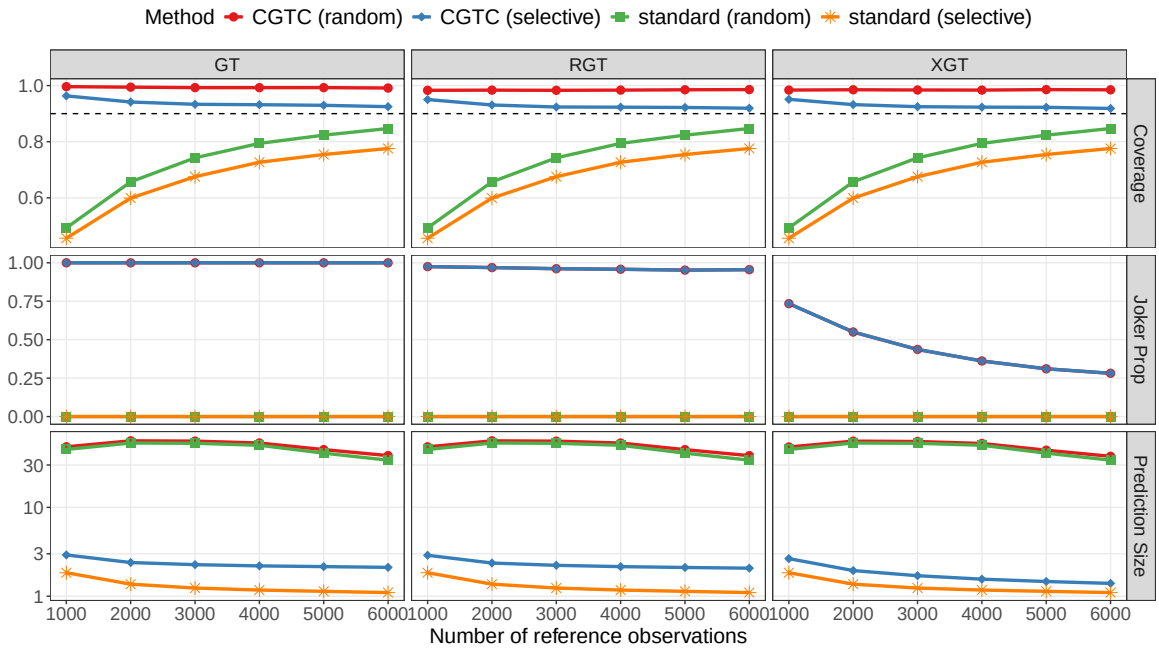

Figure A5: Performance of conformal prediction sets constructed using different methods on synthetic data generated from a Dirichlet process model with concentration $\theta = 1000$, as a function of the total number of samples in the training and calibration data. Each column represents a distinct conformal Good–Turing p-value construction: deterministic feature blind (GT), randomized feature blind (RGT), and feature-based (XGT). Other details are as in Figure 1.

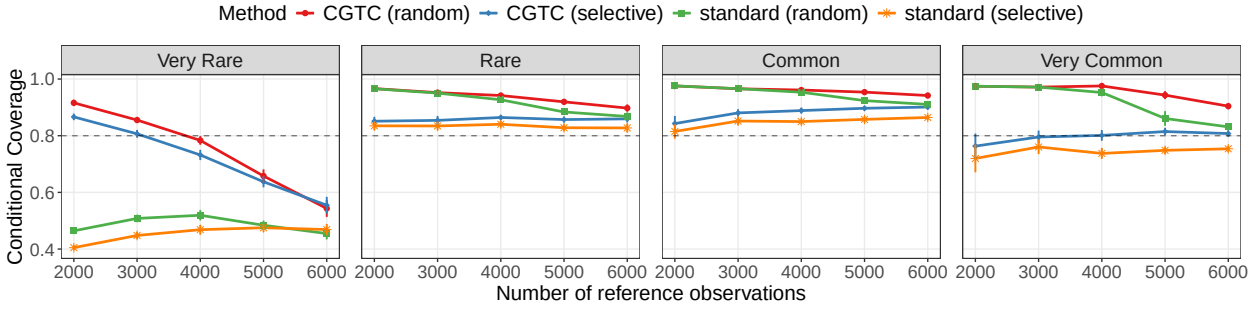

Figure A6: Coverage of conformal prediction sets constructed using different methods on face recognition data from the CelebA resource, as in Figure 4, stratified by test-label frequency. The standard approach fails to achieve coverage for very rare labels, whereas the conformal Good–Turing method attains higher conditional coverage on very rare labels. Other details are as in Figure 4.

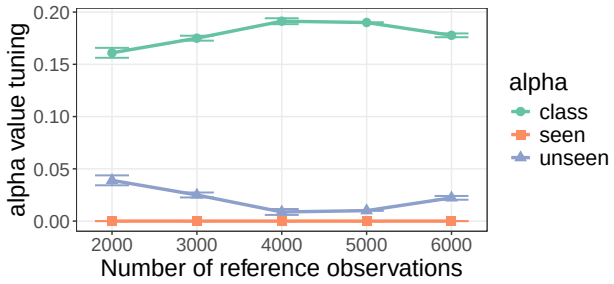

Figure A7: Allocation of significance levels $(\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}})$ for the proposed conformal Good–Turing classifier under a total budget $\alpha = 0.2$ on face recognition data from the CelebA resource, as in Figure 4. Error bars indicate 1.96 standard errors.

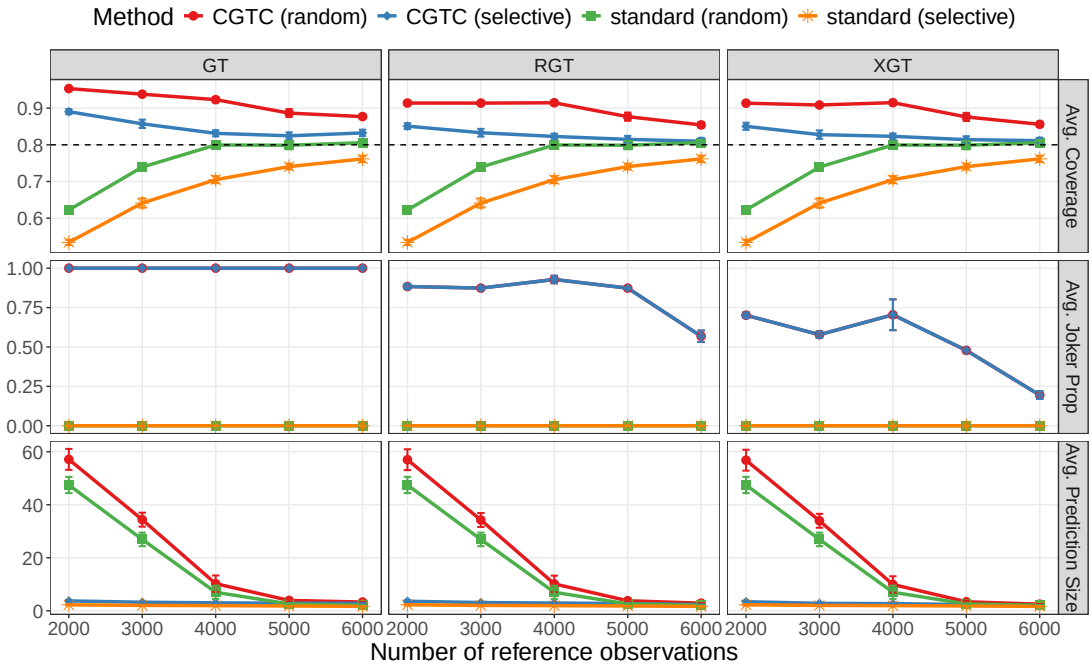

Figure A8: Performance of conformal prediction sets constructed using different methods on face recognition data from the CelebA resource, as in Figure 4. Each column corresponds to a distinct conformal Good–Turing p-value construction: deterministic feature blind (GT), randomized feature blind (RGT), and feature-based (XGT).

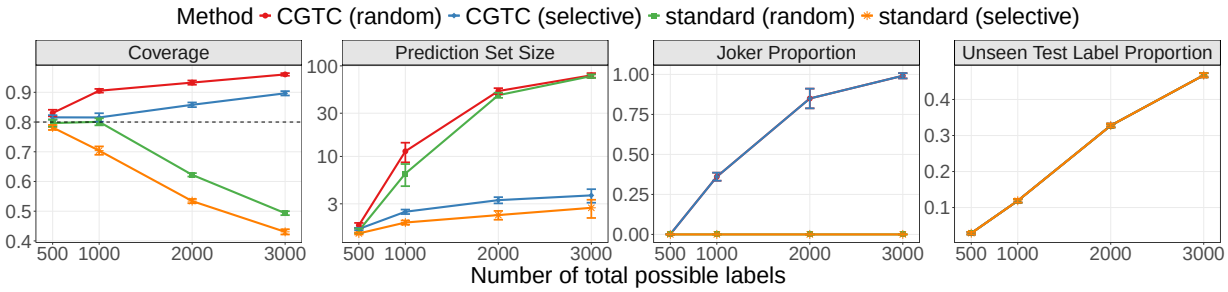

Figure A9: Performance of conformal prediction sets constructed using various methods on face recognition data from the CelebA resource, as in Figure 4. Results are shown as a function of the total number of possible labels, with the training and calibration sample sizes fixed at 2000. Other details are as in Figure 4.

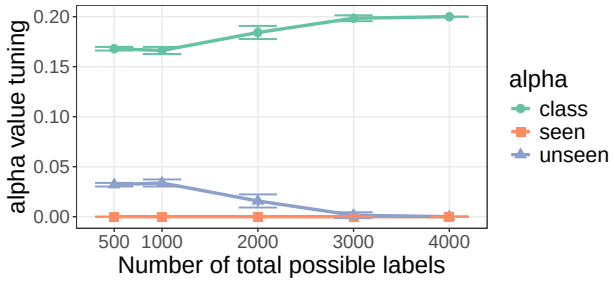

Figure A10: Allocation of significance levels ($\alpha_{\text{class}}, \alpha_{\text{unseen}}, \alpha_{\text{seen}}$) for the proposed conformal Good–Turing classifier under a total budget $\alpha = 0.2$ on face recognition data from the CelebA resource, shown as a function of the total number of possible labels, with the training and calibration sample sizes fixed at 2000. Other details are as in Figure 4.

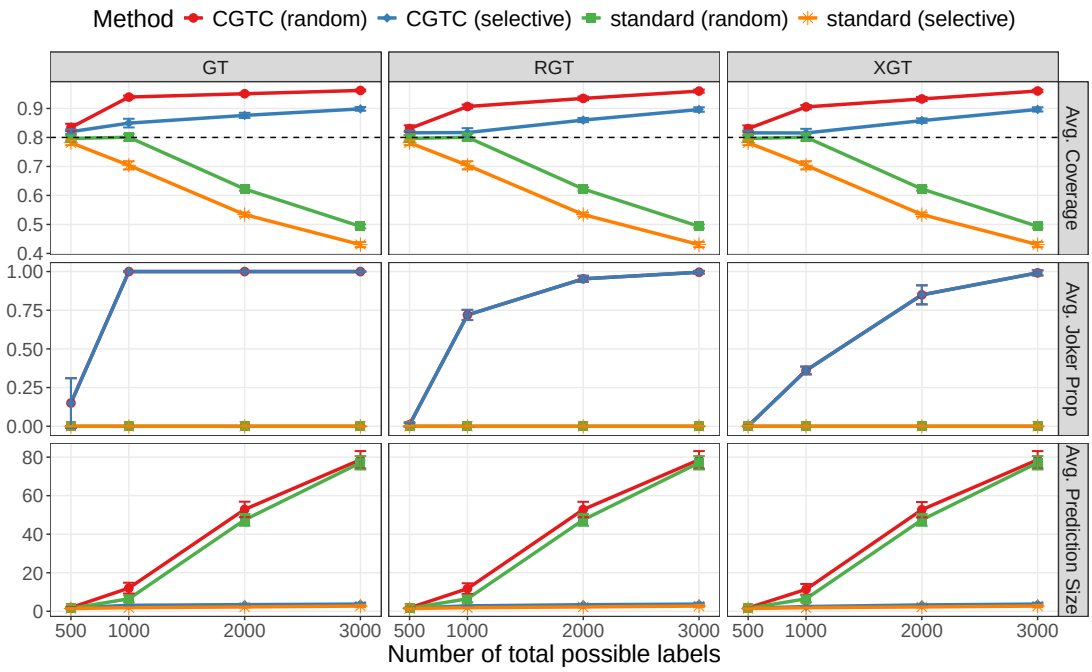

Figure A11: Performance of conformal prediction sets constructed using different methods on face recognition data from the CelebA resource, as a function of the total number of possible labels, with the training and calibration sample sizes fixed at 2000. Each column represents a distinct conformal Good–Turing p-value construction: deterministic feature blind (GT), randomized feature blind (RGT), and feature-based (XGT). Other details are as in Figure 4.