Distribution-free uncertainty quantification for classification under label shift

Aleksandr Podkopaev[1] **Aaditya Ramdas**[1]

[1]Department of Statistics & Data Science, Machine Learning Department, Carnegie Mellon University

Abstract

Trustworthy deployment of ML models requires a proper measure of uncertainty, especially in safety-critical applications. We focus on uncertainty quantification (UQ) for classification problems via two avenues — prediction sets using conformal prediction and calibration of probabilistic predictors by post-hoc binning — since these possess distribution-free guarantees for i.i.d. data. Two common ways of generalizing beyond the i.i.d. setting include handling *covariate* and *label* shift. Within the context of distribution-free UQ, the former has already received attention, but not the latter. It is known that label shift hurts prediction, and we first argue that it also hurts UQ, by showing degradation in coverage and calibration. Piggybacking on recent progress in addressing label shift (for better prediction), we examine the right way to achieve UQ by reweighting the aforementioned conformal and calibration procedures whenever some unlabeled data from the target distribution is available. We examine these techniques theoretically in a distribution-free framework and demonstrate their excellent practical performance.

1 INTRODUCTION

It is common in classification to assume access to labeled data $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y} = \{1, \ldots, K\}$ denote the covariates, or features, and the labels respectively, and the pairs (X_i, Y_i), $i = 1, \ldots, n$ are sampled i.i.d. from some unknown joint distribution P over $\mathcal{X} \times \mathcal{Y}$. Such dataset is used to learn a predictor f, a mapping from $\mathcal{X}$ to rankings or distributions over $\mathcal{Y}$, by optimizing some loss/risk. However, accurate point prediction alone can be insufficient in certain applications, e.g., medical diagnosis, where trustworthy deployment of a model requires a valid measure of uncertainty associated with corresponding predictions.

Common prediction models are mappings of the form $f : \mathcal{X} \to \Delta_K$, where Δ_K refers to the probability simplex in $\mathbb{R}^K$, and a prediction on a new (test) point $X \in \mathcal{X}$ is performed by picking the top-ranked class according to $f(X)$. One hopes that the output vector $f(X)$ reflects the true conditional probabilities of classes given the observed input, but this won't be true without additional distributional and modeling assumptions, that are typically strong and unverifiable in practice. In this work, we focus on two categories of post-processing procedures — calibration via post-hoc binning and conformal prediction — that use held-out data (referred to as *calibration* dataset) and a trained model to construct a corresponding *wrapper* that provably quantifies predictive uncertainty when no distributional assumptions are made about the data generating mechanism. (This generality comes at a certain price which we discuss further.)

We work in the context of *distribution-free* uncertainty quantification and, in particular, focus on producing prediction sets (Section 2) and calibrated probabilities (Section 3), which are complementary approaches for classifier UQ. While the former aims to produce a set of labels that contains the truth with high probability, the latter aims to amend the output of a probabilistic predictor so that it has a rigorous frequentist interpretation. It is useful to view the task through the lens of how actionable the corresponding notion is in a given setup. For example, in a binary classification setup with only 4 possible prediction sets $\{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$, if we were to observe prediction sets $\{1, 2\}$ for large fraction of data points, one might end up quite disappointed. Thus, calibration could be a better way of quantifying uncertainty in the binary case. However, mathematical guarantees on calibration degrade with growing number of classes, but the aforementioned prediction sets become an attractive option with more labels. To summarize, neither of two notions provide a complete answer to the question of UQ for classification on their own, but together they represent two of the more principled distribution-free approaches towards UQ

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

that are practically efficient and theoretically grounded.

In real-world applications, the *target* distribution (generating test data) might not be the same as the *source* distribution (generating training data) which can both hurt a model's generalization and lead to violation of the assumptions under which even assumption-lean UQ is valid. As meaningful reasoning about uncertainty on the target domain is hopeless without any additional information about the type of distribution shift, one may hope that it is possible to make simplifying assumptions which would allow us to perform appropriate corrections and construct procedures with non-trivial guarantees. Let P, Q stand for the source and target distributions defined on $\mathcal{X} \times \mathcal{Y}$, with p, q being the PDFs or PMFs associated with P and Q respectively. Two common assumptions about the type of shift include *covariate shift* [Shimodaira, 2000]: $q(x) \neq p(x)$ but $q(y \mid x) = p(y \mid x)$, and *label shift* [Saerens et al., 2002]: $q(y) \neq p(y)$ but $q(x \mid y) = p(x \mid y)$. Both assumptions allow for a tractable interpretation when viewing the data generating process as a causal or anti-causal model respectively. For example, label shift is a reasonable assumption in medical applications where diseases (Y) cause symptoms (X): it is intuitive that some sort of correction might be required when a predictor trained in ordinary conditions is deployed during extreme ones, e.g., during a pandemic.

Classic approaches for handling the aforementioned shifts make an assumption that the target support is contained in the source support, so that the covariate or label likelihood ratios (or *importance weights*) $q(x)/p(x)$ or $q(y)/p(y)$ are well-defined. In applications, true weights are never known exactly, so the construction of consistent estimators has received a lot of attention in the ML community. For label shift dominant approaches that are still computationally feasible in modern high-dimensional regimes, and that perform estimation using labeled data only from the source distribution, include: (a) Black Box Shift Estimation (BBSE) [Lipton et al., 2018] and related Regularized Learning under Label Shift (RLLS) [Azizzadenesheli et al., 2019], (b) Maximum Likelihood Label Shift (MLLS) and its variants [Saerens et al., 2002, Alexandari et al., 2020].

Within the context of distribution-free UQ, covariate shift has recently received attention. Focusing on regression, Tibshirani et al. [2019] generalize construction of conformal prediction intervals to handle the case of known covariate likelihood ratio, and empirically demonstrate that the modified procedure works reasonably well with a plug-in estimator for the importance weights. For binary classification, Gupta et al. [2020] propose a way of calibrating probabilistic predictors under covariate shift, and quantify miscalibration of the resulting estimator.

In this work, we close an existing gap for quantifying predictive uncertainty under label shift. Building on recent results about distribution-free calibration and (split-)conformal prediction, we adapt both to handling label shift through an appropriate form of reweighting. While typical application of those frameworks requires labeled data from the target to provide guarantees, we show that under reasonable assumptions one can still reason about uncertainty on the target even if only unlabeled data is available. In contrast to covariate shift where we observe X and need the covariate likelihood ratio of X to reweight, under label shift we observe X but need the likelihood ratio of Y to reweight. We also consider an alternative way of addressing label shift by performing label-conditional conformal classification [Vovk et al., 2005, 2016, Sadinle et al., 2019, Guan and Tibshirani, 2019].

2 CONFORMAL CLASSIFICATION

We begin with the notion of prediction sets as a way of quantifying predictive uncertainty. Formally, we wish to construct an uncertainty set function $C : \mathcal{X} \to 2^{\mathcal{Y}}$, such that for a new (test) data point we can guarantee that:

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

Conformal prediction [Vovk et al., 2005] has received attention recently both in regression [Lei et al., 2018, Romano et al., 2019, Barber et al., 2021] and classification [Cauchois et al., 2020, Romano et al., 2020, Angelopoulos et al., 2021] settings. It does not require making any distributional assumptions, which comes at the price of provably providing only *marginal* guarantees as stated in (1) which should be contrasted with possibly the ultimate goal of obtaining prediction sets with guarantees conditional on a given input.

Since conditional guarantees often require making restrictive and unverifiable assumptions, we instead focus on procedures that might provably provide marginal coverage guarantees but still tend to demonstrate good conditional coverage empirically. Being flexible, conformal prediction allows to proceed with both probabilistic and scoring classifiers. Within this framework, one usually defines a non-conformity score, a higher value of which on a given data point indicates that it is more 'atypical'. For example, even if a classifier outputs only the ranking of predicted classes, a rank of the true class defines a valid non-conformity score. Keeping in mind that our techniques extend to other types of classifiers, we nevertheless focus on probabilistic predictors in this work which are also dominant in modern machine learning.

2.1 EXCHANGEABLE CONFORMAL

Consider a sequence of candidate nested prediction sets $\{\mathcal{F}_\tau(x)\}_{\tau \in \mathcal{T}}$: $\mathcal{F}_{\tau_1}(x) \subseteq \mathcal{F}_{\tau_2}(x) \subseteq \mathcal{Y}$ for any $\tau_1 \leq \tau_2 \in \mathcal{T}$, with $\mathcal{F}_{\inf \mathcal{T}} = \emptyset$ and $\mathcal{F}_{\sup \mathcal{T}} = \mathcal{Y}$ [Gupta et al., 2019]. For any point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ define

$$r(x, y) := \inf \left\{\tau \in \mathcal{T} : y \in \mathcal{F}_\tau(x)\right\}, \tag{2}$$

as the smallest radius of the set in a sequence $\{\mathcal{F}_\tau(x)\}_{\tau\in\mathcal{T}}$ that captures y. Within split-conformal framework, available dataset is split at random into two parts: the first is used to construct a nested sequence and the second is used to select the smallest $\tau^\star$ that guarantees validity.

If the true class-posterior distribution $\pi_y(x) = \mathbb{P}[Y = y \mid X = x]$ is known, the optimal prediction set for any $x \in \mathcal{X}$ with conditional coverage guarantee is based on the corresponding density level sets [Vovk et al., 2005, Lei et al., 2013, Gupta et al., 2019, Sadinle et al., 2019]: one should pick the largest $\tau_\alpha(x)$ and include all labels with probabilities $\pi_y(x)$ exceeding $\tau_\alpha(x)$ so that the corresponding total probability mass is at least $1-\alpha$. When ties are present, such procedure can yield conservative sets, e.g., if for some $x \in \mathcal{X}$ all classes are equally probable in a 10-class problem, then $\tau_\alpha(x) = 0.1$ and the proposed set would simply be $\mathcal{Y}$. For the discussion that follows we assume that there are no ties or that they are broken as formally discussed in Appendix B.1. Then, to construct the optimal prediction set, one should start with an empty one and keep including labels as long as the total probability mass of labels included before is less than $1-\alpha$. Formally,

$$C_\alpha^{\text{oracle}}(x) := \{y \in \mathcal{Y} : \rho_y(x;\pi) < 1-\alpha\}, \quad (3)$$

$$\text{where} \quad \rho_y(x;\pi) := \sum_{y'=1}^{K} \pi_{y'}(x)\mathbb{1}\{\pi_{y'}(x) > \pi_y(x)\}$$

is the total probability mass of labels that are more likely than $y \in \mathcal{Y}$. Notice that for any $x \in \mathcal{X}$ and the corresponding most likely label $y^\star$ it holds that $\rho_{y^\star}(x;\pi) = 0$. When an estimator $\widehat{\pi}$ of the true conditional distribution is used, split-conformal framework provides a way of updating the threshold $1-\alpha$ in (3) in order to retain coverage guarantees. However, naive conformalization of the nested sequence suggested by the form (3) yields prediction sets with correct marginal coverage but typically inferior conditional coverage in practice. Due to that reason and a desire of consistency, i.e., recovering the oracle prediction sets from the conformal ones in the limit, we instead use a randomized version of (3) defined as

$$\widetilde{C}_\alpha^{\text{oracle}}(x) = \{y : \rho_y(x;\pi) + u \cdot \pi_y(x) \le 1-\alpha\}, \quad (4)$$

where u is a realization of $\text{Unif}([0,1])$, sampled independently of anything else [Vovk et al., 2005, Romano et al., 2020]. Note that replacing strict inequality by a non-strict does not expand the prediction set as equality happens with zero probability and that induced randomization can result in exclusion only of a single label from the set $C_\alpha^{\text{oracle}}(x)$. The form of the optimal prediction sets (4) suggests to consider the following nested sequence:

$$\mathcal{F}_\tau(x,u;\widehat{\pi}) = \{y \in \mathcal{Y} : \rho_y(x;\widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \le \tau\}, \quad (5)$$

for $\tau \in \mathcal{T} = [0,1]$. Then for any triple (X,Y,U) the corresponding radius (2), or score, is given by

$$r(X,Y,U;\widehat{\pi}) = \inf\{\tau \in \mathcal{T} : \rho_Y(X;\widehat{\pi}) + U \cdot \widehat{\pi}_Y(X) \le \tau\}$$
$$= \rho_Y(X;\widehat{\pi}) + U \cdot \widehat{\pi}_Y(X). \quad (6)$$

Adapting to label shift can be performed with other non-conformity scores proposed recently for conformal classification [Cauchois et al., 2020, Angelopoulos et al., 2021], and we further discuss the subtleties behind our choice in Appendix B.2. Assume that the dataset is split at random into two parts: training $\{(X_i,Y_i)\}_{i\in\mathcal{I}_1}$ and calibration $\{(X_i,Y_i)\}_{i\in\mathcal{I}_2}$, where for simplicity the calibration data points are indexed as $\mathcal{I}_2 = \{1,\dots,n\}$. When the data are exchangeable, the non-conformity scores $r_i = r(X_i,Y_i,U_i;\widehat{\pi}) \in [0,1], i \in \mathcal{I}_2 \cup \{n+1\}$ are exchangeable as well, which in turn implies that the prediction set

$$\mathcal{F}_{\tau^\star}(x,u;\widehat{\pi}) = \{y \in \mathcal{Y} : \rho_y(x;\widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \le \tau^\star\},$$
$$\tau^\star = Q_{1-\alpha}\left(\{r_i\}_{i\in\mathcal{I}_2} \cup \{1\}\right), \quad (7)$$

does attain the right coverage guarantee[1]. This is a classic result in conformal prediction and represents a simple fact about quantiles of exchangeable random variables, stated next for completeness.

Theorem 1. *If $\{(X_i,Y_i)\}_{i=1}^{n+1}$ are exchangeable, then:*

$$\mathbb{P}(Y_{n+1} \in \mathcal{F}_{\tau^\star}(X_{n+1},U_{n+1};\widehat{\pi}) \mid \{(X_i,Y_i)\}_{i\in\mathcal{I}_1}) \ge 1-\alpha.$$

Further, if the non-conformity scores are almost surely distinct, then the above probability is upper bounded by $1-\alpha+1/(n+1)$.

The proof is given in Appendix B.3. Notice that the randomized sequence (5) might yield empty, and thus non-actionable prediction sets, which is the consequence of deploying randomization only. Substituting the condition in (7) with $\mathbb{1}\{\rho_y(x;\widehat{\pi}) > 0\} \cdot (\rho_y(x;\widehat{\pi}) + u \cdot \widehat{\pi}_y(x))$ ensures that the prediction set always includes the most likely label. Such a construction trivially inherits the coverage guarantee stated in Theorem 1, and we refer the reader to Appendix B.2 for further details.

2.2 LABEL-SHIFTED CONFORMAL

To illustrate the necessity of accounting for label shift we consider the following toy classification task with 3 classes $\mathcal{Y} = \{1,2,3\}$ where class proportions are given as $p = (0.1, 0.6, 0.3)$ and $q = (0.3, 0.2, 0.5)$, and for each data point the covariates are sampled according to $X \mid Y = y \sim \mathcal{N}(\mu_y, \Sigma)$ where $\mu_1 = (-2;0)^\top$, $\mu_2 = (2;0)^\top$, $\mu_3 = \left(0; 2\sqrt{3}\right)^\top$, $\Sigma = \text{diag}(4,4)$. First, we perform the standard routine for constructing split-conformal prediction sets for a

[1] $Q_\beta(F) := \inf\{z : F(z) \ge \beta\}$ is β-quantile of a distribution F. For a multiset $\{z_1,\dots,z_m\}$ we write $Q_\beta(\{z_1,\dots,z_m\}) := Q_\beta\left(\frac{1}{m}\sum_{i=1}^m \delta_{z_i}\right)$, where δ_a is a point-mass distribution at a, to denote quantiles of the corresponding empirical distribution.

single draw of data from the source and target distributions using the Bayes-optimal rule as an underlying predictor. We illustrate a single draw of the test data on Figure 1a and the resulting prediction sets on Figure 1b. Next, we repeat the simulation 1000 times and track empirical coverage on the test set. Results on Figure 1c demonstrate the necessity of correcting for label shift as the classic conformal prediction sets introduced in Section 2.1 fail to achieve the correct marginal coverage.

Assume that the true likelihood ratios $w(y) = q(y)/p(y)$ are known for all $y \in \mathcal{Y}$. In order to obtain provably valid prediction sets, we consider instead:

$$\mathcal{F}^{(w)}_{\tau^\star}(x, u; \widehat{\pi}) = \{y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leq \tau^\star_w(y)\},$$
$$\tau^\star_w(y) = Q_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{p}^w_i(y)\delta_{r_i} + \tilde{p}^w_{n+1}(y)\delta_1\right), \tag{8}$$

$$\text{where} \quad \tilde{p}^w_i(y) = \frac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(y)}, \quad i = 1, \ldots, n,$$
$$\tilde{p}^w_{n+1}(y) = \frac{w(y)}{\sum_{j=1}^n w(Y_j) + w(y)}. \tag{9}$$

In addition to the fact that the empirical distribution used to calibrate the threshold in (8) is different from the one used in exchangeable setting (7), notice that the thresholds themselves now vary depending on the class label. The formal guarantee for the prediction set (8) is stated next.

Theorem 2. *For any $\alpha \in (0, 1)$, if the true likelihood ratios $w(y) = q(y)/p(y)$ are known for all $y \in \mathcal{Y}$, it holds that*

$$\mathbb{P}(Y_{n+1} \in \mathcal{F}^{(w)}_{\tau^\star}(X_{n+1}, U_{n+1}; \widehat{\pi}) \,|\, \{(X_i, Y_i)\}_{i \in \mathcal{I}_1}) \geq 1 - \alpha.$$

The proof is given in Appendix B.3. It relies on the concept of *weighted exchangeability* introduced by Tibshirani et al. [2019] to handle covariate shift in regression, and we adapt those ideas here to correct for label shift in classification. Returning to the example considered in the beginning of this section, Figure 1c illustrates that calibrating the threshold τ as in (8) with either oracle or estimated importance weights allows to achieve the target marginal coverage. Here we use BBSE [Lipton et al., 2018] to estimate the importance weights; more details are provided in Appendix A.

Next, we perform a similar experiment with the `wine quality` dataset [Cortez et al., 2009]. We refer the reader to Appendix B.4 for details regarding data pre-processing and modeling steps. The source and target class proportions are taken to be $p = (0.1, 0.4, 0.5)$ and $q = (0.4, 0.5, 0.1)$ and the data are resampled accordingly. Using a shallow multilayer perceptron as an underlying predictor and BBSE for importance weights estimation, at each iteration we repeat the routine for random splits of the original dataset and compare empirical coverage for different conformal prediction sets. Marginal coverage results given in Figure 1d support the idea that both shift-corrected conformal prediction sets demonstrate superior coverage performance compared with uncorrected ones. While conformal sets with oracle importance weights closely match the nominal coverage level, sets that proceed with estimated ones have a slightly downgraded performance. Arising basically due to an imperfect classification model and an imperfect importance weight estimation procedure, it highlights an important issue we discuss next.

While (weighted) exchangeability arguments yield a coverage guarantee in case of known importance weights, in practice one only has access to a corresponding estimator. Dominant methods, which we briefly touch upon in Appendix A, estimate importance weights using a separate labeled dataset from the source distribution and unlabeled dataset from the target. Under reasonable assumptions, such as identifiability and boundedness of the true importance weights, these estimators are known to be consistent as the size of both samples grows. For succinctness, we write $k = |\mathcal{D}_{\text{est}}|$ to denote the *total* size of the datasets used for constructing an estimator $\widehat{w}_k$ of the importance weights w.

Corollary 1. *Fix $\alpha \in (0, 1)$. Assume that $\widehat{w}_k$ is a consistent estimator of w. Further, assume that for the true w and all $y \in \mathcal{Y}$, the discrete distribution in* (8) *does not have a jump at level $1 - \alpha$. Then:*

$$\lim_{k \to \infty} \mathbb{P}\left(Y_{n+1} \in \mathcal{F}^{(\widehat{w}_k)}_{\tau^\star}(X_{n+1}, U_{n+1}; \widehat{\pi})\right) \geq 1 - \alpha.$$

The proof is given in Appendix B.3. To demonstrate why presence of a jump might cause problems, consider a simplified example. Let $Z \sim \text{Ber}(p)$ for which the quantile corresponding to any given level α is given by

$$Q_\alpha\left((1-p) \cdot \delta_0 + p \cdot \delta_1\right) = \mathbb{1}\{p > 1 - \alpha\},$$

Assume that we are given a sample of coin tosses $Z_1, \ldots, Z_n$ with the same bias parameter p. Even though the sample average $\overline{Z}_n$ is a consistent estimator of p, it nonetheless does not imply that the corresponding plug-in quantile estimator is consistent as the continuous mapping theorem cannot be invoked due to a discontinuity at $p = 1 - \alpha$. Indeed, let

$$\widehat{q}_n := Q_\alpha\left(\left(1 - \overline{Z}_n\right) \cdot \delta_0 + \overline{Z}_n \cdot \delta_1\right) = \mathbb{1}\left\{\overline{Z}_n > 1 - \alpha\right\},$$

and observe that $\widehat{q}_n \sim \text{Ber}\left(\mathbb{P}\left(\overline{Z}_n > 1 - \alpha\right)\right)$. Then by the normal approximation it follows that:

$$\mathbb{P}\left(\overline{Z}_n > 1 - \alpha\right) \approx 1 - \Phi\left(\sqrt{n}\frac{(1-\alpha) - p}{\sqrt{p(1-p)}}\right).$$

If $p > 1 - \alpha$, we can conclude that $\widehat{q}_n$ converges in probability to 1, and thus the estimator is consistent (similarly for $p < 1 - \alpha$). In case of equality, $\widehat{q}_n$ converges to $\text{Ber}(1/2)$, and thus the estimator will not be consistent. Still, for a

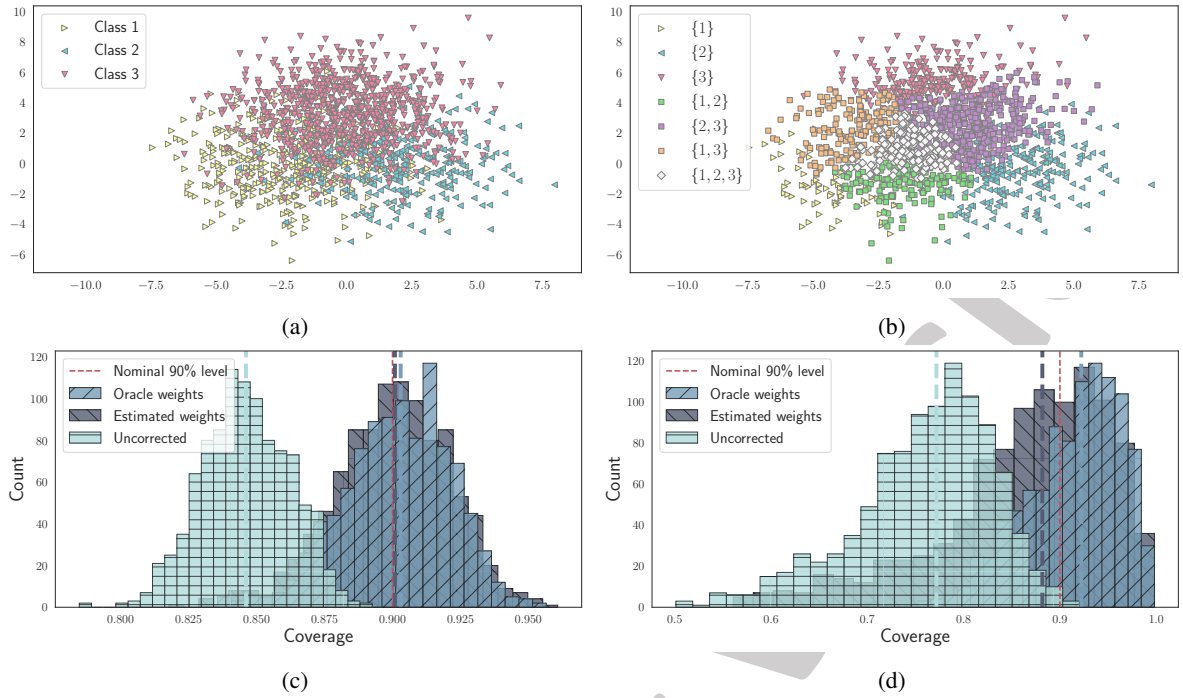


Figure 1: (a) Test data sample for the toy simulation in Section 2.2. (b) Corresponding conformal prediction sets when label shift is accounted for with oracle importance weights. (c) Empirical coverage on shifted data for the toy simulation in Section 2.2. (d): Empirical coverage on the `wine quality` dataset. Dashed vertical lines describe the median coverage values, which are significantly worse when label shift is not accounted for, while using estimated weights mimics the oracle reasonably well.

more general setting of the distribution defined in (8) it is reasonable to expect the assumption regarding absence of jumps to be satisfied as also confirmed by our conducted empirical study.

Label-conditional conformal prediction. Observing multiple points sharing the same label in a dataset makes it possible to apply the split-conformal framework in a way that makes the resulting prediction sets inherently robust to label shift [Vovk et al., 2005, 2016, Sadinle et al., 2019, Guan and Tibshirani, 2019]. Assume that a set of significance levels for each class $\{\alpha_y\}_{y\in\mathcal{Y}}$ has been chosen (e.g., $\alpha_y = \alpha$ for all y). By further splitting the calibration set $\mathcal{I}_2$ into $|\mathcal{Y}| = K$ groups depending on the corresponding labels, $\mathcal{I}_{2,y} := \{i \in \mathcal{I}_2 : Y_i = y\}$, one can consider prediction sets of the following form:

$$\mathcal{F}^c_{\tau^\star_c}(x, u; \widehat{\pi}) = \{y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \le \tau^\star_c(y)\},$$
$$\tau^\star_c(y) = Q_{1-\alpha_y}\left(\{r_i\}_{i\in\mathcal{I}_{2,y}} \cup \{1\}\right). \tag{10}$$

In other words, we separately apply split-conformal prediction framework for each label; this is like performing a separate hypothesis test for each label to determine whether there is sufficient evidence to exclude the label from the prediction set. To elaborate, the label shift assumption states that conditional distribution of X given $Y = y$ for all $y \in \mathcal{Y}$ does not change between source and target distributions. Thus for a test point (X_{n+1}, Y_{n+1}) the corresponding non-conformity score $r(X_{n+1}, Y_{n+1}, U_{n+1}; \widehat{\pi})$ together with $\{r_i\}_{i\in\mathcal{I}_{2,Y_{n+1}}}$ forms a collection of exchangeable random variables, which implies label-conditional validity, that is:

$$\mathbb{P}\left(Y_{n+1} \notin \mathcal{F}^c_{\tau^\star_c}(X_{n+1}, U_{n+1}; \widehat{\pi}) \mid Y_{n+1} = y\right) \le \alpha_y,$$

for all $y \in \mathcal{Y}$. When $\alpha_y = \alpha$ for all y, one can marginalize over y using *any* distribution (shifted or not), to yield $\mathbb{P}\left(Y_{n+1} \notin \mathcal{F}^c_{\tau^\star_c}(X_{n+1}, U_{n+1}; \widehat{\pi})\right) \le \alpha$. Thus, the label-conditional conformal framework yields a stronger guarantee than the standard (marginal) conformal and, it is automatically robust to changes in class proportions, retaining validity under label shift. The price to pay for the stronger conditional guarantee is larger prediction sets: for example, when the classes are not well-separated, label-conditional conformal can be expected to yield larger prediction sets; see Appendix B.5 for a careful empirical study. It should also be noted that the label-conditional conformal framework requires splitting available calibration data into K parts that could result in large losses of statistical efficiency when the number of classes K is large. On the other hand, such construction allows to tackle label shift in a way that does not require importance weights estimation, and thus get exact finite-sample guarantee instead of asymptotic one es-

tablished in Corollary 1. Thus, we view the label-conditional conformal framework as a complementary approach, perhaps worth utilizing when the amount of calibration data is larger relative to the number of labels.

3 CALIBRATION

While prediction sets describe a construction on top of the output of a predictor, calibration quantifies whether the output itself admits a rigorous frequentist interpretation. In contrast to the binary setting where there is usually no confusion about a definition of a calibrated predictor, there is one in the multiclass setting. First, we state a definition of a canonically calibrated predictor.

Definition 1 (Calibration)**.** A probabilistic predictor $f : \mathcal{X} \to \Delta_K$ is said to be calibrated if

$$\mathbb{P}\left(Y = y \mid f(X)\right) = f_y(X), \quad y \in \mathcal{Y},$$

where $f_y(x)$ denotes the y-th coordinate of $f(x)$.

Observe that canonical calibration requires the whole output vector to reflect the true conditional probabilities. Two extreme examples of canonically calibrated predictors include: (a) f^{Marg}: $f_y^{\text{Marg}}(x) = p(y)$, (b) f^{Bayes}: $f_y^{\text{Bayes}}(x) = \pi_y(x)$. In words, the former predictor outputs marginal probabilities of classes and the latter outputs the true class-posterior probabilities. In terms of classification efficiency, however, the first one is useless, while the second minimizes the classification risk, or the probability of incorrectly classifying a new point. Minimizing classification risk with respect to zero-one loss is computationally infeasible, and thus one refers instead to minimizing so-called *surrogate* losses, e.g., cross-entropy loss, with possibly added regularization terms. As a result, one obtains prediction models that are not calibrated out-of-the-box without making strong distributional and modeling assumptions, and thus aims to achieve it by performing post-processing using held-out data. While this topic has attracted a lot attention from practitioners recently, less results have been established on the theoretical side providing formal guarantees for common procedures that target improving model's calibration. Recognized approaches include Platt scaling [Platt, 1999], temperature scaling [Guo et al., 2017], histogram binning [Zadrozny and Elkan, 2001], isotonic regression [Zadrozny and Elkan, 2002] and others.

Model miscalibration is usually assessed using either reliability curves or related one-dimensional summary statistics. It is known that popular metrics, such as Expected Calibration Error (ECE), are not reliable since plug-in estimates can be biased if binning, or discretization, of the output of the resulting model is not performed [Kumar et al., 2019, Vaicenavicius et al., 2019]. Gupta et al. [2020] establish the necessity of binning for obtaining distribution-free calibration guarantees in a binary classification setup. Binning represents coarsening of the sample space and is defined as the partitioning of the probability simplex into non-overlapping bins: $\Delta_K = B_1 \cup \cdots \cup B_M$, $B_i \cap B_j = \emptyset$, $i \neq j$. Then a predictor f induces a partition of the sample space:

$$\mathcal{X}_m := \{x \in \mathcal{X} : f(x) \in B_m\}, \quad m \in \mathcal{M} := \{1, \ldots, M\}.$$

Since provable guarantees for canonical calibration require binning of the probability simplex, it is clear that the task becomes prohibitive with growing number of classes as each bin has to be supplied with sufficiently many data points during the calibration step for the resulting guarantees to be meaningful. One solution is given by either referring to other notions of UQ, such as the aforementioned prediction sets, or by relaxing the notion of calibration in the multiclass setting. One of well-known relaxations is class-wise, or marginal, calibration [Zadrozny and Elkan, 2002, Vaicenavicius et al., 2019, Kull et al., 2019].

Definition 2 (Class-wise calibration)**.** A probabilistic predictor $f : \mathcal{X} \to \Delta_K$ is said to be class-wise calibrated if

$$\mathbb{P}\left(Y = y \mid f_y(X)\right) = f_y(X), \quad y \in \mathcal{Y}. \tag{11}$$

Vaicenavicius et al. [2019] illustrate the difference with the canonical calibration through useful examples. In the binary setting, the two notions are equivalent with class-wise calibration being a weaker requirement for larger number of classes. It is achieved by reducing the original multiclass problem to K one-vs-all binary problems with the standard post-processing routine applied consequently to each one. We focus on canonical calibration for multiclass problems as per Definition 1 and explicitly mention important implications for the binary setting, and thus marginal calibration.

3.1 CALIBRATION FOR I.I.D. DATA

First, we assume that the binning scheme has been chosen and use $g : \mathcal{X} \to \mathcal{M}$ to denote the bin-mapping function: $g(x) = m$ if and only if $f(x) \in B_m$. The calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ is used for estimating

$$\pi_{y,m}^P := \mathbb{P}\left(Y = y \mid f(X) \in B_m\right), \quad y \in \mathcal{Y}, \tag{12}$$

for all bins $m \in \mathcal{M}$. The superscript here highlights that probabilities correspond to the source distribution P and the notation will become convenient when we talk about label shift setting. With finite data one can only estimate (12) with quantifiable measures of error, and thus provably satisfy the calibration requirement only approximately:

$$\mathbb{P}\left(Y = y \mid \widehat{\pi}_{y,g(X)}^P\right) \approx \widehat{\pi}_{y,g(X)}^P. \tag{13}$$

Let $N_m = |\{(X_i, Y_i) \in \mathcal{D}_{\text{cal}} : f(X_i) \in B_m\}|$ denote the number of calibration points that fall into bin $m \in \mathcal{M}$. Note that $\{N_m\}_{m \in \mathcal{M}}$ are random and satisfy $\sum_{m=1}^M N_m = n$.

Empirical frequencies of class labels $y \in \mathcal{Y}$ in each bin $m \in \mathcal{M}$:

$$\widehat{\pi}^P_{y,m} := \frac{1}{N_m} \sum_{i=1}^{n} \mathbb{1}\left\{Y_i = y, f(X_i) \in B_m\right\}, \qquad (14)$$

are natural candidates to satisfy the approximate calibration condition (13). For convenience, let $\pi^P_m := (\pi^P_{1,m}, \ldots, \pi^P_{K,m})^\top$ denote a vector with coordinates representing bin-conditional class probabilities and let $h : \mathcal{X} \to \Delta_K$ denote the *recalibrated* predictor, i.e., the function that maps any feature vector to the corresponding vector of *calibrated* probability estimates: $h(x) = \widehat{\pi}_{g(x)}$.

Theorem 3. *Fix $\alpha \in (0, 1)$. With probability at least $1 - \alpha$, $\left\|\widehat{\pi}^P_m - \pi^P_m\right\|_1 \leq \varepsilon_m$, simultaneously for all $m \in \mathcal{M}$, where*

$$\varepsilon_m := \frac{2}{\sqrt{N_m}} \sqrt{\frac{1}{2} \ln\left(\frac{M 2^K}{\alpha}\right)}.$$

As a consequence, with probability at least $1 - \alpha$,

$$\sum_{y=1}^{K} \left|\mathbb{P}\left(Y = y \mid h(X) = z\right) - z_y\right| \leq \max_{m \in \mathcal{M}} \varepsilon_m,$$

simultaneously for all z in the range of h.

The proof is given in Appendix C.1. In words, Theorem 3 states that as long as the least-populated bin contains sufficiently many points, the output of the recalibrated predictor will approximately satisfy condition (13). The first part of Theorem 3 justifies use of empirical frequencies in place of unknown population quantities using the language of the confidence intervals. In the binary setting, the fact that it yields the desired calibration guarantee, has been formally established by Gupta et al. [2020], and the second part of the theorem states a corresponding result for canonical calibration in the multiclass setting.

A natural question is whether one can guarantee that each bin is supplied with a sufficient number of calibration data points in order to obtain meaningful bounds. We note that in the binary setting, one way to provably spread the calibration data evenly across bins is uniform-mass, or equal frequency, binning [Kumar et al., 2019, Gupta et al., 2020, Gupta and Ramdas, 2021].

3.2 LABEL-SHIFTED CALIBRATION

For illustrating the necessity of accounting for label shift we consider the following binary classification problem: $\mathcal{Y} = \{0, 1\}$ with class probabilities given as $p(0) = p(1) = 1/2$ and $q(0) = 0.2$, $q(1) = 0.8$, i.e., while on the source domain classes are equally balanced, on the target class 1 becomes dominant. For each data point, conditionally on the corresponding label, the covariates are sampled according to $X \mid Y = y \sim \mathcal{N}(\mu_y, \Sigma)$, where

$$\mu_0 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}.$$

Similarly to the toy example from Section 2.2, here the class-posterior probabilities, and thus the Bayes-optimal rules have a closed form for both source and target domains. Not only do they minimize the probability of misclassifying a new point from the corresponding domain but also they are calibrated[2]. For the source distribution a perfect probabilistic predictor is given by

$$\pi^P_1(x) = \frac{p(1) \cdot \varphi(x; \mu_1, \Sigma)}{p(0) \cdot \varphi(x; \mu_0, \Sigma) + p(1) \cdot \varphi(x; \mu_1, \Sigma)}, \qquad (15)$$

where $\varphi(x; \mu_i, \Sigma)$, $i = 0, 1$ denotes the PDF of a Gaussian random vector with the corresponding parameters. As illustrated on Figure 2a, even though the Bayes-optimal rule is calibrated on the source, a correction is required to obtain a calibrated classifier under label shift. We sample points from the target distribution and highlight those that fall inside the area $S = \left\{x \in \mathbb{R}^2 : \pi^P_1(x) \in [0.4; 0.6]\right\}$ with boundary given by the black dashed lines. When the shift is present, predictor (15) is no longer calibrated, since otherwise one should expect roughly half of the test data points inside S to be labeled as class 1 (red squares) and half as class 0 (blue circles), which clearly does not happen.

If both the true class-posterior distribution $\pi^P_y(x)$ and the true label likelihood ratios w are known, then the form of the adjustment of the probabilistic classifier under label shift is a simple implication of the Bayes rule [Saerens et al., 2002]:

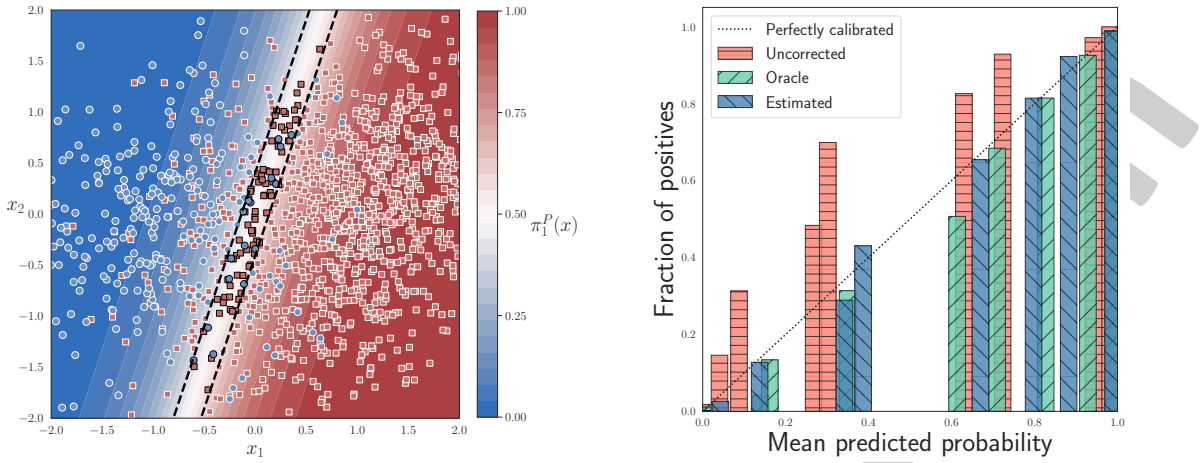
$$\pi^Q_y(x) = \frac{w(y) \cdot \pi^P_y(x)}{\sum_{k=1}^{K} w(k) \cdot \pi^P_k(x)}. \qquad (16)$$

While in the oracle setting predictor (16) is indeed calibrated on the target, in practice neither $\pi^P_y(x)$ nor w are known. Using corresponding plug-in estimators in (16) would guarantee calibration of the resulting predictor only asymptotically and under restricting modeling assumptions, and thus to obtain the distribution-free guarantees the output of the original predictor has to be discretized, or binned as in the i.i.d. setting. Relationship (16) does clearly continue to hold as formally stated next.

Proposition 1. *Under label shift, for any class label $y \in \mathcal{Y}$ and any bin B_m, $m \in \mathcal{M}$ it holds that:*

$$\pi^Q_{y,m} = \frac{w(y) \cdot \pi^P_{y,m}}{\sum_{k=1}^{K} w(k) \cdot \pi^P_{k,m}}.$$

[2]Recall that in the binary setting, canonical and class-wise calibration are equivalent.



(a) (b)

Figure 2: (a) Sampled points from the target distribution plotted against the true source class-posterior probabilities. (b) Reliability curves for Fisher's LDA calibrated via binning with/without taking label shift into account. The deviation of uncorrected probabilities from the diagonal line (perfect calibration) reflects the need to correct for label shift; recalibration based on estimated weights is almost identical to using oracle weights, both of which result in near-perfect calibration.

In Section 3.1 we justified the use of empirical frequencies of class labels $\{\widehat{\pi}_m^P\}_{m\in\mathcal{M}}$ for achieving canonical calibration of a predictor on the source domain and, as it has been noted in Section 2.2, there are estimators of the importance weights which are known to be provably consistent under reasonable assumptions. Thus, with an estimator $\widehat{w}$ at hand, Proposition 1 suggests an appropriate correction to provably obtain asymptotically calibrated predictors on the target:

$$\widehat{\pi}_{y,m}^{(\widehat{w})} = \frac{\widehat{w}(y)\cdot\widehat{\pi}_{y,m}^P}{\sum_{k=1}^K \widehat{w}(k)\cdot\widehat{\pi}_{k,m}^P}, \quad y\in\mathcal{Y}, \tag{17}$$

for all bins $m\in\mathcal{M}$. Theorem 3 quantifies the error when the empirical label frequencies are used as estimators for the true unknown bin-conditional class probabilities on the source domain. However, different bounds on ε_m could be available depending on chosen binning scheme, and thus we instead quantify how this estimation error on the source domain translates into the estimation error on the target for the cases when the importance weights are known and when they are rather estimated. As we shall see, the performance depends on the ratio of the largest to the smallest nonzero importance weight. Define the *condition number*:

$$\kappa := \frac{\sup_k w(k)}{\inf_{k:w(k)\neq 0} w(k)},$$

with $\kappa = 1$ corresponding to label shift not being present. Next, we quantify the miscalibration of the predictor (17).

Theorem 4. *Let $\widehat{w}$ be an estimator of w and let $\widehat{\pi}_{y,m}^{(\widehat{w})}$ denote the reweighted empirical frequencies* (17) *for all labels $y\in\mathcal{Y}$ and bins $m\in\mathcal{M}$. For any bin $m\in\mathcal{M}$, it holds that:*

$$\left\|\widehat{\pi}_m^{(\widehat{w})} - \pi_m^Q\right\|_1 \leq \underbrace{2\kappa\cdot\left\|\widehat{\pi}_m^P - \pi_m^P\right\|_1}_{(a)} + \underbrace{\frac{2\left\|\widehat{w}-w\right\|_\infty}{\inf_{l:w(l)\neq 0} w(l)}}_{(b)}. \tag{18}$$

The proof is given in Appendix C.1. In words, the calibration error on the target decomposes into two terms where (a) is controlled by the calibration error on the source and (b) is controlled by the importance weights estimation error. Further, under reasonable assumptions common procedures, such as BBSE and RLLS, construct estimators of the importance weights which are not only known to be consistent but also have quantifiable error [Lipton et al., 2018, Azizzadenesheli et al., 2019]. Similarly, any proper binning scheme that provably controls number of calibration points in each bin, e.g., uniform-mass binning in the binary setting [Kumar et al., 2019], yields finite-sample guarantees for the calibration error on the source [Gupta et al., 2020]. Thus, finite-sample guarantees for the miscalibration of the resulting predictor on the target domain trivially follow by virtue of Theorem 4 via invoking simple probabilistic arguments.

Within the same binary classification setup from the beginning of Section 3.2, we also compare calibration via uniform-mass binning with and without accounting for label shift but this time we use Fisher's LDA as an underlying classifier, which differs from the Bayes-optimal rule by using estimators of the corresponding means and covariance matrices in (15). Results illustrated on Figure 2b via the reliability curves indicate that shift-corrected binning with either true or estimated importance weights yields a calibrated predictor on the target domain while uncorrected

fails to do so as expected. To complete the empirical study, Appendix C.2 further examines calibration with and without accounting for label shift on the `wine quality` dataset from Section 2.2.

4 DISCUSSION

For safety-critical applications model's prediction must be supported with a proper measure of uncertainty. As various ad-hoc procedures provide valid inference only under assumptions that are either unrealistic or unverifiable, it is essential to understand whether non-trivial guarantees can be obtained in an assumption-lean manner. Guided by this principle, we analyzed distribution-free uncertainty quantification for classification via two complementary notions: prediction sets and calibration.

We focused on a less studied — but still highly relevant to real-world scenarios — setting of label shift. While it is evident that label shift does hurt model's calibration, the corresponding impact on prediction sets is less obvious. In the extreme example of almost perfectly separable data, prediction sets are usually expected to contain the most likely label only, and thus coverage is not expected to suffer much no matter how the class proportions change for the test data. Still, as we illustrated, in less idealized settings, a correction for label shift is necessary. By adapting conformal prediction sets and calibration via binning to label shift, we close an existing gap for distribution-free uncertainty quantification under two standard ways of generalizing beyond the classic i.i.d. setting. Importantly, those adaptations do not require labeled data from the target domain which can be useful in applications where the labeling process is expensive. We note that handling label shift should be expected to be an easier task rather than handling another common setting — covariate shift — as the latter typically involves estimating a high-dimensional, and usually continuous, likelihood ratio.

With theoretical results available for calibration in the binary setting, and thus class-wise (coordinatewise) calibration in a more general multiclass setting, establishing meaningful guarantees for "full" canonical calibration in the latter setting remains an intriguing future research direction. One particular example is related to the question of the importance weights estimation under label shift. While approaches based on confusion matrices, e.g., BBSE and RLLS, provably yield consistent estimators under relatively mild assumptions, alternative approaches, such as MLLS with preceding ad-hoc calibration on the source domain, tend to perform better empirically [Alexandari et al., 2020]. Theoretical foundations for MLLS developed recently by Garg et al. [2020] require the underlying predictor to be canonically calibrated which is itself, unfortunately, hard to guarantee provably which creates a (somewhat circular) gap between theory and practice.

Acknowledgements

The authors would like to thanks Chirag Gupta and the anonymous UAI 2021 reviewers for comments on an initial version of this paper.

References

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*, 2020.

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019.

Foygel Rina Barber, J. Emmanuel Candes, Aaditya Ramdas, and J. Ryan Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 2021.

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

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 2009.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, 2020.

Leying Guan and Rob Tibshirani. Prediction and outlier detection in classification problems. *arXiv preprint: 1905.04396*, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Chirag Gupta and Aaditya Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, 2021.

Chirag Gupta, Arun K. Kuchibhotla, and Aaditya K. Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *arXiv preprint: 1910.10562*, 2019.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *Advances in Neural Information Processing Systems*, 2020.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, 2019.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 2019.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 2013.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 2018.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, 2018.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 1999.

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

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

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.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 2002.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 2000.

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

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B Schön. Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics*, 2019.

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

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alex Gammerman. Criteria of efficiency for conformal prediction. In *Symposium on Conformal and Probabilistic Prediction with Applications*, 2016.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *International Conference on Machine Learning*, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *International Conference on Knowledge Discovery and Data Mining*, 2002.