Distribution-free uncertainty quantification for classification under label shift

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

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, and in this work we focus on the latter setting. It is known that label shift hurts prediction, and we argue that it also hurts UQ, by showing degradation in coverage and calibration. We examine, both theoretically and empirically, the right way to achieve UQ by reweighting the aforementioned conformal and calibration procedures. **This work has been accepted to UAI 2021.**

1 Introduction

It is common in classification to assume access to a dataset $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y} = \{1, \dots, K\}$ denote the features and the labels respectively, and the pairs (X_i, Y_i), $i = 1, \dots, n$ are sampled i.i.d. from some unknown joint distribution P, and it is used to learn a predictor f, a mapping from $\mathcal{X}$ to rankings or distributions over $\mathcal{Y}$, by optimizing some loss/risk. Still, in some applications accurate point prediction alone is insufficient and an associated measure of uncertainty is required.

For common predictors that are mappings of the form $f : \mathcal{X} \to \Delta_K$, where Δ_K refers to the probability simplex in $\mathbb{R}^K$, the output vector $f(X)$ is expected to reflect the true conditional probabilities of classes given the observed input, which won't be true without, typically strong and unverifiable in practice, assumptions. We focus on two complementary categories of post-processing procedures — calibration via post-hoc binning and conformal prediction — that use held-out data (called *calibration* dataset) and any trained model (e.g., deep NNs) to construct a *wrapper* that provably quantifies uncertainty without any distributional assumptions about the data generating mechanism. While the former aims to amend the output of a predictor so that it has a rigorous frequentist interpretation, the latter aims to produce a set of labels that contains the truth with high probability. While calibration could be a better way of UQ in the binary setup, corresponding mathematical guarantees degrade with growing number of classes, and prediction sets become an attractive option for UQ. 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 that are practically efficient and theoretically grounded.

In real-world applications, the *target* distribution Q (generating test data) might not be the same as the *source* distribution P (generating training data) that leads to violation of the assumptions under which even assumption-lean UQ is valid. One may hope that it is possible to make simplifying assumptions that would allow us to perform appropriate corrections for the procedures to retain guarantees. Common assumptions about the type of shift include *covariate shift* (Shimodaira, 2000) and *label shift* (Saerens et al., 2002). Within the context of distribution-free UQ, covariate shift has recently received attention (Tibshirani et al., 2019; Gupta et al., 2020), and we close an existing gap for quantifying predictive uncertainty under label shift. Classic approaches for handling label shift make an assumption that the target support is contained in the source support, so that the label likelihood ratio (or *importance weights*) $q(y)/p(y)$ is well-defined, and computationally attractive estimation procedures that use labeled data only from the source distribution are available (Lipton et al., 2018; Azizzadenesheli et al., 2019; Saerens et al., 2002; Alexandari et al., 2020).

Building on recent progress in distribution-free UQ, we adapt both procedures to handling label shift through a proper form of reweighting that makes use only of unlabeled data from the target. We also consider an alternative way of addressing label shift by performing label-conditional conformal classification (Sadinle et al., 2019; Vovk et al., 2016; Guan and Tibshirani, 2019; Vovk et al., 2005).

[1]Department of Statistics & Data Science, Machine Learning Department, Carnegie Mellon University. Correspondence to: Aleksandr Podkopaev <podkopaev@cmu.edu>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

2 Conformal classification

We wish to construct an uncertainty set function $C : \mathcal{X} \to 2^{\mathcal{Y}}$, such that for a new (test) data point:

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

Conformal prediction (Vovk et al., 2005; Lei et al., 2018; Barber et al., 2021; Romano et al., 2019; Cauchois et al., 2020; Angelopoulos et al., 2021; Romano et al., 2020; Gupta et al., 2019) does not make any distributional assumptions at the price of provably providing only *marginal* guarantees as stated in (1), but not conditional on a given input.

Exchangeable conformal If the true distribution $\pi_y(x) = \mathbb{P}\left[Y = y \mid X = x\right]$ is known, the *oracle* prediction set is obtained by including the next most-likely label as long as the total mass of those included before is less than $1 - \alpha$ (Lei et al., 2013; Sadinle et al., 2019):

$$C_\alpha^{\text{oracle}}(x) := \left\{y \in \mathcal{Y} : \rho_y(x; \pi) < 1 - \alpha\right\},$$
$$\text{where} \quad \rho_y(x; \pi) := \sum_{y'=1}^{K} \pi_{y'}(x) \mathbb{1}\left\{\pi_{y'}(x) > \pi_y(x)\right\} \tag{2}$$

is the total mass of labels that are more likely than $y \in \mathcal{Y}$. Split-conformal framework describes a way of updating the threshold $1-\alpha$ in (2) in order to obtain coverage guarantees when an estimator $\widehat{\pi}$ is used in place of the true π. Conformalization applied naively to (2) yields prediction sets with typically inferior conditional coverage in practice. We instead consider conformalization of a sequence of nested prediction sets motivated by the randomized version of (2):

$$\mathcal{F}_\tau(x, u; \widehat{\pi}) = \left\{y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leq \tau\right\}, \tag{3}$$

where $\tau \in \mathcal{T} = [0, 1]$ and u is a realization of $\text{Unif}\left([0, 1]\right)$, sampled independently of anything else (Vovk et al., 2005; Romano et al., 2020). For any triple (X, Y, U) the smallest radius of a set that captures label Y, or score, is given by

$$r(X, Y, U; \widehat{\pi}) = \rho_Y(X; \widehat{\pi}) + U \cdot \widehat{\pi}_Y(X). \tag{4}$$

We further discuss our score choice in place of alternative ones (Angelopoulos et al., 2021; Cauchois et al., 2020) 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 $\mathcal{I}_2 = \{1, \ldots, n\}$. Exchangeability of the data implies exchangeability of the scores $r_i = r(X_i, Y_i, U_i; \widehat{\pi})$, $i \in \mathcal{I}_2 \cup \{n+1\}$. Thus

$$\mathcal{F}_{\tau^\star}\left(x, u; \widehat{\pi}\right) = \left\{y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leq \tau^\star\right\},$$
$$\tau^\star = Q_{1-\alpha}\left(\{r_i\}_{i \in \mathcal{I}_2} \cup \{1\}\right), \tag{5}$$

attains the right coverage guarantee which is a classic result in conformal prediction[1].

2.1 Label-shifted conformal

When the true likelihood ratios $w(y) = q(y)/p(y)$ are known for all $y \in \mathcal{Y}$, we show that the prediction sets:

$$\mathcal{F}_{\tau^\star}^{(w)}\left(x, u; \widehat{\pi}\right) = \left\{y \in \mathcal{Y} : \rho_y\left(x; \widehat{\pi}\right) + u \cdot \widehat{\pi}_y(x) \leq \tau_w^\star(y)\right\},$$
$$\tau_w^\star(y) = Q_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{p}_i^w(y) \delta_{r_i} + \tilde{p}_{n+1}^w(y) \delta_1\right), \tag{6}$$

$$\text{where} \quad \tilde{p}_i^w(y) = \frac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(y)}, \quad i = 1, \ldots, n,$$
$$\tilde{p}_{n+1}^w(y) = \frac{w(y)}{\sum_{j=1}^n w(Y_j) + w(y)}, \tag{7}$$

are provably valid. Note that the thresholds themselves now vary depending on the class label. The formal guarantee for the prediction set (6) 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.

Theorem 1. *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*

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

Weighted exchangeability argument guarantees correct coverage in case of known importance weights, but in practice one only has access to a corresponding estimator. Dominant methods estimate importance weights using a separate labeled dataset from the source distribution and unlabeled dataset from the target (see Appendix A), and under reasonable assumptions yield consistent estimators 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 (6) does not have a jump at level $1 - \alpha$. Then:*

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

The performance of reweighting is illustrated on Figure 1 using the `wine quality` dataset (Cortez et al., 2009) (Appendix B.5 contains additional details). Both shift-corrected conformal prediction sets demonstrate superior coverage performance compared with uncorrected ones, and the procedure with estimated importance weights has a slightly downgraded performance. We also refer the reader to a simulated experiment in Appendix B.4.

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

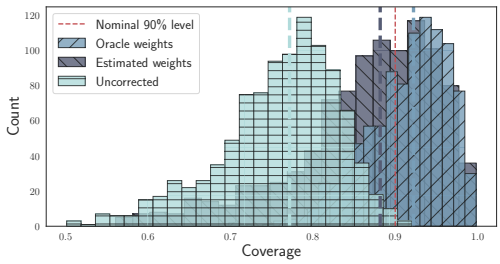

Figure 1. Empirical coverage on the `wine quality` dataset. Dashed vertical lines describe the median coverage values, which are significantly worse when label shift is not accounted for, while using estimated weights mimics the oracle reasonably well.

Label-conditional conformal (LCC) Conformal framework can be applied in a way that makes the resulting prediction sets inherently robust to label shift (Sadinle et al., 2019; Vovk et al., 2016; Guan and Tibshirani, 2019; Vovk et al., 2005). By fixing a set of significance levels per class $\{\alpha_y\}_{y\in\mathcal{Y}}$ (e.g., $\alpha_y = \alpha$ for all y) and further splitting of 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:

$$\mathcal{F}^{\mathrm{c}}_{\tau^\star_{\mathrm{c}}}(x, u; \widehat{\pi}) = \{y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \le \tau^\star_{\mathrm{c}}(y)\},$$
$$\tau^\star_{\mathrm{c}}(y) = Q_{1-\alpha_y}\left(\{r_i\}_{i\in\mathcal{I}_{2,y}} \cup \{1\}\right). \tag{8}$$

In words, for each label a separate hypothesis test is performed to determine whether there is enough evidence to exclude it from the prediction set. If the conditional distribution of X given $Y = y$ for all $y \in \mathcal{Y}$ does not change, the non-conformity score $r(X_{n+1}, Y_{n+1}, U_{n+1}; \widehat{\pi})$ together with $\{r_i\}_{i\in\mathcal{I}_{2,Y_{n+1}}}$ will form a collection of exchangeable random variables, which implies label-conditional validity:

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

for all $y \in \mathcal{Y}$. If $\alpha_y = \alpha$ for all y, marginalizing over y using *any* distribution (shifted or not) yields that LCC is robust to label shift. The price to pay for the stronger conditional guarantee is the size of the prediction sets: for example, for not well-separated classes LCC can be expected to yield larger prediction sets; see Appendix B.6 for an empirical study. Moreover, splitting available calibration data into K parts (for large K) could result in large losses of statistical efficiency. Still, LCC does not require importance weights estimation, and thus has exact finite-sample guarantee instead of asymptotic (Corollary 1), and 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

First, we state the definition of a canonical calibration which is of central interest to us.

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

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

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

Canonical calibration requires the whole output vector to reflect the true conditional probabilities, and two extreme examples of calibrated predictors include the one that outputs marginal probabilities of classes and the one that outputs the true class-posterior probabilities, which also represent two extremes in terms of classification efficiency. Prediction models are typically not calibrated out-of-the-box, and additional post-processing is performed, but many common procedures lack formal theoretical guarantees (Platt, 1999; Guo et al., 2017; Zadrozny and Elkan, 2001; 2002). Miscalibration is assessed using either reliability curves or related one-dimensional summary statistics, and popular metrics, such as Expected Calibration Error (ECE), are known to be unreliable if discretization of the output of the resulting model is not performed (Kumar et al., 2019; Vaicenavicius et al., 2019). Discretization is also needed for obtaining distribution-free calibration guarantees in the binary case (Gupta et al., 2020). Binning represents a 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: $\mathcal{X}_m := \{x \in \mathcal{X} : f(x) \in B_m\}$, $m \in \mathcal{M} := \{1, \ldots, M\}$. Binning step makes achieving canonical calibration prohibitive for large number of classes as each bin has to be supplied with sufficiently many data points 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, e.g., class-wise/marginal, calibration (Zadrozny and Elkan, 2002; Kull et al., 2019), a weaker requirement in the non-binary setup. Vaicenavicius et al. (2019) illustrate the difference with the canonical calibration through useful examples.

3.1 Calibration for i.i.d. data

Assume that the binning scheme has been chosen and with $g : \mathcal{X} \to \mathcal{M}$ denoting 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

$$\pi^P_{y,m} := \mathbb{P}(Y = y \mid f(X) \in B_m), \quad y \in \mathcal{Y}, \tag{9}$$

for $m \in \mathcal{M}$. Thus, with finite data, it is possible to provably satisfy the calibration requirement only approximately:

$$\mathbb{P}\left(Y = y \mid \widehat{\pi}^P_{y,g(X)}\right) \approx \widehat{\pi}^P_{y,g(X)}. \tag{10}$$

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

$$\widehat{\pi}^P_{y,m} = \frac{1}{N_m} \sum_{i=1}^{n} \mathbb{1}\left\{Y_i = y, f(X_i) \in B_m\right\}, v \qquad (11)$$

for $m \in \mathcal{M}$, is a natural candidate to satisfy the approximate calibration (10). Let $\pi^P_m := (\pi^P_{1,m}, \ldots, \pi^P_{K,m})^\top$ denote a vector with bin-conditional class probabilities as coordinates and let $h : \mathcal{X} \to \Delta_K$ denote the 'recalibrated' predictor: $h(x) = \widehat{\pi}_{g(x)}$. The formal guarantee for the i.i.d. case (Theorem 3 in Appendix C.1) states that the recalibrated predictor will satisfy (10) as long as the least-populated bin has sufficiently many points. One way to provably spread the calibration data evenly across bins for the binary setting is uniform-mass binning (Kumar et al., 2019).

3.2 Label-shifted calibration

If the true distribution $\pi^P_y(x)$ and weights w are known, the form of the adjustment of a predictor under label shift is a simple implication of the Bayes rule (Saerens et al., 2002):

$$\pi^Q_y(x) = \frac{w(y) \cdot \pi^P_y(x)}{\sum_{k=1}^{K} w(k) \cdot \pi^P_k(x)}. \qquad (12)$$

Using estimators of $\pi^P_y(x)$ and w in (12) would yield asymptotically calibrated predictor only under strong modeling assumptions. As in the i.i.d. setting, to obtain the distribution-free guarantees the output of a predictor has to be discretized, and the relationship (12), which will still continue to hold (Appendix C.1), suggests an appropriate correction for obtaining a canonically calibrated predictor on the target:

$$\widehat{\pi}^{(\widehat{w})}_{y,m} = \frac{\widehat{w}(y) \cdot \widehat{\pi}^P_{y,m}}{\sum_{k=1}^{K} \widehat{w}(k) \cdot \widehat{\pi}^P_{k,m}}, \quad y \in \mathcal{Y}, m \in \mathcal{M}. \qquad (13)$$

We quantify how the estimation error of $\pi^P_{y,m}$ on the source translates into the estimation error on the target when the importance weights are known/estimated, and the performance depends on the ratio of the largest to the smallest nonzero importance weight. Define the *condition number*: $\kappa := \sup_k w(k) / \inf_{k:w(k)\neq 0} w(k)$, with $\kappa = 1$ corresponding to label shift not being present.

Theorem 2. *Let $\widehat{w}$ be an estimator of w and let $\widehat{\pi}^{(\widehat{w})}_{y,m}$ denote the reweighted empirical frequencies (13) for all labels $y \in \mathcal{Y}$ and bins $m \in \mathcal{M}$. For any bin $m \in \mathcal{M}$, it holds that:*

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

Miscalibration on the target decomposes into two terms where (a) is controlled by miscalibration on the source and (b) is controlled by the importance weights estimation error. Finite-sample guarantees for miscalibration of the resulting predictor trivially follow by virtue of Theorem 2 for any chosen binning scheme and importance weights estimation procedure. We compare calibration with/without accounting for label shift for Fisher's LDA (see Appendix C.2 for details) on Figure 2. The reliability curves indicate that shift-corrected binning with true/estimated weights yields a calibrated predictor on the target while uncorrected fails to do so. To complete the empirical study, Appendix C.3 further examines calibration with and without accounting for label shift on the `wine quality` dataset from Section 2.1.

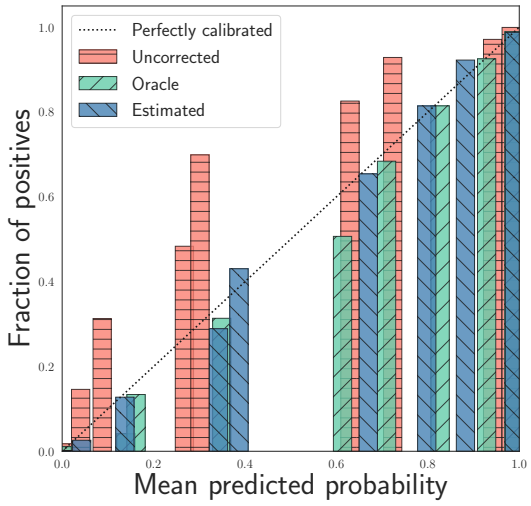

Figure 2. Reliability curves for binning with/without accounting for label shift. The deviation from the diagonal line reflects the need to correct for label shift; recalibration with both oracle/estimated weights results in near-perfect calibration.

4 Discussion

For safety-critical applications model's prediction must be supplemented with a proper measure of uncertainty. While 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 and illustrated that a correction for label shift might be necessary and established ways of adapting related procedures to this setting.

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, 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.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence*. Springer, 1996.

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.

A Importance weights estimation under label shift

Below we provide details about importance weights estimation procedures which are relevant mainly to Sections 2.1 and 3.2 of the paper. Estimation of the importance weights is performed using a held-out labeled set from the source distribution and an unlabeled set from the target distribution. Procedures, such as BBSE (Lipton et al., 2018) or RLLS (Azizzadenesheli et al., 2019), are based on estimation of the confusion matrix and yield consistent importance weights estimators with quantifiable estimation error under relatively mild assumptions. First, given a black-box predictor $f : \mathcal{X} \to \Delta_K$, define the corresponding expected confusion matrix $C_P(f) \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$:

$$[C_P(f)]_{ij} := \mathbb{E}_P \left[\mathbb{1}\left\{\text{argmax}_k f_k(X) = i\right\} \cdot \mathbb{1}\left\{Y = j\right\}\right].$$

We assume that

(A1) for every label $y \in \mathcal{Y}$, it holds that $q(y) > 0 \implies p(y) > 0$,

(A2) expected confusion matrix $C_P(f)$ is full-rank.

Assumption (A1) states that target label distribution is absolutely continuous with respect to the source. Indeed, reasoning properly about a class in the target domain which is not represented in the source domain is not possible. Assumption (A2) simply represents an identifiability condition. Lipton et al. (2018) show that under label shift assumption: $\mathbb{P}_Q\left(f(X) = i\right) = \sum_{j\in\mathcal{Y}} \left[C_P(f)\right]_{ij} w(j)$, or in matrix-vector notation:

$$\mu = C_P(f)w.$$

where $\mu \in \mathbb{R}^{|\mathcal{Y}|} : \mu_i = \mathbb{P}_Q\left(f(X) = i\right)$. BBSE is a simple plug-in procedure, which yields the following estimator of the importance weights:

$$\widehat{w} = \widehat{C}^{-1}\,\widehat{\mu},$$

$$\text{where} \quad \widehat{C}_{ij} = \frac{1}{m}\sum_{p=1}^{m} \mathbb{1}\left\{f(X_p^s) = i \text{ and } Y_p^s = j\right\},$$

$$\widehat{\mu}_i = \frac{1}{l}\sum_{p=1}^{l} \mathbb{1}\left\{f(X_p^t) = i\right\},$$

where $\{(X_i^s, Y_i^s)\}_{i=1}^m$ is a labeled dataset from the source distribution and $\{(X_i^t)\}_{i=1}^l$ is unlabeled data from the target distribution. BBSE-hard described above can be trivially modified to the whole probability distribution output of f which is referred to as BBSE-soft procedure. Under aforementioned assumptions, Lipton et al. (2018) establish results with respect to consistency of BBSE and corresponding convergence rates.

A well-known alternative approach to directly estimate the importance weights which performs well in practice is MLLS (Saerens et al., 2002) and its recent variations that combine it with preceding calibration on the source domain (Alexandari et al., 2020). We refer the reader to Garg et al. (2020) for the theoretical analysis of MLLS and a detailed overview of the results for the importance weights estimation under label shift. For all simulations in this work we use BBSE-soft procedure, and such choice is motivated simply by its satisfactory empirical performance throughout all of the simulations we performed. Our modular approach to UQ allows a practitioner to replace BBSE with any alternative of their choice.

B Conformal classification

Below, Section B.1 includes details about the tie-breaking rules for the oracle prediction sets, Section B.2 includes a discussion regarding the role of randomization for conformal classification, Section B.3 includes all necessary proofs for Sections 2. Section B.5 includes details about the simulation on a real dataset mentioned in Section 2.1.

B.1 Tie-breaking rules for the oracle prediction set

In practice, when an estimator $\widehat{\pi}_y(x)$ is used in place of $\pi_y(x)$, one does not expect ties to be present but for completeness it is important to consider such scenario in the oracle setting. First, note that for any $\alpha \in (0,1)$, the oracle prediction set clearly never include labels $y \in \mathcal{Y} : \pi_y(x) = 0$. Now, presence of ties can lead to a conservative prediction set for some $x \in \mathcal{X}$ if there is a subset of class labels $S(x) \subseteq \mathcal{Y}$ of size $L = |S(x)| > 1$, such that $\forall y, y' \in S(x) : \pi_y(x) = \pi_{y'}(x) > 0$ and

$$\begin{cases} \mathbb{P}\left(Y \in C_\alpha^{\text{oracle}}(X) \backslash S(X) \mid X = x\right) < 1 - \alpha, \\ \mathbb{P}\left(Y \in C_\alpha^{\text{oracle}}(X) \mid X = x\right) \geq (1 - \alpha). \end{cases}$$

In the oracle case ties can be broken arbitrarily in order to preserve the conditional coverage. One option is to break ties randomly, i.e. one can fix a random permutation of labels in $S(x)$: $\widetilde{y}_{i_1}, \ldots, \widetilde{y}_{i_l}$, and output a smaller oracle prediction set:

$$C_\alpha^{\text{oracle,new}}(X) := \left(C_\alpha^{\text{oracle}}(X) \backslash S(X)\right) \cup \{\widetilde{y}_{i_1}, \ldots, \widetilde{y}_{i_{l^\star}}\},$$

where $l^\star$ is the smallest index in $\{1, \ldots, l\}$ such that

$$\mathbb{P}\left(Y \in C_\alpha^{\text{oracle}}(X) \backslash S(X) \mid X = x\right) + \sum_{k=1}^{l^\star} \pi_{i_k}(x) \geq 1 - \alpha.$$

B.2 Note on randomization and conditional coverage

As the number of works on conformal classification has seen a recent spurt, it is important to understand what exactly might be the benefits of using one nested sequence over another. For example, Angelopoulos et al. (2021) state in their Appendix B that "randomization is of little practical importance, since... output by the randomized procedure will differ from that of the non-randomized procedure by at most one element". However, we do not quite agree with their sentiment about it being of little practical importance for the following reason. While their observation is indeed accurate in the oracle setting, there is a noticeable difference in the empirical conditional coverage when the nested sequences are conformalized in practice (non-oracle setting). Roughly speaking, randomized scores better handle the heterogeneity of the conditional distribution of the response variable across the sample space. Note that this type of randomization has a different role from that of a randomized conformal p-value (Vovk et al., 2005) which deals with conservative coverage due to possible ties among the non-conformity scores. We believe that the reasoning below complements the one given in Romano et al. (2020) and, in particular, might help an unfamiliar reader to gain some useful insights (as well as arguably having simpler notation). For completeness, we start with an example of randomization in action. Consider a binary classification problem: $\mathcal{Y} = \{0, 1\}$, and fix target miscoverage level $\alpha = 0.05$. Now, assume that for some $x \in \mathcal{X}$:

- $\pi_0(x) = 0.99$, $\pi_1(x) = 0.01$. Then with probability $95/99$, we have $\tilde{C}_\alpha^{\text{oracle}}(x, u) = \{0\}$ and $\tilde{C}_\alpha^{\text{oracle}}(x, u) = \{\emptyset\}$ otherwise.
- $\pi_0(x) = 0.9$, $\pi_1(x) = 0.1$. Then with probability $1/2$, $\tilde{C}^{\text{oracle}}(x, u) = \{0, 1\}$ and $\tilde{C}^{\text{oracle}}(x, u) = \{0\}$ otherwise.

First, consider the marginal coverage of conformal prediction sets in the "null" case when $\widehat{\pi} \equiv \pi$. The marginal coverage guarantee of conformal prediction sets is due to Lemma 5 which states a classic result for quantiles of exchangeable random variables and is tight when these variables are almost surely distinct. In the non-randomized setting for any point (X, Y), the corresponding non-conformity score are given by $\rho_Y(X; \pi)$. Such form might suggest that the marginal coverage could be conservative due to possible ties as whenever the predicted most likely label appears to be the correct one, it holds that $\rho_Y(X; \pi) = 0$. However, if ties among non-conformity scores are present, they would typically occur only between zero-valued scores, and thus in a reasonable classification setup one should expect the marginal coverage to be tight even for non-randomized nested sequence as the calibrated threshold would typically be nonzero.

Next, before reasoning about conditional coverage of conformal sets, recall that the conditional distribution of the response is discrete in classification setting, and thus even

in the null case it is hard to reason meaningfully about the distribution of non-conformity scores $\rho_Y(X;\pi)$. However, Romano et al. (2020) noticed that if randomization of the scores is used, then it becomes possible to do at least in the null case. If $\widehat{\pi} \equiv \pi$, it is trivial to see the distribution of corresponding non-conformity scores $\rho_Y(X;\pi) + U \cdot \pi(X)$ is uniform conditional on X. Then, as the authors conjecture, it is intuitive that conformal prediction sets would recover the oracle ones under some consistency assumptions for $\widehat{\pi}$.

However, randomization is also performed when the prediction set is a singleton containing the most likely label only, and thus might yield non-interpretable and non-actionable empty prediction sets being purely the consequence of deploying randomization. Thus one might consider abstaining from dropping a label from the prediction set whenever it forms a singleton and perform randomization if and only if the oracle prediction set contains more than one label. While that decision can be embedded into either prediction step only or calibration step as well, we state explicitly that it should be done at the prediction step only for the aforementioned reasons.

Consider the binary toy example from Section 3.2 with focus on the source distribution only. As the true class-posterior probability $\pi_1^P(x)$ is known, we construct the non-randomized oracle prediction set C^{oracle} and compare it visually with the randomized version $\widetilde{C}^{\text{oracle}}$ on Figures 3 and 4 where randomization demonstrates desired behavior.

Consequently, we consider conformal prediction sets based on non-randomized sequence:

$$\begin{aligned} \mathcal{F}_{\tau^\star}(x,u;\widehat{\pi}) &= \{y \in \mathcal{Y} : r'(x,y) \leq \tau^\star\}, \\ \tau^\star &= Q_{1-\alpha}\left(\{r_i'\}_{i\in\mathcal{I}_2} \cup \{1\}\right), \\ r'(x,y) &= \rho_y(x;\widehat{\pi}), \end{aligned} \tag{14}$$

and two randomized sequences where Scheme 1 performs randomization for all labels and was introduced before for conformal prediction sets (5) and Scheme 2 (added for completeness of comparison) performs randomization for all labels except the most likely one:

$$\begin{aligned} \mathcal{F}_{\tau^\star}(x,u;\widehat{\pi}) &= \{y \in \mathcal{Y} : r''(x,y) \leq \tau^\star\}, \\ \tau^\star &= Q_{1-\alpha}\left(\{r_i''\}_{i\in\mathcal{I}_2} \cup \{1\}\right), \end{aligned} \tag{15}$$

where

$$r''(x,y) = \mathbb{1}\{\rho_y(x;\widehat{\pi}) > 0\} \cdot (\rho_y(x;\widehat{\pi}) + u \cdot \widehat{\pi}_y(x)).$$

We again use the Bayes-optimal classifier $\pi_y(x)$, and thus ignore the results that are due to estimation and focus purely on effects that are due to conformalization. For a single data draw we illustrate the resulting conformal prediction sets on Figures 5, 6 and 7. While at first sight it might seem that non-randomized nested sequences is superior in terms of yielding prediction sets with smaller cardinality, it should be taken with a grain of salt. We repeatedly draw calibration and test data and track marginal characteristics for those sets. As expected, all three resulting prediction sets inherit $1-\alpha$ (marginal) coverage guarantee as confirmed on Figure 8. Moreover, Figure 9 indeed confirms that randomization could yield larger prediction sets for not perfectly separable data. But Figure 10 confirms that randomization proposed by Romano et al. (2020) (Scheme 1) demonstrates superior conditional coverage since for this example the true $\pi_y(x)$ is used, and thus the oracle prediction sets are recovered if $\tau^\star = 1-\alpha$. Figure 11 confirms that oracle prediction sets are not recovered even when the size of the calibration set is increased.

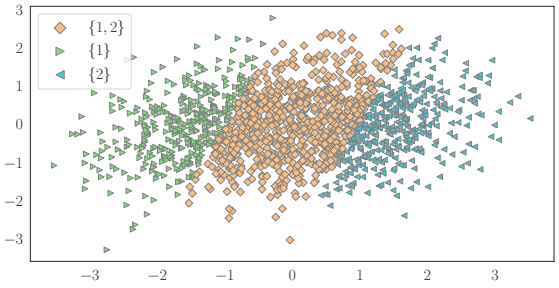

Figure 3. Prediction sets corresponding to the non-randomized oracle from (2). See Section B.2 for more details.

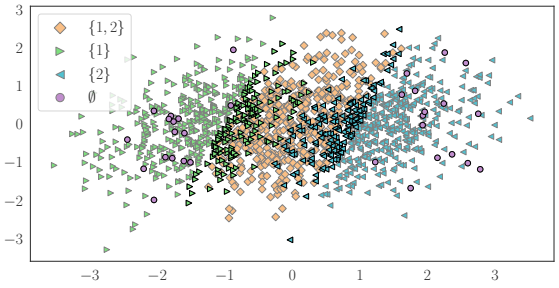

Figure 4. Prediction sets corresponding to the randomized oracle. See Section B.2 for more details.

B.3 Proofs

Proof of Theorem 1. First, recall the definition of weighted exchangeability (Tibshirani et al., 2019).

Definition 2 (Weighted exchangeability)**.** Random variables $Z_1, \ldots, Z_n$ are said to be *weighted exchangeable*, with weight functions $\omega_1, \ldots, \omega_n$, if the density f of their

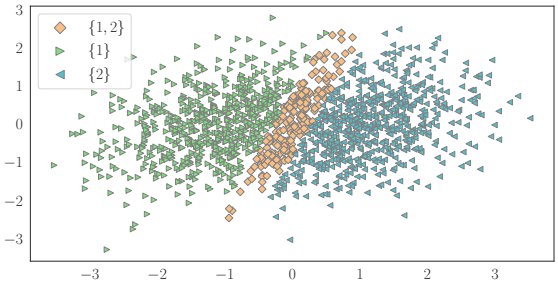

Figure 5. Prediction sets corresponding to the non-randomized conformal method (14). See Section B.2 for more details.

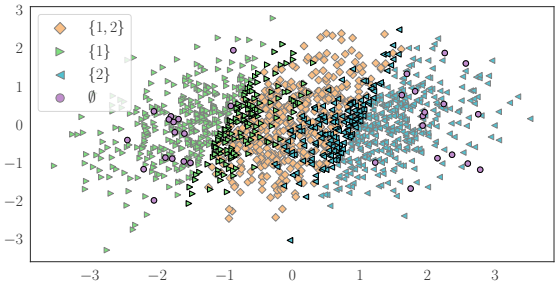

Figure 6. Prediction sets corresponding to the randomized conformal (scheme 1) method (5). See Section B.2 for more details.

joint distribution can be factorized as:

$$f(z_1, \ldots, z_n) = \prod_{i=1}^{n} \omega_i(z_i) \cdot g(z_1, \ldots, z_n),$$

where g is any function that that invariant to permutations of its arguments, i.e., $g(z_{\sigma(1)}, \ldots, z_{\sigma(n)})$ for any permutation σ of $1, \ldots, n$.

Independent draws are always weighted exchangeable and it is easy to see that under label shift setting $Z_i = (X_i, Y_i, U_i)$, $i = 1, \ldots, n+1$ are weighted exchangeable with $\omega_i \equiv 1$, $i = 1, \ldots, n$ and $\omega_{n+1}((x, y)) = q(y)/p(y)$, for any pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Let $r_{n+1} := r\left(X_{n+1}, Y_{n+1}, U_{n+1}; \widehat{\pi}\right)$. By construction $Y_{n+1} \in \mathcal{F}_{\tau^\star}^{(w)}\left(X_{n+1}, U_{n+1}; \widehat{\pi}\right)$ if and only if:

$$r_{n+1} \leq Q_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{p}_i^w(Y_{n+1})\delta_{r_i} + \tilde{p}_{n+1}^w(Y_{n+1})\delta_1\right).$$

Under label shift assumption, weights (22) do simplify as

$$\begin{aligned} p_i^w(Z_1, \ldots, Z_{n+1}) &= \frac{\sum_{\sigma:\sigma(n+1)=i} w_{n+1}(Z_i)}{\sum_\sigma w_{n+1}(Z_{\sigma(n+1)})} \\ &= \frac{w(Y_i)}{\sum_{j=1}^{n} w(Y_j) + w(Y_{n+1})} \\ &= \tilde{p}_i^w(Y_{n+1}), \end{aligned}$$

Figure 7. Prediction sets corresponding to the randomized conformal (scheme 2) method (15). See Section B.2 for more details.

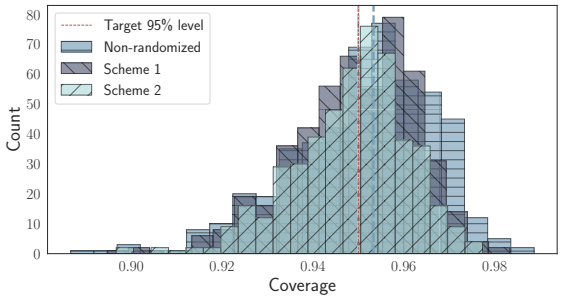

Figure 8. Average marginal coverage of conformal prediction sets for the simulation in Section B.2.

for $i = 1, \ldots, n+1$ matching the ones stated in (7). The result follows by invoking Lemma 6. As $\widehat{\pi}$ is fixed at the calibration step being pre-computed on a separate part of the dataset split, the result is conditional on $\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}$. □

Proof of Corollary 1. As for the other results, here it is also conditional on the training data, and thus we omit writing $\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}$ for succinctness and we use $r_{n+1} = r\left(X_{n+1}, Y_{n+1}, U_{n+1}; \widehat{\pi}\right)$ to denote the radius for the test point. Choose an arbitrary $\varepsilon > 0$. We have:

$$\begin{aligned} & \mathbb{P}\left(Y_{n+1} \notin \mathcal{F}_{\tau^\star}^{(\widehat{w}_k)}\left(X_{n+1}, U_{n+1}; \widehat{\pi}\right)\right) \\ = \; & \mathbb{P}\left(r_{n+1} > \tau_{\widehat{w}_k}^\star(Y_{n+1})\right) \\ = \; & \mathbb{P}\left(\left\{r_{n+1} > \tau_{\widehat{w}_k}^\star(Y_{n+1})\right\} \cap \left\{r_{n+1} + \varepsilon > \tau_w^\star(Y_{n+1})\right\}\right) \\ + \; & \mathbb{P}\left(\left\{r_{n+1} > \tau_{\widehat{w}_k}^\star(Y_{n+1}\right\} \cap \left\{r_{n+1} + \varepsilon \leq \tau_w^\star(Y_{n+1})\right\}\right). \end{aligned} \tag{16}$$

We have that:

$$\mathbb{P}\left(r_{n+1} \geq \tau_w^\star(Y_{n+1})\right) = \mathbb{P}\left(r_{n+1} > \tau_w^\star(Y_{n+1})\right) < \alpha,$$

where equality is due to the fact that r_{n+1} in the randomized scheme has a continuous distribution and inequality is due

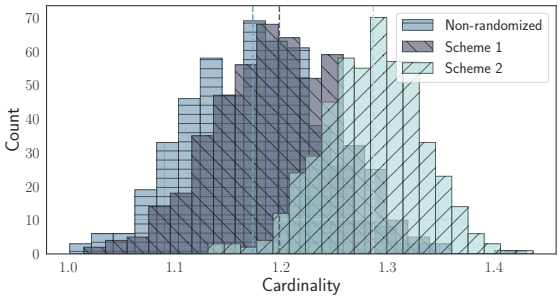

Figure 9. Average cardinality of conformal prediction sets for the simulation in Section B.2.

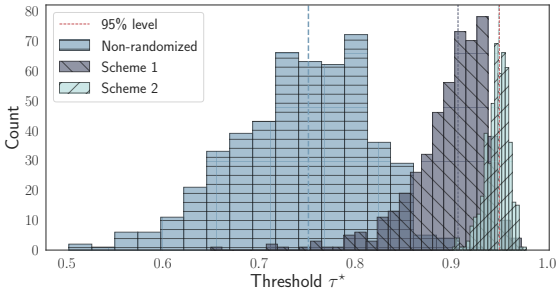

Figure 10. Learned cut-off thresholds in each setting (appending empty prediction sets with the most-likely label does not impact the threshold) for conformal prediction sets for the simulation in Section B.2.

to Theorem 1. For the first term in (16) we have:

$$\begin{aligned} & \mathbb{P}\left(\left\{r_{n+1} > \tau^\star_{\widehat{w}_k}(Y_{n+1}\right\} \cap \left\{r_{n+1} + \varepsilon > \tau^\star_w(Y_{n+1})\right\}\right) \\ = \;& \mathbb{P}\left(\left\{r_{n+1} > \tau^\star_{\widehat{w}_k}(Y_{n+1}\right\} \cap \left\{r_{n+1} > \tau^\star_w(Y_{n+1}) - \varepsilon\right\}\right) \\ \leq \;& \mathbb{P}\left(r_{n+1} > \tau^\star_w(Y_{n+1}) - \varepsilon\right), \end{aligned}$$

and for the second term we have that:

$$\begin{aligned} & \mathbb{P}\left(\left\{r_{n+1} > \tau^\star_{\widehat{w}_k}(Y_{n+1})\right\} \cap \left\{r_{n+1} \leq \tau^\star_w(Y_{n+1}) - \varepsilon\right\}\right) \\ \leq \;& \mathbb{P}\left(\left|\tau^\star_{\widehat{w}_k}(Y_{n+1}) - \tau^\star_w(Y_{n+1})\right| \geq \varepsilon\right). \end{aligned}$$

Note that ε was chosen arbitrarily, so we can let $\varepsilon \to 0$. By the continuous mapping theorem, consistency of $\widehat{w}_k$ implies that of $\tau^\star_{\widehat{w}_k}(y)$, $y \in \mathcal{Y}$. Thus,

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

which concludes the proof of the Corollary. □

Remark 1. To demonstrate why presence of a jump might cause problems, consider a simplified example. Let $Z \sim \mathrm{Ber}(p)$ for which the quantile corresponding to any given level α is given by

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

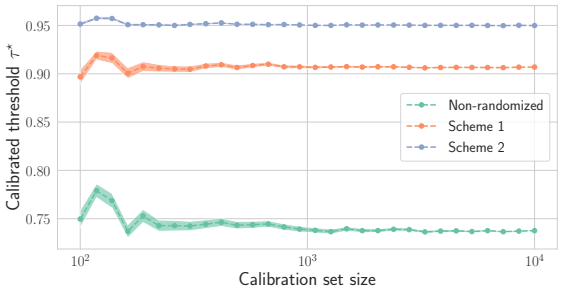

Figure 11. Learned cut-off thresholds in each setting when increasing the size of the calibration set for conformal prediction sets for the simulation in Section B.2.

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 $\widehat{q}_n \sim \mathrm{Ber}\left(\mathbb{P}\left(\overline{Z}_n > 1-\alpha\right)\right)$. 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 $\mathrm{Ber}(1/2)$, and thus the estimator will not be consistent. Still, for a more general setting of the distribution defined in (6) it is reasonable to expect the assumption regarding absence of jumps to be satisfied as also confirmed by our conducted empirical study.

B.4 Necessity of accounting for label shift

To illustrate the necessity of accounting for label shift in conformal classification, 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 $\mu_1 = (-2; 0)^\top$, $\mu_2 = (2; 0)^\top$, $\mu_3 = \left(0; 2\sqrt{3}\right)^\top$, $\Sigma = \mathrm{diag}(4, 4)$. First, we perform the standard routine for constructing split-conformal prediction sets for a 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 12 and the resulting prediction sets

on Figure 13. Next, we repeat the simulation 1000 times and track empirical coverage on the test set. Results on Figure 14 demonstrate the necessity of correcting for label shift as the classic (exchangeable) conformal prediction sets fail to achieve the correct marginal coverage.

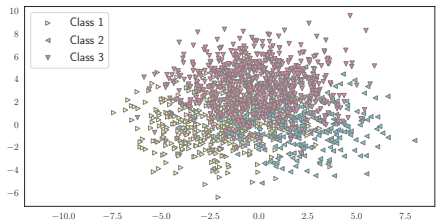

Figure 12. Test data sample for the toy simulation in Section B.4.

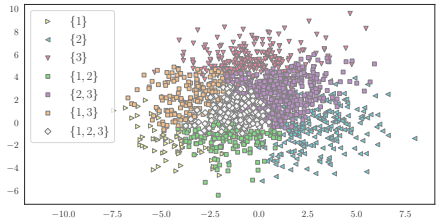

Figure 13. Conformal prediction sets when label shift is accounted with oracle importance weights for the toy simulation in Section B.4.

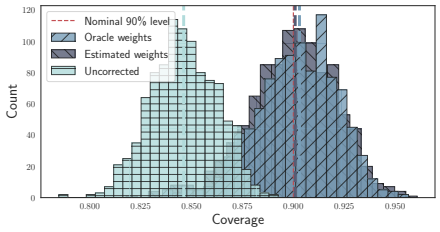

Figure 14. Empirical coverage on shifted data for the toy simulation in Section B.4. Dashed vertical lines describe the median coverage values, which are significantly worse when label shift is not accounted for, while using estimated weights mimics the oracle reasonably well.

B.5 Simulation on real data

For the simulation in Section 2.1 we use `wine quality` dataset (Cortez et al., 2009) to illustrate the performance of the conformal prediction sets when label shift is (not) taken into account. We focus on white wine dataset only, which has 4898 instances with 11 features and construct a 3-class classification problem by keeping classes 5,6,7 only to avoid complications arising due to high imbalance in the dataset (less than 10% of the data points were removed). Other important aspects include

1. 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.

2. **Data Split:** First, the original dataset $\mathcal{D}$ is split into two disjoint and approximately equal sets $\mathcal{D}_1$ and $\mathcal{D}_2$. Then label shift is simulated via resampling according to considered class proportions yielding $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$ of the same size. Finally, the former dataset is split at random into sets for training ($\approx$ 1000 instances), calibration ($\approx$ 100 instances) and importance weights estimation ($\approx$ 700 instances) and the latter is split is split at random into importance weight estimation ($\approx$ 1000 instances; recall that only labels from the target are used) and test ($\approx$ 1600 instances) sets.

3. **Model:** We use a standard Feed Forward Neural Network with 3 hidden layers with (128,64,32) neurons and ℓ_2-regularization in each as an underlying model. We use Adam optimizer with default parameters, set the maximum number of training epochs to 500 and deploy Early Stopping with patience for 25 epochs.

4. **Estimating label shift:** We use BBSE-soft (Lipton et al., 2018) for estimating importance weights.

B.6 Marginal (standard) conformal versus label-conditional conformal

Various procedures of performing label-conditional conformal prediction have been proposed in a series of works (Vovk et al., 2005; 2016; Sadinle et al., 2019; Guan and Tibshirani, 2019). Those are based on a slight modification of the standard conformal p-value used to determine whether there is enough evidence to exclude given label from the prediction set. Roughly speaking, for each candidate label y instead of looking whether a pair (X_{n+1}, y) conforms well to the whole collection of points $\widetilde{\mathcal{D}} = \{(X_i, Y_i)\}_{i \in \mathcal{I}}$, one analyzes only the subcollection that shares the same label y. Since the standard exchangeability argument immediately implies validity, the difference then lies in a particular choice for the underlying (non-)conformity score. For example, one could design a score that aims to minimize expected size of the prediction set (Sadinle et al., 2019; Guan and Tibshirani, 2019).

We now apply label-conditional split-conformal framework to the setting discussed in this work and focus on the case of not well-separated data. Consider, for example, the data simulation pipeline from Section 2.1. First, we fix $\alpha_y = \alpha = 0.1$ for all $y \in \mathcal{Y}$ and illustrate the difference between label-conditional conformal (8) and standard conformal (5)

prediction sets with the same randomized non-conformity scores (4) for a fair comparison on Figures 15, 16. In both cases a shallow MLP (two layers with 100 hidden units in each) is used. In this example a stronger requirement of conditional validity forces many prediction sets to be larger in a non-negligible area and to contain the least populated class 1.

Then we perform 1000 simulations and compare label-conditional conformal against marginal conformal in two settings (in all cases prediction sets are forced to contain at least the most likely label for a fair comparison). First, we set the calibration set size to be ≈ 350 data points and compare two procedures depending on whether class proportions change, and in the former case we perform reweighting of the non-conformity scores as described in Section 2.1. On Figures 17, 18 we observe that when class proportions do not change label-conditional conformal yields larger prediction sets as opposed to standard marginal conformal due to a stronger coverage requirement. However, when class proportions change, after performing the reweighting with the true label likelihood ratios, both procedures output prediction sets of similar size on average as illustrated on Figure 19, 20. Motivated by the fact that in practice one does not have infinite data resources, and thus keeping a sufficiently large held-out set per label could become prohibitive, we also consider a setting when the calibration set contains ≈ 100 data points (total). Smaller calibration set size results in losses of statistical power when testing whether a given label should be included into the prediction set, and thus, might yield larger prediction sets as we observe on Figure 21, 22.

To summarize, label-conditional conformal is a complementary (and powerful) technique to the one described in this work that inherently adapts to label shift. It does not require knowledge or estimation of the importance weights but has certain limitations: (a) it might be potentially a bit conservative in certain areas of the sample space when classes overlap, (b) it requires further splitting of the calibration data based on labels that could result in the loss of statistical power, especially when the number of classes K is large, which a common setting for the modern datasets.

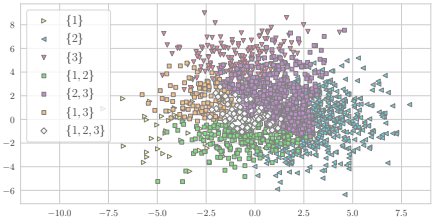

Figure 15. Conformal prediction sets with marginal coverage guarantee.

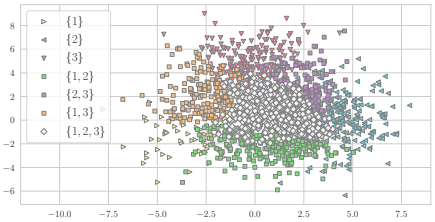

Figure 16. Conformal prediction sets with class-specific coverage guarantee. Stronger coverage comes at the price of larger the prediction sets in certain areas.

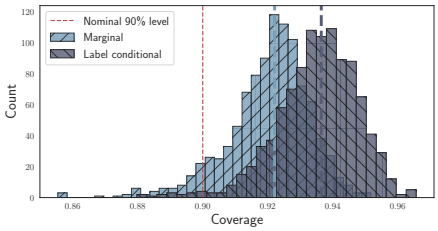

Figure 17. Empirical coverage of conformal prediction sets on source distribution and ≈ 350 calibration data points total. See Section B.6 for more details.

C Calibration

Section C.1 includes all proofs for Sections 3.1 and 3.2. Section C.2 includes details about the simulation on a toy example that illustrates the necessity of accounting for label shift. Section C.3 includes details about the simulation on a real dataset mentioned in Section 3.2.

C.1 Proofs

Theorem 3. *Fix $\alpha \in (0,1)$. With probability at least $1-\alpha$, $\left\|\widehat{\pi}_m^P - \pi_m^P\right\|_1 \leq \varepsilon_m$, 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.

Proof of Theorem 3. Recall that $g : \mathcal{X} \to \mathcal{M}$ denotes the bin-mapping function. Let E be the event that $(g(X_1), \ldots, g(X_n)) = (g(x_1), \ldots, g(x_n))$. On this event, the number of calibration points N_m within each bin B_m is known and for each bin labels are i.i.d.

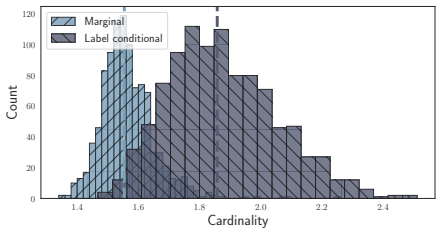

Figure 18. Average cardinality of conformal prediction sets on the source distribution and ≈ 350 calibration data points total. See Section B.6 for more details.

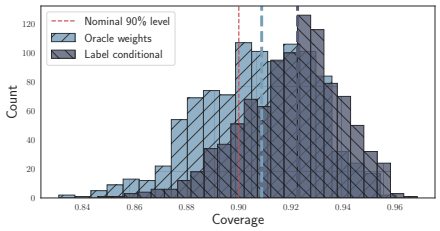

Figure 19. Empirical coverage of conformal prediction sets on target distribution and ≈ 350 calibration data points total. See Section B.6 for more details.

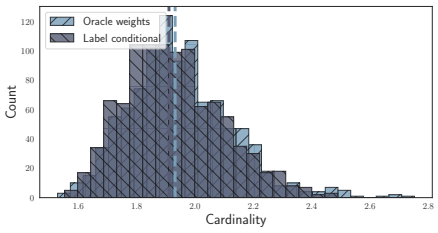

Figure 20. Average cardinality of conformal prediction sets on target distribution and ≈ 350 calibration data points total. See Section B.6 for more details.

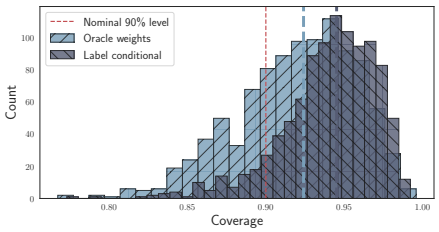

Figure 21. Empirical coverage of conformal prediction sets on target distribution and ≈ 100 calibration data points total. See Section B.6 for more details.

with corresponding class probabilities given by $\pi^P_{y,m} = \mathbb{P}(Y = y \mid f(X) \in B_m)$ for all $y \in \mathcal{Y}$. Thus, a vector corresponding of label frequencies has multinomial distribution with parameters N_m and $\left\{\pi^P_{y,m}\right\}_{y\in\mathcal{Y}}$. Theorem 4 yields that conditional on E

$$\sum_{y=1}^{K} \left|\widehat{\pi}^P_{y,m} - \pi^P_{y,m}\right| \geq \frac{2}{\sqrt{N_m}} \sqrt{\frac{1}{2} \ln\left(\frac{M2^K}{\alpha}\right)},$$

with probability at most α/M. Invoking union bound, we get that, conditional on E, with probability at least $1-\alpha$,

$$\sum_{y=1}^{K} \left|\widehat{\pi}^P_{y,m} - \pi^P_{y,m}\right| \leq \frac{2}{\sqrt{N_m}} \sqrt{\frac{1}{2} \ln\left(\frac{M2^K}{\alpha}\right)},$$

simultaneously for all $m \in \mathcal{M}$. Since it is true for any E, we can marginalize to obtain the first assertion of the Proposition. The second assertion simply represents a consideration of the case when multiple bins happen to have the same calibrated output which is needed to state the desired calibration guarantee. Let

$$\varepsilon^\star = \sup_{m\in\mathcal{M}} \varepsilon_m$$

denote the worst-case bound. Note that $\varepsilon^\star$ is in fact random and to be fully rigorous we, first, perform next steps conditional on E and then marginalize to obtain the assertion.

Now, for any $y \in \mathcal{Y}$:

$$\begin{aligned}
& \left|\mathbb{P}(Y = y \mid h(X)) - h_y(X)\right| \\
= \; & \left|\mathbb{E}\left[\mathbb{1}\{Y = y\} \mid h(X)\right] - h_y(X)\right| \\
\overset{(a)}{=} \; & \left|\mathbb{E}\left[\mathbb{1}\{Y = y\} \mid h(X)\right] - \mathbb{E}\left[h_y(X) \mid h(X)\right]\right| \\
\overset{(b)}{=} \; & \left|\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{Y = y\} \mid g(X)\right] \mid h(X)\right] - \mathbb{E}\left[h_y(X) \mid h(X)\right]\right| \\
\overset{(c)}{=} \; & \left|\mathbb{E}\left[\left[\pi^P_{y,g(X)} - h_y(X)\right] \mid h(X)\right]\right| \\
\overset{(d)}{\leq} \; & \mathbb{E}\left[\left|\pi^P_{y,g(X)} - \widehat{\pi}_{y,g(X)}\right| \mid h(X)\right],
\end{aligned} \tag{17}$$

where $(a), (b)$ are due to the tower rule (h is a function of g), (c) is due to linearity of conditional expectation and due to definition of $\pi^P_{y,m}$ and, finally, (d) is due to Jensen's inequality. Consider the event:

$$E_1 : \quad \left\|\widehat{\pi}^P_m - \pi^P_m\right\|_1 \leq \varepsilon_m,$$

simultaneously for all $m \in \mathcal{M}$. Note that the first assertion of the Proposition states event E_1 happens with probability at least $1-\alpha$ for chosen ε_m: $\mathbb{P}(E_1) \geq 1-\alpha$. Let E_2 be the following event:

$$E_2 : \quad \sum_{y=1}^{K} \left|\mathbb{P}(Y = y \mid h(X)) - h_y(X)\right| \leq \varepsilon^\star.$$

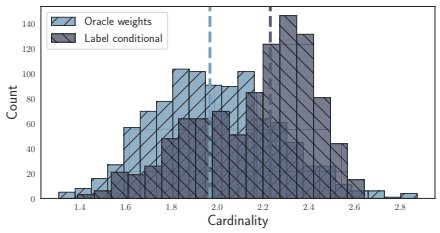

Figure 22. Average cardinality of conformal prediction sets on target distribution and ≈ 100 calibration data points total. See Section B.6 for more details.

Summing up over labels $y \in \mathcal{Y}$, (17) yields that on E_1 it holds with probability 1:

$$\begin{aligned} & \sum_{y=1}^{K} \left| \mathbb{P}\left(Y = y \mid h(X)\right) - h_y(X) \right| \\ \leq \;& \mathbb{E}\left[\left\| \pi^P_{g(X)} - \widehat{\pi}_{g(X)} \right\|_1 \mid h(X) \right] \\ \leq \;& \mathbb{E}\left[\varepsilon^\star \mid h(X) \right] = \varepsilon^\star, \end{aligned}$$

since $\varepsilon^\star$ is a constant. We get that $E_1 \subseteq E_2$, and thus $\mathbb{P}(E_2) \geq \mathbb{P}(E_1)$, and the assertion of the Proposition follows. □

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:*

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

Proof of Proposition 1. The Proposition is a straightforward combination of the Bayes rule and label shift assumption. Given a predictor f, for any class label $y \in \mathcal{Y}$ and any bin B_m, $m \in \mathcal{M} = \{1, \ldots, M\}$ one can equivalently represent conditional probabilities with respect to the target distribution as:

$$\begin{aligned} & \mathbb{P}_Q\left(Y = y \mid f(X) \in B_m\right) \\ \overset{(a)}{=} \;& \mathbb{P}_Q\left(f(X) \in B_m \mid Y = y\right) \cdot \frac{\mathbb{P}_Q\left(Y = y\right)}{\mathbb{P}_Q\left(f(X) \in B_m\right)} \\ \overset{(b)}{=} \;& \mathbb{P}_P\left(f(X) \in B_m \mid Y = y\right) \cdot \frac{\mathbb{P}_Q\left(Y = y\right)}{\mathbb{P}_Q\left(f(X) \in B_m\right)} \\ \overset{(c)}{=} \;& \mathbb{P}_P\left(Y = y \mid f(X) \in B_m\right) \\ & \cdot \frac{\mathbb{P}_Q\left(Y = y\right)}{\mathbb{P}_P\left(Y = y\right)} \cdot \frac{\mathbb{P}_P\left(f(X) \in B_m\right)}{\mathbb{P}_Q\left(f(X) \in B_m\right)} \\ = \;& \mathbb{P}_P\left(Y = y \mid X \in B_m\right) \cdot w(y) \cdot V_m, \end{aligned}$$

where $w(y)$ is the importance weight of label y and V_m is the 'relative volume' of bin B_m. Steps (a), (c) are due to the Bayes rule, (b) is due to label shift assumption. Normalization: $\sum_{k=1}^{K} \mathbb{P}_Q\left(Y = k \mid f(X) \in B_m\right) = 1$, implies:

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

Thus for all bins $m \in \mathcal{M}$ and labels $y \in \mathcal{Y}$ it holds:

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

which concludes the proof of the Proposition. □

Proof of Theorem 2. By triangle inequality, one obtains that for any bin $m \in \mathcal{M}$:

$$\begin{aligned} & \sum_{y=1}^{K} \left| \widehat{\pi}^{(\widehat{w})}_{y,m} - \pi^Q_{y,m} \right| \\ \leq \;& \sum_{y=1}^{K} \left| \widehat{\pi}^{(w)}_{y,m} - \pi^Q_{y,m} \right| + \sum_{y=1}^{K} \left| \widehat{\pi}^{(\widehat{w})}_{y,m} - \widehat{\pi}^{(w)}_{y,m} \right|. \end{aligned} \tag{18}$$

Consider the first term in (18). For any $y \in \mathcal{Y}$:

$$\begin{aligned} & \left| \widehat{\pi}^{(w)}_{y,m} - \pi^Q_{y,m} \right| \\ = \;& \left| \frac{w(y) \cdot \widehat{\pi}^P_{y,m}}{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}^P_{k,m}} - \frac{w(y) \cdot \pi^P_{y,m}}{\sum_{l=1}^{K} w(l) \cdot \pi^P_{l,m}} \right| \\ = \;& w(y) \cdot \left| \frac{\widehat{\pi}^P_{y,m}}{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}^P_{k,m}} - \frac{\pi^P_{y,m}}{\sum_{l=1}^{K} w(l) \cdot \pi^P_{l,m}} \right| \\ = \;& w(y) \cdot \left| \frac{\widehat{\pi}^P_{y,m}}{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}^P_{k,m}} - \frac{\pi^P_{y,m} - \widehat{\pi}^P_{y,m} + \widehat{\pi}^P_{y,m}}{\sum_{l=1}^{K} w(l) \cdot \pi^P_{l,m}} \right| \\ \overset{(a)}{\leq} \;& w(y) \cdot \widehat{\pi}^P_{y,m} \cdot \left| \frac{1}{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}^P_{k,m}} - \frac{1}{\sum_{l=1}^{K} w(l) \cdot \pi^P_{l,m}} \right| \\ + \;& w(y) \cdot \left| \frac{\pi^P_{y,m} - \widehat{\pi}^P_{y,m}}{\sum_{l=1}^{K} w(l) \cdot \pi^P_{l,m}} \right|, \end{aligned}$$

where (a) is due to triangle inequality. We infer that:

$$\begin{aligned}
&\sum_{y=1}^{K}\left|\widehat{\pi}_{y,m}^{(w)}-\pi_{y,m}^{Q}\right| \\
\leq\ & \left|1-\frac{\sum_{k=1}^{K}w(k)\cdot\widehat{\pi}_{k,m}^{P}}{\sum_{l=1}^{K}w(l)\cdot\pi_{l,m}^{P}}\right| \\
+\ & \frac{\sum_{y=1}^{K}w(y)\left|\pi_{y,m}^{P}-\widehat{\pi}_{y,m}^{P}\right|}{\sum_{l=1}^{K}w(l)\cdot\pi_{l,m}^{P}} \\
=\ & \frac{\left|\sum_{k=1}^{K}w(k)\cdot\left(\widehat{\pi}_{k,m}^{P}-\pi_{l,m}^{P}\right)\right|}{\sum_{l=1}^{K}w(l)\cdot\pi_{l,m}^{P}} \\
+\ & \frac{\sum_{y=1}^{K}w(y)\left|\pi_{y,m}^{P}-\widehat{\pi}_{y,m}^{P}\right|}{\sum_{l=1}^{K}w(l)\cdot\pi_{l,m}^{P}} \\
\overset{(a)}{\leq}\ & 2\cdot\frac{\sum_{y=1}^{K}w(y)\left|\pi_{y,m}^{P}-\widehat{\pi}_{y,m}^{P}\right|}{\sum_{l=1}^{K}w(l)\cdot\pi_{l,m}^{P}} \\
\overset{(b)}{\leq}\ & 2\cdot\frac{\left(\sup_{k}w(k)\right)\cdot\sum_{y=1}^{K}\left|\pi_{y,m}^{P}-\widehat{\pi}_{y,m}^{P}\right|}{\sum_{l=1}^{K}w(l)\cdot\pi_{l,m}^{P}},
\end{aligned}$$

where (a) is due to triangle inequality and (b) is due to Hölder's inequality. Observe that for any $m\in\mathcal{M}$:

$$\begin{aligned}
&\frac{1}{\sum_{k=1}^{K}w(k)\cdot\pi_{k,m}^{P}} \\
\leq\ & \frac{1}{\left(\inf\limits_{k:w(k)\neq 0}w(k)\right)\cdot\sum_{l=1}^{K}\pi_{l,m}^{P}} \\
=\ & \frac{1}{\inf\limits_{k:w(k)\neq 0}w(k)},
\end{aligned}$$

as $\sum_{l=1}^{K}\pi_{l,m}^{P}=1,\forall m\in\mathcal{M}$. Hence, for any $m\in\mathcal{M}$,

$$\begin{aligned}
&\sum_{y=1}^{K}\left|\widehat{\pi}_{y,m}^{(w)}-\pi_{y,m}^{Q}\right| \\
\leq\ & 2\cdot\frac{\sup_{k}w(k)}{\inf\limits_{k:w(k)\neq 0}w(k)}\cdot\sum_{y=1}^{K}\left|\pi_{y,m}^{P}-\widehat{\pi}_{y,m}^{P}\right|.
\end{aligned}\tag{19}$$

Now, consider the second term in (18). Observe that:

$$\begin{aligned}
&\left|\widehat{\pi}_{y,m}^{(\widehat{w})}-\widehat{\pi}_{y,m}^{(w)}\right| \\
=\ & \left|\frac{\widehat{w}(y)\cdot\widehat{\pi}_{y,m}^{P}}{\sum_{k=1}^{K}\widehat{w}(k)\cdot\widehat{\pi}_{k,m}^{P}}-\frac{w(y)\cdot\widehat{\pi}_{y,m}^{P}}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}}\right| \\
=\ & \widehat{\pi}_{y,m}^{P}\cdot\left|\frac{\widehat{w}(y)}{\sum_{k=1}^{K}\widehat{w}(k)\cdot\widehat{\pi}_{k,m}^{P}}-\frac{w(y)}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}}\right| \\
=\ & \widehat{\pi}_{y,m}^{P}\cdot\left|\frac{\widehat{w}(y)}{\sum_{k=1}^{K}\widehat{w}(k)\cdot\widehat{\pi}_{k,m}^{P}}-\frac{w(y)-\widehat{w}(y)+\widehat{w}(y)}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}}\right| \\
\overset{(a)}{\leq}\ & \left|\frac{1}{\sum_{k=1}^{K}\widehat{w}(k)\cdot\widehat{\pi}_{k,m}^{P}}-\frac{1}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}}\right|\cdot\widehat{\pi}_{y,m}^{P}\cdot\widehat{w}(y) \\
+\ & \frac{\widehat{\pi}_{y,m}^{P}\cdot|w(y)-\widehat{w}(y)|}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}},
\end{aligned}$$

where (a) is due to triangle inequality. Thus,

$$\begin{aligned}
&\sum_{y=1}^{K}\left|\widehat{\pi}_{y,m}^{(\widehat{w})}-\widehat{\pi}_{y,m}^{(w)}\right| \\
\leq\ & \left|\frac{1}{\sum_{k=1}^{K}\widehat{w}(k)\cdot\widehat{\pi}_{k,m}^{P}}-\frac{1}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}}\right|\cdot\sum_{y=1}^{K}\widehat{\pi}_{y,m}^{P}\cdot\widehat{w}(y) \\
+\ & \frac{\sum_{y=1}^{K}\widehat{\pi}_{y,m}^{P}\cdot|w(y)-\widehat{w}(y)|}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}} \\
=\ & \left|1-\frac{\sum_{y=1}^{K}\widehat{w}(y)\cdot\widehat{\pi}_{y,m}^{P}}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}}\right| \\
+\ & \frac{\sum_{y=1}^{K}\widehat{\pi}_{y,m}^{P}\cdot|w(y)-\widehat{w}(y)|}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}} \\
=\ & \frac{\left|\sum_{y=1}^{K}\left(w(y)-\widehat{w}(y)\right)\cdot\widehat{\pi}_{y,m}^{P}\right|}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}} \\
+\ & \frac{\sum_{y=1}^{K}\widehat{\pi}_{y,m}^{P}\cdot|w(y)-\widehat{w}(y)|}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}} \\
\leq\ & \frac{2\left\|\widehat{w}-w\right\|_{\infty}}{\sum_{l=1}^{K}w(l)\cdot\widehat{\pi}_{l,m}^{P}},
\end{aligned}$$

since $\sum_{k=1}^{K}\widehat{\pi}_{k,m}^{P}=1,\forall m\in\mathcal{M}$. Similarly, for any $m\in\mathcal{M}$:

$$\begin{aligned}
&\frac{1}{\sum_{k=1}^{K}w(k)\cdot\widehat{\pi}_{k,m}^{P}} \\
\leq\ & \frac{1}{\left(\inf_{l:w(l)\neq 0}w(l)\right)\cdot\sum_{k=1}^{K}\widehat{\pi}_{k,m}^{P}}=\frac{1}{\inf_{l:w(l)\neq 0}w(l)}.
\end{aligned}$$

Thus, we get that for any $m\in\mathcal{M}$:

$$\sum_{y=1}^{K}\left|\widehat{\pi}_{y,m}^{(\widehat{w})}-\widehat{\pi}_{y,m}^{(w)}\right|\leq\frac{2\left\|\widehat{w}-w\right\|_{\infty}}{\inf_{l:w(l)\neq 0}w(l)}.\tag{20}$$

Combining bounds (19) and (20) with the bound (18), we obtain that for any $m \in \mathcal{M}$:

$$\sum_{y=1}^{K} \left| \widehat{\pi}^{(\widehat{w})}_{y,m} - \pi^{Q}_{y,m} \right| \leq 2\kappa \cdot \sum_{y=1}^{K} \left| \widehat{\pi}^{P}_{y,m} - \pi^{P}_{y,m} \right| + \frac{2 \left\| \widehat{w} - w \right\|_{\infty}}{\inf_{l: w(l) \neq 0} w(l)},$$

which concludes the proof of the Theorem. □

C.2 Necessity of accounting for label shift

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 2.1, 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

$$\pi^{P}_{1}(x) = \frac{p(1) \cdot \varphi(x; \mu_1, \Sigma)}{p(0) \cdot \varphi(x; \mu_0, \Sigma) + p(1) \cdot \varphi(x; \mu_1, \Sigma)}, \quad (21)$$

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 23, 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 (21) 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), which clearly does not happen.

C.3 Simulation on real data

For the simulation mentioned in Section 3.2 we use `wine quality` dataset (Cortez et al., 2009). The original dataset contains ratings for white wines and we reduce it to a binary classification problem by treating wine as good if the corresponding rating is at least 7 on a 10-point scale. Logistic regression is used as an underlying predictor and for each pass the original dataset $\mathcal{D}$ is, first, split into two disjoint and approximately equal sets $\mathcal{D}_1$ and $\mathcal{D}_2$. Label shift is simulated via resampling of $\widetilde{\mathcal{D}}_1$ with class proportions $p = (0.8, 0.2)$ and $\widetilde{\mathcal{D}}_2$ with class proportions $(0.5, 0.5)$. Final splitting resulted in $\approx$ 1350 instances used for both training and calibration, $\approx$ 700 and $\approx$ 400 instances used for importance weights estimation on the source and the target respectively and $\approx$ 1100 instances used for the test. Uniform-mass binning with 10 bins was used for calibration purposes. For 4 random data splits the resulting reliability curves are presented on Figures 24, 25, 26, 27 illustrating that calibration with proper reweighting leads to approximate calibration on the target domain and uncorrected fails to do so.

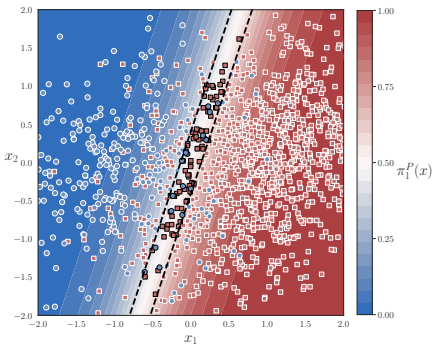

Figure 23. Sampled points from the target distribution plotted against the true source class-posterior probabilities.

D Auxiliary results

Note Lemma 5 and Lemma 6 were originally formulated for possibly unbounded non-conformity scores. It is easy to see that we can safely replace point masses δ_∞ by δ_1 in the conformal classification setting considered in this work.

Theorem 4 (Bretagnolle-Huber-Carol inequality (van der Vaart and Wellner, 1996))**.** *If the random vector $(N_1, \ldots, N_k)$ is multinomially distributed with parameters n and $(p_1, \ldots, p_k)$, then*

$$\mathbb{P}\left(\sum_{i=1}^{k} |N_i - np_i| \geq 2\sqrt{n}\lambda \right) \leq 2^k e^{-2\lambda^2}, \quad \lambda > 0.$$

Lemma 5 (Lemma 1 (Tibshirani et al., 2019))**.** *Assume $Z_1, \ldots, Z_{m+1}$ are exchangeable random variables sup-*

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

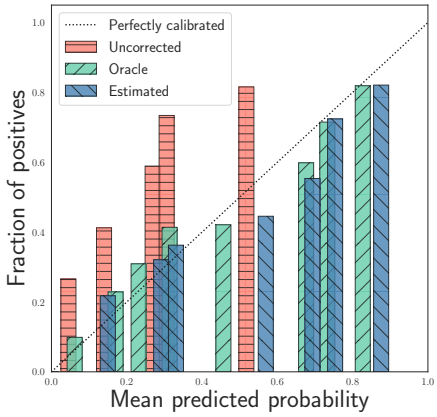

Figure 24. Reliability curve for the simulation on the `wine quality` dataset. Notice that the bars indicating calibration using oracle and estimated importance weights are quite similar to each other, but most importantly that both are very close to the ideal diagonal line (perfect calibration). In contrast, the uncorrected bars are poorly calibrated, demonstrating both the need for handling label shift and the relative success of our procedures in doing so. See Section C.3 for details.

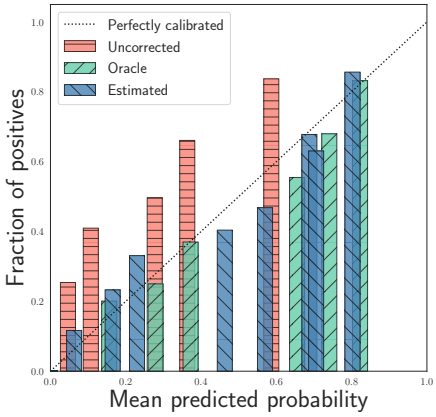

Figure 25. Reliability curve for the simulation on the `wine quality` dataset.

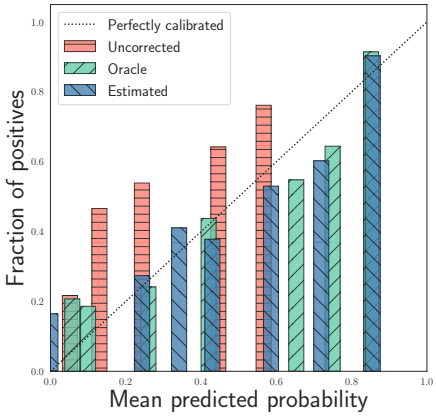

Figure 26. Reliability curve for the simulation on the `wine quality` dataset.

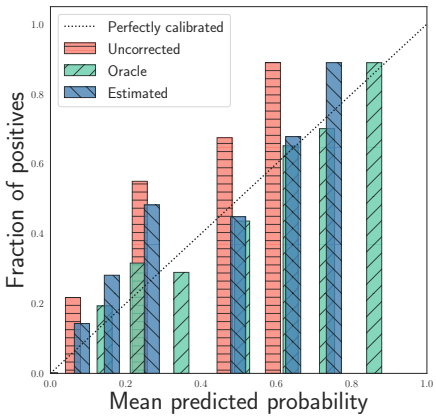

Figure 27. Reliability curve for the simulation on the `wine quality` dataset.

ported on $[0,1]$. *Then for any* $\beta \in (0,1)$,

$$\mathbb{P}\left(Z_{m+1} \leq Q_\beta\left(Z_{1:m} \cup \{1\}\right)\right) \geq \beta.^{3}$$

Moreover, if Z_i, $i = 1, \ldots, m+1$ *are almost surely distinct, then the above probability is upper bounded by* $\beta + \frac{1}{m+1}$.

Lemma 6 (Lemma 3 (Tibshirani et al., 2019))**.** *Let* Z_i, $i = 1, \ldots, n+1$ *be weighted exchangeable random variables with weight functions* $w_1, \ldots, w_{n+1}$ *and supported on* $[0,1]$. *Let* $V_i = S\left(Z_i, Z_{-i}\right)$, *where* $Z_{-i} = Z_{1:(n+1)} \backslash \{Z_i\}$, $i = 1, \ldots, n+1$ *and* S *is an arbitrary score function. Define*

$$p_i^w(z_1, \ldots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}{\sum_\sigma \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}, \tag{22}$$

for $i = 1, \ldots, n+1$, *where summations are taken over permutations* σ *of* $1, \ldots, n+1$. *Then for any* $\beta \in (0,1)$,

$$\mathbb{P}\left(V_{n+1} \leq Q_\beta\left(G_n\right)\right) \geq 1 - \beta,$$

where the distribution G_n *is defined as*

$$G_n := \sum_{i=1}^{n} p_i^w(Z_1, \ldots, Z_{n+1})\delta_{V_i} + p_{n+1}^w(Z_1, \ldots, Z_{n+1})\delta_1.$$

[3] In this case, $Q_\beta\left(Z_{1:m} \cup \{1\}\right)$ can be equivalently defined as the $\lceil \beta(m+1) \rceil$-th smallest element of the set $\{Z_i\}_{i=1}^m$ if $\beta \leq \frac{m}{m+1}$, and as 1 otherwise.