Transductive conformal inference with adaptive scores

Ulysse Gazin, Gilles Blanchard, Etienne Roquain

Transductive conformal inference with adaptive scores

Ulysse Gazin* Gilles Blanchard† Etienne Roquain‡

March 19, 2024

Abstract

Conformal inference is a fundamental and versatile tool that provides distribution-free guarantees for many machine learning tasks. We consider the transductive setting, where decisions are made on a test sample of m new points, giving rise to m conformal p-values. While classical results only concern their marginal distribution, we show that their joint distribution follows a Pólya urn model, and establish a concentration inequality for their empirical distribution function. The results hold for arbitrary exchangeable scores, including *adaptive* ones that can use the covariates of the test+calibration samples at training stage for increased accuracy. We demonstrate the usefulness of these theoretical results through uniform, in-probability guarantees for two machine learning tasks of current interest: interval prediction for transductive transfer learning and novelty detection based on two-class classification.

1 Introduction

Conformal inference is a general framework aiming at providing sharp uncertainty quantification guarantees for the output of machine learning algorithms used as "black boxes". A central tool of that field is the construction of a "(non)-conformity score" S_i for each sample point. The score functions can be learnt on a training set using various machine learning methods depending on the task at hand. The scores observed on a data sample called "calibration sample" $\mathcal{D}_{\text{cal}}$ serve as references for the scores of a "test sample" $\mathcal{D}_{\text{test}}$ (which may or may not be observed, depending on the setting). The central property of these scores is that they are an exchangeable family of random variables.

1.1 Motivating tasks

To be more concrete, we start with two specific settings serving both as motivation and as application.

(PI) Prediction intervals: we observe $\mathcal{D}_{\text{cal}} = (X_1, Y_1), \ldots, (X_n, Y_n)$ a sample of i.i.d. variables with unknown distribution P, where $X_i \in \mathbb{R}^d$ is a regression covariate and $Y_i \in \mathbb{R}$ is the outcome. Given a new independent datum (X_{n+1}, Y_{n+1}) generated from P, the task is to build a prediction interval for Y_{n+1} given X_{n+1} and $\mathcal{D}_{\text{cal}}$. More generally, in the *transductive* conformal setting (Vovk, 2013), the task is repeated $m \geq 1$ times: given m new data points $\mathcal{D}_{\text{test}} = (X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$ i.i.d from P, build m prediction intervals for $Y_{n+1}, \ldots, Y_{n+m}$ given $X_{n+1}, \ldots, X_{n+m}$ and $\mathcal{D}_{\text{cal}}$.

*Université Paris Cité and Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation. Email: ugazin@lpsm.paris

†Université Paris Saclay, Institut Mathématique d'Orsay. Email: gilles.blanchard@universite-paris-saclay.fr

‡Sorbonne Université and Université Paris Cité, CNRS, Laboratoire de Probabilités, Statistique et Modélisation. Email: etienne.roquain@upmc.fr

(ND) Novelty detection: we observe $\mathcal{D}_{\text{cal}} = (X_1, \ldots, X_n)$, a sample of nominal data points in $\mathbb{R}^d$, drawn i.i.d. from an unknown (null) distribution P_0, and a test sample $\mathcal{D}_{\text{test}} = (X_{n+1}, \ldots, X_{n+m})$ of independent points in $\mathbb{R}^d$, each of which is distributed as P_0 or not. The task is to decide if each X_{n+i} is distributed as the training sample (i.e., from P_0) or is a "novelty".

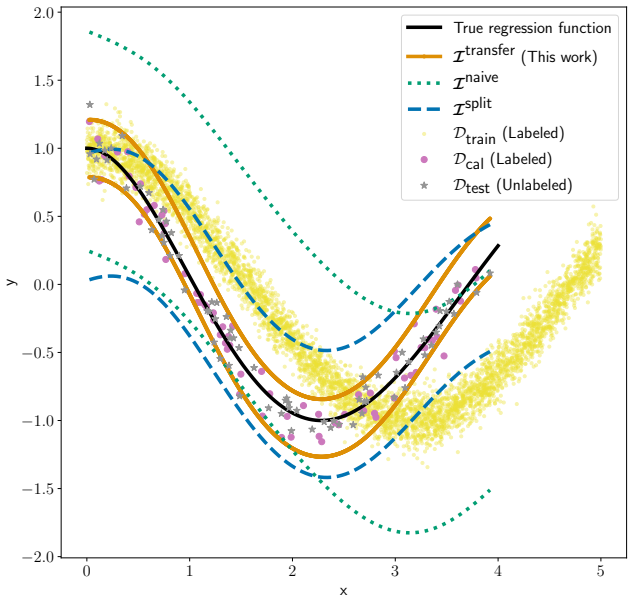

Figure 1: Task (PI) with adaptive scores in a non-parametric regression setting with domain shift between train and calibration+test samples (proof-of-concept model, see Section 3.5). Our contribution is both to propose adaptive scores and predictions relying on transfer learning (this figure), and uniform bounds on the false coverage proportion, see Figure 2.

For both inference tasks, the usual pipeline is based on the construction of non-conformity real-valued scores $S_1, \ldots, S_{n+m}$ for each member of $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$, which requires an additional independent training sample $\mathcal{D}_{\text{train}}$ (in the so-called "split conformal" approach):

(PI) the scores are (for instance) the regression residuals $S_i = |Y_i - \mu(X_i; \mathcal{D}_{\text{train}})|$, $1 \leq i \leq n+m$, where the function $\mu(x; \mathcal{D}_{\text{train}})$ is a point prediction of Y_i given $X_i = x$, learnt from the sample $\mathcal{D}_{\text{train}}$.

(ND) the scores are of the form $S_i = g(X_i; \mathcal{D}_{\text{train}})$, $1 \leq i \leq n+m$, where the score function $g(\cdot; \mathcal{D}_{\text{train}})$ is learnt using the sample $\mathcal{D}_{\text{train}}$; $g(x)$ is meant to be large if x is fairly different from the members of $\mathcal{D}_{\text{train}}$ (so that it is "not likely" to have been generated from P_0).

In both cases, inference is based on the so-called split conformal p-values (Vovk et al., 2005):

$$p_i = (n+1)^{-1}\left(1 + \sum_{j=1}^{n} \mathbf{1}\{S_j \geq S_{n+i}\}\right), \quad i \in [\![m]\!]. \tag{1}$$

In other words, $(n+1)p_i$ is equal to the rank of S_{n+i} in the set of values $\{S_1, \ldots, S_n, S_{n+i}\}$, and a small p-value p_i indicates that the test score S_{n+i} is abnormally high within the set of reference scores. The link to the two above tasks is as follows: for (PI), the prediction interval $\boldsymbol{\mathcal{C}}(\alpha)$ for Y_{n+i} with coverage probability $(1-\alpha)$ is obtained by inverting the inequality $p_i > \alpha$ w.r.t. Y_{n+i}, see (13) below. For (ND), the members of the test sample declared as novelties are those with a p-value $p_i \leq t$ for some threshold t.

Studying the behavior of the conformal p-value family is thus a cornerstone of conformal inference. Still, classical results only concern the *marginal* distribution of the p-values while the joint distribution remains largely unexplored in full generality.

1.2 Contributions and overview of the paper

In Section 2, we present new results for the joint distribution of the conformal p-values (1) for general exchangeable scores (for any sample sizes n and m). First, in Section 2.2, we show that the dependence structure involved only depends on n and m, and follows a Pólya urn model; this entails both explicit formula and useful characterizations. Second, we deduce a new finite sample DKW-type concentration inequality (Massart, 1990) for the empirical distribution function (ecdf) of the conformal p-values.

We illustrate the interest of the theoretical results through the application cases (PI) in Section 3 and (ND) in Section 4, for which dedicated numerical experiments are also provided[1].

Our theory provides *in-probability (i.e. confidence) bounds* for the error proportion when m decisions are taken simultaneously (transductive setting); furthermore, these bounds are *uniform* over a certain class of decisions. More precisely, the proportion of errors among the m decisions corresponds to the false coverage proportion (FCP) for (PI), resp. the false discovery proportion (FDP) for (ND). We develop upper confidence bounds for these quantities, in dependence of prediction interval length for (PI), resp. the rejection threshold for (ND), and valid *uniformly* over the choice of these parameters. This is in contrast to marginal guarantees in previous literature only providing in-expectation guarantees of FCP/FDP at a fixed level α, and for specific procedures. Obtaining in-probability bounds for the FDP is a classical and active theme of multiple testing theory: in contrast to an in-expectation control, it takes into account the fluctuations of the error proportion. It thus brings more fine-grained reliability, while the uniform guarantee also offers the practitioner more flexibility for taking a data-driven decision that is still theoretically backed up. These guarantees can in particular be crucial when handling sensible data. Similarly, obtaining a sharp confidence bound on the actual (random) number of false inferences for (PI) for repeated decisions is much more informative than a bound on its expectation.

We insist that we only assume that the scores are exchangeable to obtain in-probability guarantees. Exchangeability is a classical assumption in conformal theory, though some recent works have sometimes dropped it in favor of i.i.d. scores. Deriving results under the weaker exchangeable assumption is crucial in the considered applications, because while the *data* is assumed i.i.d., we rely on *adaptive conformal scores* which depend not only on the training sample (arbitrarily), but also *on the calibration+test sample* in an exchangeable way. Adaptive scores offer superior performance in practice (see Figure 1 for our approach to transductive transfer PI), are indeed exchangeable (thus, our theory applies) but *not* i.i.d. This illustrates the interest to develop the joint distribution theory under the weaker exchangeable assumption *even* if the underlying data is assumed to be i.i.d., a standard setting (our results also hold if the data is only assumed exchangeable).

1.3 Relation to previous work

For fundamentals on conformal prediction, see Vovk et al. (2005); Balasubramanian et al. (2014). We only consider the *split conformal* approach, also named inductive conformal approach in the seminal work of Papadopoulos et al. (2002). The split conformal approach uses a separate training set but is considered the most practically amenable approach for big data (in contrast to the "full conformal" approach which can be sharper but computationally intractable).

The most important consequence of score exchangeability is that the marginal distribution of a conformal p-value is a discrete uniform under the joint (calibration and test) data distribution. There has been significant recent interest for the *conditional* distribution of a marginal p-value, conditional to the calibration sample, under the stronger assumption of i.i.d. scores. The corresponding results take the form of bounds on $\mathbb{P}(p_1 \leq t \mid \mathcal{D}_{\text{cal}})$ holding with high probability over $\mathcal{D}_{\text{cal}}$ (Vovk, 2012; Bian and Barber, 2022; Sarkar and Kuchibhotla, 2023; Bates et al., 2023, where in the two latter references the results are in addition uniformly valid in t). However, the i.i.d.

[1]The code used in all our experiments is made publicly available at `https://github.com/ulyssegazin/TransductiveAdaptive_CP`.

scores assumption prevents handling adaptive scores , for which only exchangeability is guaranteed; moreover, these works only handle a single prediction.

Simultaneous inference for the (PI) task has been proposed by Vovk (2013) (see also Saunders et al., 1999 for an earlier occurrence for one p-value with multiple new examples), referred to as transductive conformal inference. It includes a bound on the family-wise error rate (the probability of committing one or more PI errors) based on a Bonferroni-type correction. In the present work we allow the number of PI errors to be positive but aim at a tight control of this number in probability (uniformly valid over the choice of PI length).

Closest to our work, Marques F. (2023); Huang et al. (2023) analyze the false coverage proportion (FCP) of the usual prediction interval family $\boldsymbol{\mathcal{C}}(\alpha)$ repeated over m test points: the exact distribution of the FCP under data exchangeability is provided, and related in Marques F. (2023) to a Pòlya urn model with two colors. We show the more general result that the *full joint* distribution of $(p_1, \ldots, p_m)$ follows a Pòlya urn model with $(n+1)$ colors, which entails the result of Marques F. (2023); Huang et al. (2023) as a corollary (see Supplemental A). This brings substantial innovations: our bounds on FCP are *uniform* in α, and we provide both the exact joint distribution and an explicit non-asymptotic approximation via a DKW-type concentration bound. Finally, Bao et al. (2024) also established an in-expectation control of the FCP after a selection stage. By contrast, we provide FCP bounds in probability. In addition, while our theory is stated without selection stage, it can be applied to a permutation invariant selection, see Remark 3.2.

The (ND) setting is alternatively referred to as Conformal Anomaly Detection (see Chapter 4 of Balasubramanian et al., 2014). We specifically consider here the (transductive) setting of Bates et al. (2023) where the test sample contains novelties, and the corresponding p-values for 'novelty' entries are not discrete uniform but expected to be stochastically smaller. Due to strong connections to multiple testing, ideas and procedures stemming from that area can be adapted to address (ND), specifically by controlling the false discovery rate (FDR, the expectation of the FDP), such as as the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). Use of adaptive scores and corresponding FDR control has been investigated by Marandon et al. (2022). Our contribution with respect to that work comes from getting uniform and in-probability bounds for the FDP (rather than only in expectation, for the FDR).

2 Main results

2.1 Setting

We denote integer ranges using $[\![i]\!] = \{1, \ldots, i\}$, $[\![i, j]\!] = \{i, \ldots, j\}$. Let $(S_i)_{i \in [\![n+m]\!]}$ be real random variables corresponding to non-conformity scores, for which $(S_j)_{j \in [\![n]\!]}$ are the "reference" scores and $(S_{n+i})_{i \in [\![m]\!]}$ are the "test" scores. We assume

$$\text{The score vector } (S_i)_{i \in [\![n+m]\!]} \text{ is exchangeable.} \tag{Exch}$$

Under (Exch), the p-values (1) have super-uniform marginals (see, e.g., Romano and Wolf, 2005). In addition, the marginal distributions are all equal and uniformly distributed on $\{\ell/(n+1), \ell \in [\![n+1]\!]\}$ under the additional mild assumption:

$$\text{The score vector } (S_i)_{i \in [\![n+m]\!]} \text{ has no ties a.s.} \tag{NoTies}$$

While the marginal distribution is well identified, the joint distribution of the p-values is not well studied yet. In particular, we will be interested in the empirical distribution function of the p-value family, defined as

$$\widehat{F}_m(t) := m^{-1} \sum_{i=1}^{m} \mathbf{1}\{p_i \leq t\}, \quad t \in [0, 1]. \tag{2}$$

Note that the p-values are not i.i.d. under (Exch), so that most classical concentration inequalities, such as DKW's inequality (Massart, 1990), or Bernstein's inequality, cannot be directly used. Instead, we should take into account the specific dependence structure underlying these p-values.

2.2 Key properties

We start with a straightforward result, under the stronger assumption

$$\text{The variables } S_i, i \in [\![n+m]\!], \text{ are i.i.d.} \tag{IID}$$

For this, introduce, for any fixed vector $U = (U_1, \ldots, U_n) \in [0,1]^n$, the discrete distribution P^U on the set $\left\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\right\}$, defined as

$$P^U(\{\ell/(n+1)\}) = U_{(\ell)} - U_{(\ell-1)}, \quad \ell \in [\![n+1]\!], \tag{3}$$

where $0 = U_{(0)} \leq U_{(1)} \leq \cdots \leq U_{(n)} \leq U_{(n+1)} = 1$ are the increasingly ordered values of $U = (U_1, \ldots, U_n)$. In words, the n values of U divide the interval $[0,1]$ into $(n+1)$ distinct cells (labeled $\frac{\ell}{n+1}, \ell \in [\![n+1]\!]$), and P^U is the probability distribution of the label of the cell a Unif$[0,1]$ variable would fall into.

Note that P^U has for c.d.f.

$$F^U(x) = U_{(\lfloor (n+1)x \rfloor)}, \quad x \in [0,1]. \tag{4}$$

The following result can be considered as well known from previous literature (see, e.g., proof of Theorem 6 in Bates et al., 2023); we include a short proof for completeness.

Proposition 2.1. *Assume* (IID) *and* (NoTies) *and consider the p-values* $(p_i, i \in [\![m]\!])$ *given by* (1). *Then conditionally on* $\mathcal{D}_{cal} = (S_1, \ldots, S_n)$, *the p-values are i.i.d. of common distribution given by*

$$p_1 \,|\, \mathcal{D}_{cal} \sim P^U,$$

where $U = (U_1, \ldots, U_n) = \big(1 - F(S_1), \ldots, 1 - F(S_n)\big)$ *are pseudo-scores and* F *is the common c.d.f. of the scores of* $\mathcal{D}_{cal}$, *that is,* $F(s) = \mathbb{P}(S_1 \leq s)$, $s \in \mathbb{R}$. *In addition the pseudo-score vector* U *is i.i.d.* Unif$[0,1]$ *distributed.*

Proof sketch. The conditional distribution of p_i only depends on score ordering which is unambiguous due to (NoTies), and is thus invariant by monotone transformation of the scores by $(1 - F)$. Writing explicitly the cdf of p_i from the uniformly distributed transformed scores yields (4). See Supplemental C.1 for details.

In the literature, such a result is used to control the conditional failure probability $\mathbb{P}(p_1 \leq \alpha \,|\, \mathcal{D}_{\text{cal}})$ around its expectation (which is ensured to be smaller than, and close to, α) with concentration inequalities valid under an i.i.d. assumption (Bates et al., 2023; Sarkar and Kuchibhotla, 2023; Bian and Barber, 2023).

By integration over U, a direct consequence of Proposition 2.1 is that, under (IID) and (NoTies), and *unconditionally on* $\mathcal{D}_{cal}$, the family of conformal p-values $(p_i, i \in [\![m]\!])$ has the "universal" distribution $P_{n,m}$ on $[0,1]^m$ defined as follows:

$$P_{n,m} = \mathcal{D}\big(q_i, i \in [\![m]\!]\big), \text{ where} \tag{5}$$

$$\begin{cases} (q_1, \ldots, q_m | U) & \overset{\text{i.i.d.}}{\sim} P^U; \\ \text{and } U = (U_1, \ldots, U_n) & \overset{\text{i.i.d.}}{\sim} \text{Unif}([0,1]). \end{cases} \tag{6}$$

Our first result is to note that the latter holds beyond the i.i.d. assumption.

Proposition 2.2. *Assume* (Exch) *and* (NoTies), *then the family of p-values* $(p_i, i \in [\![m]\!])$ *given by* (1) *has joint distribution* $P_{n,m}$, *which is defined by* (5)-(6) *and is independent of the specific score distribution.*

Proof sketch. The joint distribution of the p-values only depends on the ranks of the $(n+m)$ scores. Since the scores have exchangeable distribution and (NoTies) holds, their ranks form a random permutation of $[\![n+m]\!]$. Thus, the same rank distribution (and consequently joint p-value distribution) is generated when the scores are i.i.d. Applying Proposition 2.1, the p-value distribution can be represented as (5)-(6). See also Supplemental C.2.

The next proposition is an alternative and useful characterization of the distribution $P_{n,m}$.

Proposition 2.3. *$P_{n,m}$ is the distribution of the colors of m successive draws in a standard Pólya urn model with $n+1$ colors labeled $\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\}$.*

Proposition 2.3 is proved in Supplemental A, where several explicit formulas for $P_{n,m}$ are also provided. We also show that this generalizes the previous works of Marques F. (2023); Huang et al. (2023).

Comparing Proposition 2.1 and Proposition 2.2, we see that having i.i.d. scores is more favorable because guarantees are valid conditionally on $\mathcal{D}_{\text{cal}}$ (with an explicit expression for $U = U(\mathcal{D}_{\text{cal}})$). However, as we will see in Sections 3 and 4, the class of exchangeable scores is much more flexible and includes adaptive scores, which can improve substantially inference sharpness in specific situations. For this reason, we work with the unconditional distribution as in Proposition 2.2 in the sequel.

2.3 Consequences

We now provide a DKW-type envelope for the empirical distribution function (2) of conformal p-values. Let us introduce the discretized identity function

$$I_n(t) = \lfloor (n+1)t \rfloor / (n+1) = \mathbb{E}\widehat{F}_m(t), \quad t \in [0,1], \tag{7}$$

and the following bound:

$$B^{\text{DKW}}(\lambda, n, m) := \mathbf{1}_{\{\lambda<1\}} \left[1 + \frac{2\sqrt{2\pi}\lambda\tau_{n,m}}{(n+m)^{1/2}} \right] e^{-2\tau_{n,m}\lambda^2}, \tag{8}$$

where $\tau_{n,m} := nm/(n+m) \in [(n \wedge m)/2, n \wedge m]$ is an "effective sample size".

Theorem 2.4. *Let us consider the process $\widehat{F}_m$ defined by (2), the discrete identity function $I_n(t)$ defined by (7), and assume (Exch) and (NoTies). Then we have for all $\lambda > 0$, $n, m \geq 1$,*

$$\mathbb{P}\Big(\sup_{t\in[0,1]} (\widehat{F}_m(t) - I_n(t)) > \lambda \Big) \leq B^{DKW}(\lambda, n, m). \tag{9}$$

In addition, $B^{DKW}(\lambda^{DKW}_{\delta,n,m}, n, m) \leq \delta$ for

$$\begin{aligned} \lambda^{DKW}_{\delta,n,m} &= \Psi^{(r)}(1); \\ \Psi(x) &= 1 \wedge \left(\frac{\log(1/\delta) + \log\left(1 + \sqrt{2\pi}\frac{2\tau_{n,m}x}{(n+m)^{1/2}}\right)}{2\tau_{n,m}} \right)^{1/2}, \end{aligned} \tag{10}$$

where $\Psi^{(r)}$ denotes the function Ψ iterated r times (for an arbitrary integer $r \geq 1$).

Proof sketch. Use the representation (6), apply the DKW inequality separately to $(U_1, \ldots, U_n)$ and to $(q_1, \ldots, q_m)$ conditional to U, and integrate over U. See supplemental Section C.4 for details (a slightly more accurate bound is also proposed).

In supplemental Section E, we illustrate the sharpness of the inequality (9).

Remark 2.5. The Simes inequality (Simes, 1986) is true for conformal p-values (Bates et al., 2023), which provides a different confidence envelope on $\widehat{F}_m$. A comparison with the new DKW bound (8) is provided in Supplemental G. It shows that the latter is sharper in a wide range of situations.

Remark 2.6. Since the distribution $P_{n,m}$ can be easily sampled from, $\lambda^{\text{DKW}}_{\delta,n,m}$ in (10) can be further improved by considering the sharper but implicit quantile

$$\begin{aligned} \lambda^{\text{num-DKW}}(\delta, n, m) &= \min\Big\{ x \geq 0 \,:\, \pi_{n,m,x} \leq \delta \Big\}, \text{with} \\ \pi_{n,m,x} &:= P_{n,m}\Big(\sup_{\ell \in [\![n+1]\!]} \Big(\widehat{F}_m\Big(\frac{\ell}{n+1}\Big) - \frac{\ell}{n+1} \Big) > x \Big). \end{aligned}$$

In addition, numerical confidence envelopes for $\widehat{F}_m$ with other shapes can be investigated. For instance, for any set $\mathcal{K} \subset [\![m]\!]$ of size K, we can calibrate thresholds $t_1, \ldots, t_K > 0$ such that

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{p} \sim P_{n,m}}(\forall k \in \mathcal{K},\, p_{(k+1)} > t_k) \\ &= \mathbb{P}_{\boldsymbol{p} \sim P_{n,m}}(\forall k \in \mathcal{K},\, \widehat{F}_m(t_k) \leq k/m) \geq 1 - \delta. \end{aligned} \tag{11}$$

A method is to start from a "template" one-parameter family $(t_k(\lambda))_{k \in \mathcal{K}}$ and then adjust λ to obtain the desired control (Blanchard et al., 2020; Li et al., 2022). This approach is developed in detail in Suppl. B.

3 Application to prediction intervals

In this section, we apply our results to build simultaneous conformal prediction intervals, with an angle towards adaptive scores and transfer learning.

3.1 Setting

Let us consider a conformal prediction framework for a regression task, see, e.g., Lei et al. (2018), with three independent samples of points (X_i, Y_i), where $X_i \in \mathbb{R}^d$ is the covariable and $Y_i \in \mathbb{R}$ is the outcome:

- Training sample $\mathcal{D}_{\text{train}}$: observed and used to build predictors;
- Calibration sample $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i), i \in [\![n]\!]\}$;
  observed and used to calibrate the size(s) of the prediction intervals;
- Test sample $\mathcal{D}_{\text{test}} = \{(X_{n+i}, Y_{n+i}), i \in [\![m]\!]\}$; only the X_i's are observed and the aim is to provide prediction intervals for the labels.

In addition, we consider the following transfer learning setting: while the data points are i.i.d. within each sample and the distributions of $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{test}}$ are the same, the distribution of $\mathcal{D}_{\text{train}}$ can be different. However, $\mathcal{D}_{\text{train}}$ can still help to build a good predictor by using a transfer learning toolbox, considered here as a black box (see, e.g., Zhuang et al., 2020 for a survey on transfer learning). A typical situation of use is when the training labeled data $\mathcal{D}_{\text{train}}$ is abundant but there is a domain shift for the test data, and we have a limited number of labeled data $\mathcal{D}_{\text{cal}}$ from the new domain.

3.2 Adaptive scores and procedures

Formally, the aim is to build $\boldsymbol{\mathcal{I}} = (\mathcal{I}_i)_{i \in [\![m]\!]}$, a family of m random intervals of $\mathbb{R}$ such that the amount of coverage errors $(\mathbf{1}\{Y_{n+i} \notin \mathcal{I}_i\})_{i \in [\![m]\!]}$ is controlled. The construction of a rule $\boldsymbol{\mathcal{I}}$ is based on non-conformity scores S_i, $1 \leq i \leq n+m$, corresponding to residuals between Y_i and the prediction at point X_i:

$$S_i := |Y_i - \hat{\mu}(X_i; (\mathcal{D}_{\text{train}}, \mathcal{D}^X_{\text{cal+test}}))|, \;\; i \in [\![n+m]\!], \tag{12}$$

where the predictor $\hat{\mu}$ is learnt using $\mathcal{D}_{\text{train}}$ and the calibration+test covariates $\mathcal{D}^X_{\text{cal+test}} = (X_1, \ldots, X_{n+m})$. More sophisticated scores than the residuals have been proposed in earlier literature (Romano et al., 2019; Gupta et al., 2022), in particular allowing for conditional variance or quantile prediction and resulting prediction intervals of varying length. Our theory extends to those as well and we consider here (12) for simplicity. We call the scores (12) *adaptive* because they can use the unlabeled data $\mathcal{D}^X_{\text{cal+test}}$, which is particularly suitable in the transfer learning framework where the covariates of $\mathcal{D}_{\text{train}}$ should be mapped to those of $\mathcal{D}^X_{\text{cal+test}}$ to build a good predictor. Classical scores can also be recovered via (12) if the predictor ignores $\mathcal{D}^X_{\text{cal+test}}$. The predictor $\widehat{\mu}$ can be any

"black box" (an unspecified transfer learning algorithm) provided the following mild assumption is satisfied, ensuring score exchangeability:

$$\forall x \in \mathbb{R}^d,\ \hat{\mu}\big(x; (\mathcal{D}_{\text{train}}, \mathcal{D}^X_{\text{cal+test}})\big) \text{ is invariant by permutation of } \mathcal{D}^X_{\text{cal+test}}. \quad \text{(PermInv)}$$

Since $(X_i, Y_i)_{i\in[\![n+m]\!]}$ are i.i.d. and thus exchangeable, one can easily show that (Exch) holds for the adaptive scores (12) when the predictor satisfies (PermInv). Predictors based on transfer machine learning procedures typically satisfy (PermInv). In addition, (NoTies) is a mild assumption: add a negligible noise to the scores is an appropriate tie breaking that makes (NoTies) hold.

Given the scores (12), we build the conformal p-values via (1) and define the specific conformal procedure $\boldsymbol{\mathcal{C}}(\alpha) = (\mathcal{C}_i(\alpha))_{i\in[\![m]\!]}$ obtained by inverting $\{p_i > \alpha\}$ with respect to Y_{n+i}, that is, $\{p_i > \alpha\} = \{Y_{n+i} \in \mathcal{C}_i(\alpha)\}$ almost surely with

$$\mathcal{C}_i(\alpha) := \big[\hat{\mu}(X_{n+i}; (\mathcal{D}_{\text{train}}, \mathcal{D}^X_{\text{cal+test}})) \pm S_{(\lceil (n+1)(1-\alpha)\rceil)}\big], \tag{13}$$

where $S_{(1)} \le \cdots \le S_{(n)} \le S_{(n+1)} := +\infty$ denote the order statistics of the calibration scores $(S_1, \ldots, S_n)$. Observe that the radius of the interval $S_{(\lceil (n+1)(1-\alpha)\rceil)}$ can be equivalently described as the $(1-\alpha)$-quantile of the distribution $\sum_{i=1}^n \frac{1}{n+1}\delta_{S_i} + \frac{1}{n+1}\delta_{+\infty}$. Note also that $\boldsymbol{\mathcal{C}}(\alpha) = \mathbb{R}^m$ if $\alpha < 1/(n+1)$, that is, if the desired coverage error is too small w.r.t. the size of the calibration sample.

3.3 Transductive error rates

By Proposition 2.2, the following marginal control holds for the conformal procedure $\boldsymbol{\mathcal{C}}(\alpha)$ (13):

$$\mathbb{P}(Y_{n+i} \notin \mathcal{C}_i(\alpha)) \le \alpha,\ \ i \in [\![m]\!]. \tag{14}$$

This is classical for non-adaptive scores and our result already brings an extension to adaptive scores in the transfer learning setting.

In addition, we take into account the prediction multiplicity by considering *false coverage proportion* (FCP) of some procedure $\boldsymbol{\mathcal{I}} = (\mathcal{I}_i)_{i\in[\![m]\!]}$, given by

$$\mathrm{FCP}(\boldsymbol{\mathcal{I}}) := m^{-1} \sum_{i=1}^m \mathbf{1}\{Y_{n+i} \notin \mathcal{I}_i\}. \tag{15}$$

It is clear from (14) that the procedure $\boldsymbol{\mathcal{C}}(\alpha)$ (13) controls the *false coverage rate*, that is, $\mathrm{FCR}(\boldsymbol{\mathcal{C}}(\alpha))) := \mathbb{E}[\mathrm{FCP}(\boldsymbol{\mathcal{C}}(\alpha))] \le \alpha$. However, the error $\mathrm{FCP}(\boldsymbol{\mathcal{C}}(\alpha))$ naturally fluctuates around its mean and the event $\{\mathrm{FCP}(\boldsymbol{\mathcal{C}}(\alpha)) \le \alpha\}$ is not guaranteed. Hence, we aim at the following control in probability of the FCP:

$$\mathbb{P}[\mathrm{FCP}(\boldsymbol{\mathcal{C}}(\alpha)) \le \overline{\alpha}] \ge 1 - \delta. \tag{16}$$

Several scenarios can be considered: α is fixed and we want to find a suitable bound $\overline{\alpha} = \overline{\mathrm{FCP}}_{\alpha,\delta}$ for the "traditional" conformal procedure $\boldsymbol{\mathcal{C}}(\alpha)$; or conversely, $\overline{\alpha}$ is fixed and we want to adjust the parameter $\alpha = t_{\overline{\alpha},\delta}$ of the procedure to ensure the probabilistic control at target level $\overline{\alpha}$. For $\overline{\alpha} = 0$, this reduces to $\mathbb{P}[\forall i \in [\![m]\!],\ Y_{n+i} \in \mathcal{I}_i] \ge 1 - \delta$, i.e., no false coverage with high probability. By applying a union bound, the procedure $\boldsymbol{\mathcal{C}}(\delta/m)$ satisfies the latter control, as already proposed by Vovk (2013). However, in this case the predicted intervals can be trivial, that is, $\boldsymbol{\mathcal{C}}(\delta/m) = \mathbb{R}^m$, if the test sample is too large, namely, $m > \delta(n+1)$. Moreover, in a more general scenario the practitioner may want to adjust the parameter $\alpha = \widehat{\alpha}$ on their own depending on the data, for example based on some personal tradeoff between the probabilistic control obtained and the length of the corresponding prediction intervals — this is the common practice of a "post-hoc" choice (made after looking at the data). This motivates us to aim at a uniform (in α) bound, that is, find a family of random variables $(\overline{\mathrm{FCP}}_{\alpha,\delta})_{\alpha\in(0,1)}$ such that

$$\mathbb{P}\big[\forall \alpha \in (0,1),\ \mathrm{FCP}(\boldsymbol{\mathcal{C}}(\alpha)) \le \overline{\mathrm{FCP}}_{\alpha,\delta}\big] \ge 1 - \delta\,. \tag{17}$$

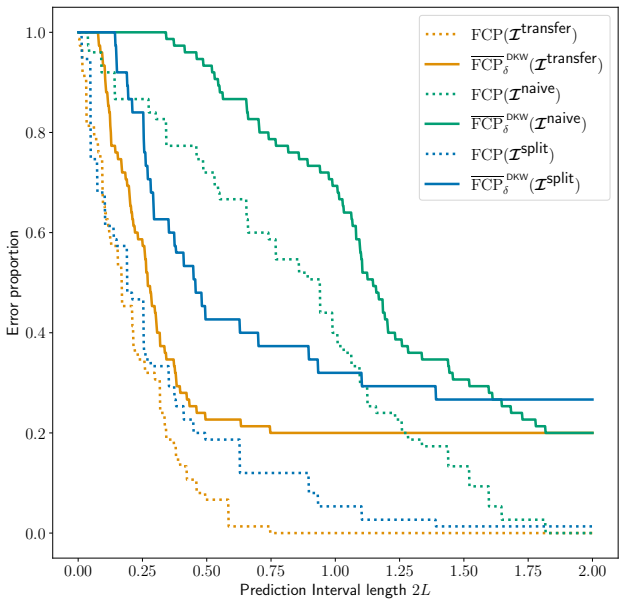

Figure 2: Plot of $\text{FCP}(\boldsymbol{\mathcal{I}})$ (15) (dashed) and bound $\overline{\text{FCP}}^{\text{DKW}}_{\widehat{\alpha}(L),\delta}$ (21) (18) (solid, $\delta = 0.2$) in function of interval length $2L$ in the same setting and procedures as in Figure 1.

Establishing such bounds is investigated in the next section. This gives a guarantee on the FCP in any of the above scenarios, in particular a post-hoc choice of the parameter $\widehat{\alpha}$. As a concrete example, one may want to choose a data-dependent $\widehat{\alpha}$ to ensure prediction intervals $\boldsymbol{\mathcal{C}}(\alpha)$ of radius at most L, namely,

$$\widehat{\alpha}(L) = (n+1)^{-1}\Big(1 + \sum_{i=1}^{n} \mathbf{1}\{S_i > L\}\Big). \tag{18}$$

Guarantee (17) yields a $(1-\delta)$-confidence error bound $\overline{\text{FCP}}_{\widehat{\alpha}(L),\delta}$ for this choice.

3.4 Controlling the error rates

To establish (16) and (17), we use that from (2), (13) and (15), $\text{FCP}(\boldsymbol{\mathcal{C}}(t)) = \widehat{F}_m(t)$ and thus for all $t \in [0,1]$,

$$\begin{aligned}\{\text{FCP}(\boldsymbol{\mathcal{C}}(t)) \le \overline{\alpha}\} &= \big\{\widehat{F}_m(t) \le \overline{\alpha}\big\} \\ &= \big\{m\widehat{F}_m(t) \le \lfloor \overline{\alpha} m \rfloor\big\} \\ &= \big\{p_{(\lfloor \overline{\alpha} m \rfloor + 1)} > t\big\},\end{aligned}$$

where $p_{(1)} \le \cdots \le p_{(m)}$ denote an ordered conformal p-values. We deduce the following result.

Corollary 3.1. *Let $n, m \ge 1$. Consider the setting of Section 3.1, the conformal procedure $\boldsymbol{\mathcal{C}}(\alpha)$ given by (13) and $P_{n,m}$ given by (5). Then the following holds:*

(i) for any $\overline{\alpha} \in [0,1]$, $\delta \in (0,1)$, $\boldsymbol{\mathcal{C}}(\alpha = t_{\overline{\alpha},\delta})$ satisfies (16) provided that $t_{\overline{\alpha},\delta}$ is chosen s.t.

$$\mathbb{P}_{\boldsymbol{p}\sim P_{n,m}}\big(p_{(\lfloor \overline{\alpha} m \rfloor + 1)} \le t_{\overline{\alpha},\delta}\big) \le \delta. \tag{19}$$

(ii) for any $\delta \in (0,1)$, $\big(\overline{\text{FCP}}_{\alpha,\delta}\big)_{\alpha\in(0,1)}$ satisfies (17) provided that

$$\mathbb{P}_{\boldsymbol{p}\sim P_{n,m}}\big(\exists \alpha \in (0,1) \,:\, \widehat{F}_m(\alpha) > \overline{\text{FCP}}_{\alpha,\delta}\big) \le \delta. \tag{20}$$

Applying Corollary 3.1 (i), for conformal prediction with guaranteed FCP, we obtain an adjusted level parameter which can be computed numerically (an explicit formula can also be given for $\alpha = 0$, see Supplemental D). Applying Corollary 3.1 (ii), and thanks to (9), the following family bound $(\overline{\mathrm{FCP}}_{\alpha,\delta})_{\alpha\in(0,1)}$ is valid for (17)

$$\overline{\mathrm{FCP}}^{\mathrm{DKW}}_{\alpha,\delta} = \big(\alpha + \lambda^{\mathrm{DKW}}_{\delta,n,m}\big)\mathbf{1}\{\alpha \geq 1/(n+1)\}, \tag{21}$$

with $\lambda^{\mathrm{DKW}}_{\delta,n,m} > 0$ given by (10). Obviously, numerical bounds can also be developed according to Remark 2.6.

Remark 3.2. Our FDP bounds extend to a selective inference framework where $\mathcal{S} = \mathcal{S}(\mathcal{D}_{\mathrm{train}}, \mathcal{D}^X_{\mathrm{cal+test}})) \subset [\![n+m]\!]$ is a selection rule invariant by permutation of $\mathcal{D}^X_{\mathrm{cal+test}}$, typically $\mathcal{S} = \{i \in [\![n+m]\!] : \hat{\mu}(X_i) \geq 0\}$. The calibration and test samples *over the selection* S are $\mathcal{D}_{\mathcal{S},\mathrm{cal}} = \{(X_i, Y_i), i \in \mathcal{S} \cap [\![n]\!]\}$ and $\mathcal{D}_{\mathcal{S},\mathrm{test}} = \{(X_{n+i}, Y_{n+i}), i \in \mathcal{S} \cap [\![n+1, n+m]\!]\}$, respectively. Defining the p-values accordingly, our envelopes are also valid for the FCP *over the selection* S by simply replacing n by $|\mathcal{S} \cap [\![n]\!]|$ and m by $|\mathcal{S} \cap [\![n+1, n+m]\!]|$. This complements the recent work of Bao et al. (2024), where only in-expectation results were established.

3.5 Numerical experiments

To illustrate the performance of the method, we consider the following proof-of-concept regression model: (W_i, Y_i) i.i.d. with $Y_i \mid W_i \sim \mathcal{N}(\mu(W_i), \sigma^2)$ for some unknown function μ and parameter $\sigma > 0$. To accommodate the transfer learning setting, we assume that we observe $X_i = f_1(W_i)$ in $\mathcal{D}_{\mathrm{train}}$ and $X_i = f_2(W_i)$ in $\mathcal{D}_{\mathrm{cal}} \cup \mathcal{D}_{\mathrm{test}}$ for some transformations f_1 and f_2. Three conformal procedures[2] $\boldsymbol{\mathcal{I}} = \boldsymbol{\mathcal{C}}(\alpha) = (\mathcal{C}_i(\alpha))_{i\in[\![m]\!]}$ are considered which differ only in the construction of the scores: first, $\boldsymbol{\mathcal{I}}^{\mathrm{naive}}$ consists in using a predictor of the usual form $\hat{\mu}(\cdot, \mathcal{D}_{\mathrm{train}})$ hence ignoring the distribution difference between $\mathcal{D}_{\mathrm{train}}$ and $\mathcal{D}_{\mathrm{cal}} \cup \mathcal{D}_{\mathrm{test}}$ (no transfer) with a RBF kernel ridge regression; the second procedure $\boldsymbol{\mathcal{I}}^{\mathrm{split}}$ ignores completely $\mathcal{D}_{\mathrm{train}}$ and works by splitting $\mathcal{D}_{\mathrm{cal}}$ in two new samples of equal size to apply the usual approach with these new (reduced) samples (transfer not needed); the third approach $\boldsymbol{\mathcal{I}}^{\mathrm{transfer}}$ is the proposed one, and uses the transfer predictor $\hat{\mu}(\cdot; (\mathcal{D}_{\mathrm{train}}, \mathcal{D}^X_{\mathrm{cal+test}}))$ based on optimal transport proposed by Courty et al. (2017). While all methods provide the correct $(1-\alpha)$ marginal coverage, we see from Figure 1 that $\boldsymbol{\mathcal{I}}^{\mathrm{transfer}}$ is much more accurate, which shows the benefit of using transfer learning and adaptive scores. Here, $|\mathcal{D}_{\mathrm{train}}| = 5000$, $n = m = 75$, $\mu(x) = \cos(x)$, $W_i \sim \mathcal{U}(0, 5)$, $f_1(x) = x$, $f_2(x) = 0.6x + x^2/25$ and $\sigma = 0.1$. Next, for each of the three methods, the FCP and corresponding bounds (21) are displayed in Figure 2. This illustrates both that each bound is uniformly valid in L and that transfer learning reduces the FCP (and thus also the FCP bounds).

4 Application to novelty detection

4.1 Setting

In the novelty detection problem, we observe the two following independent samples:

- a training null sample $\mathcal{D}_{\mathrm{null}}$ of n_0 nominal data points in $\mathbb{R}^d$ which are i.i.d. with common distribution P_0;
- a test sample $\mathcal{D}_{\mathrm{test}} = (X_i, i \in [\![m]\!])$ of independent points in $\mathbb{R}^d$ either distributed as P_0 or not.

The aim is to decide if each X_i is distributed as the training sample (that is, as P_0) or not. This long standing problem in machine learning has been recently revisited with the aim of controlling the proportion of errors among the items declared as novelties Bates et al. (2023);

[2]Python code for (PI) based on implementation of Boyer and Zaffran (2023).

let $\mathcal{H}_0 = \{i \in [\![m]\!] \,:\, X_i \sim P_0\}$ corresponding to the set of non-novelty in the test sample and consider the false discovery proportion

$$\mathrm{FDP}(R) = \frac{|R \cap \mathcal{H}_0|}{|R| \vee 1}, \tag{22}$$

for any (possibly random) subset $R \subset [\![m]\!]$ corresponding to the X_i's declared as novelties. The advantage of considering $\mathrm{FDP}(R)$ for measuring the errors has been widely recognized in the multiple testing literature since the fundamental work of Benjamini and Hochberg (1995) and its popularity is nowadays increasing in large scale machine learning theory, see Bates et al. (2023); Marandon et al. (2022); Jin and Candès (2023); Bashari et al. (2023), among others. The main advantage of $\mathrm{FDP}(R)$ is that the number of errors $|R \cap \mathcal{H}_0|$ is rescaled by the number of declared novelties $|R|$, which makes it scale invariant with respect to the size m of the test sample, so that novelty detection can still be possible in large scale setting.

4.2 Adaptive scores

Following Bates et al. (2023); Marandon et al. (2022), we assume that scores are computed as follows:

1. Split the null sample $\mathcal{D}_{\text{null}}$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}} = (X_i, i \in [\![n]\!])$ for some chosen $n \in (1, n_0)$;
2. Compute novelty scores $S_i = g(X_i)$, $i \in [\![n+m]\!]$, for some score function $g : \mathbb{R}^d \to \mathbb{R}$ (discussed below);
3. Compute conformal p-values as in (1).

In the work of Bates et al. (2023), the score function is built from $\mathcal{D}_{\text{train}}$ only, using a one-class classification method (classifier solely based on null examples), which makes the scores independent conditional to $\mathcal{D}_{\text{train}}$. The follow-up work Marandon et al. (2022) considers a score function depending both on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$ (in a permutation-invariant way of the sample $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$), which allows to use a two-class classification method including test examples. Doing so, the scores are adaptive to the form of the novelties present in the test sample, which significantly improves novelty detection (in a nutshell: it is much easier to detect an object when we have some examples of it). While the independence of the scores is lost, an appropriate exchangeability property is maintained so that we can apply our theory in that case, by assuming in addition (NoTies).

4.3 Methods and FDP bounds

Let us consider any thresholding novelty procedure

$$\mathcal{R}(t) := \{i \in [\![m]\!] \,:\, p_i \le t\}, \;\; t \in (0,1). \tag{23}$$

Then the following result holds true.

Corollary 4.1. *In the above novelty detection setting and under Assumption NoTies, the family of thresholding novelty procedures* (23) *is such that, with probability at least* $1-\delta$*, we have for all* $t \in (0,1)$*,*

$$\mathrm{FDP}(\mathcal{R}(t)) \le \frac{m I_n(t) + m\lambda^{DKW}_{\delta,n,m}}{1 \vee |\mathcal{R}(t)|} =: \overline{\mathrm{FDP}}^{DKW}_{t,\delta}, \tag{24}$$

and with an estimation of m_0*,*

$$\mathrm{FDP}(\mathcal{R}(t)) \le \frac{\hat{m}_0 I_n(t) + \max_{r \in [\![\hat{m}_0]\!]} \{r\lambda^{DKW}_{\delta,n,r}\}}{1 \vee |\mathcal{R}(t)|} =: \overline{\mathrm{FDP}}^{DKW}_{t,\delta}, \tag{25}$$

where $\lambda^{DKW}_{\delta,n,r}$ *is given by* (10) *and* $\hat{m}_0$ *is any random variable such that*

$$\hat{m}_0 \geq \max\left\{r : \inf_t \frac{\sum_{i=1}^m \mathbf{1}\{p_i > t\} + \max_{u\in[\![r]\!]}\{u\lambda^{DKW}_{\delta,n,u}\}}{1 - I_n(t)} \geq r\right\}, \tag{26}$$

where r *is in the range* $[\![m]\!]$ *and the maximum is equal to* m *if the set is empty.*

The proof is provided in Supplemental F.

Remark 4.2. Among thresholding procedures (23), AdaDetect (Marandon et al., 2022) is obtained by applying the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) to the conformal p-values. It is proved to control the expectation of the FDP (that is, the false discovery rate, FDR) at level α. Applying Corollary 4.1 provides in addition an FDP bound for AdaDetect, uniform in α, see Supplemental H.

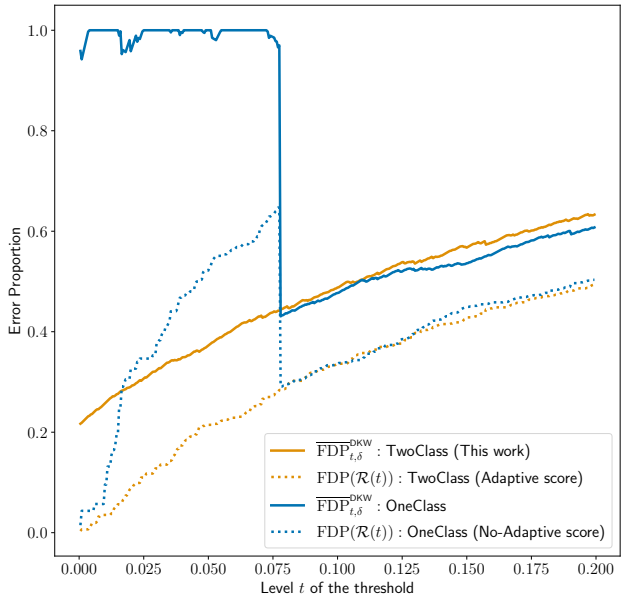

Figure 3: Plot of $\mathrm{FDP}(\mathcal{R}(t))$ (22)(23) (dashed) and bound $\overline{\mathrm{FDP}}^{\mathrm{DKW}}_{t,\delta}$ (25) (solid, $\delta = 0.2$) in function of the threshold t for $\mathcal{R}(t)$ (23) with a score obtained either with a one-class classification (non-adaptive) or a two-class classification (adaptive).

4.4 Numerical experiments

We follow the numerical experiments on "Shuttle" datasets of Marandon et al. (2022)[3]. In Figure 3, we displayed the true FDP and the corresponding bound (25) when computing p-values based on different scores: the non-adaptive scores of Bates et al. (2023) obtained with isolation forest one-class classifier; and the adaptive scores of Marandon et al. (2022) obtained with random forest two-class classifier. While the advantage of considering adaptive scores is clear (smaller FDP and bound) , it illustrates that the bound is correct simultaneously on t. Additional experiments are provided in Supplemental I.

5 Conclusion

The main takeaway from this work is the characterization of a "universal" joint distribution $P_{n,m}$ for conformal p-values based on n calibration points and m test points. We derived as

[3]The Python code uses the implementation of the procedure AdaDetect of Marandon (2022).

a consequence a non-asymptotic concentration inequality for the p-value empirical distribution function; numerical procedures can also be of use for calibration in practice. This entails uniform error bounds on the false coverage/false discovery proportion that hold with high probability, while standard results are only marginal or in expectation and not uniform in the decision. Since the results hold under the score exchangeability assumption only, they are applicable to *adaptive* score procedures using the calibration and test sets for training.

Acknowledgements

We would like to thank Anna Benhamou for constructive discussions and Romain Périer for pointing out an error in a previous version of the manuscript. The authors acknowledge the grants ANR-21-CE23-0035 (ASCAI), ANR-23-CE40-0018-01 (BACKUP) and ANR-23-CE40-0018-01 (BISCOTTE) of the French National Research Agency ANR, and the Emergence project MARS of Sorbonne Université.

References

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). Conformal prediction for reliable machine learning: theory, adaptations and applications. Morgan Kaufmann books.

Bao, Y., Huo, Y., Ren, H., and Zou, C. (2024). Selective conformal inference with false coverage-statement rate control. Biometrika, page asae010.

Bashari, M., Epstein, A., Romano, Y., and Sesia, M. (2023). Derandomized novelty detection with FDR control via conformal E-values. arXiv preprint 2302.07294.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. Ann. Statist., 51(1):149–178.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Statist., 29(4):1165–1188.

Bian, M. and Barber, R. F. (2022). Training-conditional coverage for distribution-free predictive inference. arXiv preprint arXiv:2205.03647.

Bian, M. and Barber, R. F. (2023). Training-conditional coverage for distribution-free predictive inference. Electronic Journal of Statistics, 17(2):2044–2066.

Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. Ann. Statist., 48(3):1281–1303.

Boyer, C. and Zaffran, M. (2023). Tutorial on conformal prediction. `https://claireboyer.github.io/tutorial-conformal-prediction/`.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. In Advances in neural information processing systems 30 (NIPS 2017), volume 30.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., and Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In 2015 IEEE symposium series on computational intelligence, pages 159–166. IEEE.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. Pattern Recognition, 127:108496.

Huang, K., Jin, Y., Candes, E., and Leskovec, J. (2023). Uncertainty quantification over graph with conformalized graph neural networks. Advances in Neural Information Processing Systems, 36.

Jin, Y. and Candès, E. J. (2023). Model-free selective inference under covariate shift via weighted conformal p-values. arXiv preprint arXiv:2307.09291.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. J. Amer. Stat. Assoc., 113(523):1094–1111.

Li, J., Maathuis, M. H., and Goeman, J. J. (2022). Simultaneous false discovery proportion bounds via knockoffs and closed testing. arXiv preprint arXiv:2212.12822.

Marandon, A. (2022). Machine learning meets FDR. `https://github.com/arianemarandon/adadetect#machine-learning-meets-fdr`.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2022). Machine learning meets false discovery rate. arXiv preprint 2208.06685.

Marques F., P. C. (2023). On the universal distribution of the coverage in split conformal prediction. arXiv preprint 2303.02770.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. Ann. Probab., 18(3):1269–1283.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In 13th European Conference on Machine Learning (ECML 2002), pages 345–356. Springer.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. J. Amer. Statist. Assoc., 100(469):94–108.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. Advances in neural information processing systems, 32.

Sarkar, S. and Kuchibhotla, A. K. (2023). Post-selection inference for conformal prediction: Trading off coverage for precision. arXiv preprint arXiv:2304.06158.

Saunders, C., Gammerman, A., and Vovk, V. (1999). Transduction with confidence and credibility. In 16th International Joint Conference on Artificial Intelligence (IJCAI 1999), pages 722–726.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. Biometrika, 73(3):751–754.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2013). Openml: networked science in machine learning. SIGKDD Explorations, 15(2):49–60.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In 4th Asian conference on machine learning (ACML 2012), pages 475–490. PMLR.

Vovk, V. (2013). Transductive conformal predictors. In Artificial Intelligence Applications and Innovations: 9th IFIP WG 12.5 International Conference (AIAI 2013), pages 348–360. Springer.

Vovk, V., Gammerman, A., and Shafer, G. (2005). Algorithmic learning in a random world. Springer.

Woods, K. S., Doss, C. C., Bowyer, K. W., Solka, J. L., Priebe, C. E., and Kegelmeyer Jr, W. P. (1993). Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. International Journal of Pattern Recognition and Artificial Intelligence, 7(06):1417–1436.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1):43–76.

A Exact formulas for $P_{n,m}$

In this section, we provide new formulas for the distribution $P_{n,m}$ given by (5). First let for $\boldsymbol{j} = (j_1, \dots, j_m) \in [\![n+1]\!]^m$, $\boldsymbol{M}(\boldsymbol{j}) := (M_k(\boldsymbol{j}))_{k \in [\![n+1]\!]}$ where $M_k(\boldsymbol{j}) := |\{i \in [\![m]\!] : j_i = k\}|$ is the number of coordinates of $\boldsymbol{j}$ equal to k, for $k \in [\![n+1]\!]$, and $\boldsymbol{M}(\boldsymbol{j})! := \prod_{k=1}^{n+1} (M_k(\boldsymbol{j})!)$.

Theorem A.1. *$P_{n,m}$ corresponds to the distribution of the colors of m successive draws in a standard Pólya urn model with $n+1$ colors labeled as $\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\}$ (with an urn starting with 1 ball of each color). That is, for $\boldsymbol{p} \sim P_{n,m}$ in (5), we have*

(i) Sequential distribution: for all $i \in [\![0, m-1]\!]$, the distribution of p_{i+1} conditionally on $p_1, \dots, p_i$ does not depend on m and is given by

$$\mathcal{D}(p_{i+1} \mid p_1, \dots, p_i) = \sum_{j=1}^{n+1} \frac{1 + \sum_{k=1}^{i} \mathbf{1}\{p_k = j/(n+1)\}}{n+1+i} \delta_{j/(n+1)}. \tag{27}$$

(ii) Joint distribution: for all vectors $\boldsymbol{j} \in [\![n+1]\!]^m$,

$$\mathbb{P}\left(\boldsymbol{p} = \frac{\boldsymbol{j}}{n+1}\right) = \boldsymbol{M}(\boldsymbol{j})! \frac{n!}{(n+m)!}, \tag{28}$$

(iii) Histogram distribution: the histogram of $\boldsymbol{p}$ is uniformly distributed on the set of histograms of m-sample into $n+1$ bins, that is, for all $\boldsymbol{m} = (m_1, \dots, m_{n+1}) \in [\![0, m]\!]^{n+1}$ with $m_1 + \cdots + m_{n+1} = m$,

$$\mathbb{P}\big(\boldsymbol{M}\big((n+1)\boldsymbol{p}\big) = \boldsymbol{m}\big) = \binom{n+m}{m}^{-1}. \tag{29}$$

In particular, conditionally on $\boldsymbol{M}\big((n+1)\boldsymbol{p}\big)$, the variable $\boldsymbol{p}$ is uniformly distributed on the set of possible trajectories, that is, for all vectors $\boldsymbol{j} \in [\![n+1]\!]^m$,

$$\mathbb{P}\left(\boldsymbol{p} = \frac{\boldsymbol{j}}{n+1} \;\middle|\; \boldsymbol{M}\big((n+1)\boldsymbol{p}\big) = \boldsymbol{M}\big(\boldsymbol{j}\big)\right) = \frac{\boldsymbol{M}\big(\boldsymbol{j}\big)!}{m!}. \tag{30}$$

Theorem A.1 is proved in Section C.3 for completeness. Theorem A.1 (i) gives the mechanism of the Pólya urn model: Namely, the urn first contains one ball of each of the $n+1$ colors, so p_1 has a uniform distributed on $\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\}$; then, given $p_1 = \ell/(n+1)$, we have drawn a ball of color ℓ and we put back this ball in the urn with another one of the same color ℓ, so p_2 is generated according to the distribution on $\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\}$ with equal chance $(= 1/(n+2))$ of generating $k/(n+1)$, $k \neq \ell$, and twice more chance $(= 2/(n+2))$ of generating $\ell/(n+1)$. Recursively, given $p_1, \dots, p_i$, the random variable p_{i+1} is generated in $\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\}$ according to the sizes of the histogram of the sample $((n+1)p_1, \dots, (n+1)p_i)$, see Figure 4.

Theorem A.1 (ii) provides the exact dependency structure between the p-values: for instance, $\boldsymbol{M}(\boldsymbol{j})! = 1$ when the coordinates of $\boldsymbol{j} = (j_1, \dots, j_m)$ are all distinct, while $\boldsymbol{M}(\boldsymbol{j})! = m!$ when the coordinates of $\boldsymbol{j} = (j_1, \dots, j_m)$ are the same. This means that the distribution slightly favors the $\boldsymbol{j}$ with repeated entries. This shows that the conformal p-values are not i.i.d. but have a positive structure of dependency. This is in accordance with the specific positive dependence property (called PRDS) already shown by Bates et al. (2023); Marandon et al. (2022).

Theorem A.1 (iii) shows an interesting non-concentration behavior of $P_{n,m}$ when n is kept small: if the p_i's were i.i.d. uniform on $\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\}$ then the histogram $\boldsymbol{M}((n+1)\boldsymbol{p})$ would follow a multinomial distribution and the histogram would concentrate around the uniform histogram as m tends to infinity. Rather, the p_i's are here only exchangeable, not i.i.d., and the histogram does not concentrate when m tends to infinity while n is small. As a case in point, for $n = 1$, $M_1((n+1)\boldsymbol{p})$ is uniform on $[\![m]\!]$, whatever m is, see (29). Nevertheless, we will show in the next section that a concentration occurs when *both m and n* tend to infinity.

Remark A.2. Note that P^U in (3) is the conditional distribution that one would get by applying the de Finetti theorem to the infinite exchangeable sequence $(p_i)_{i \geq 1}$ with $(p_1, \dots, p_m) \sim P_{n,m}$ for all m.

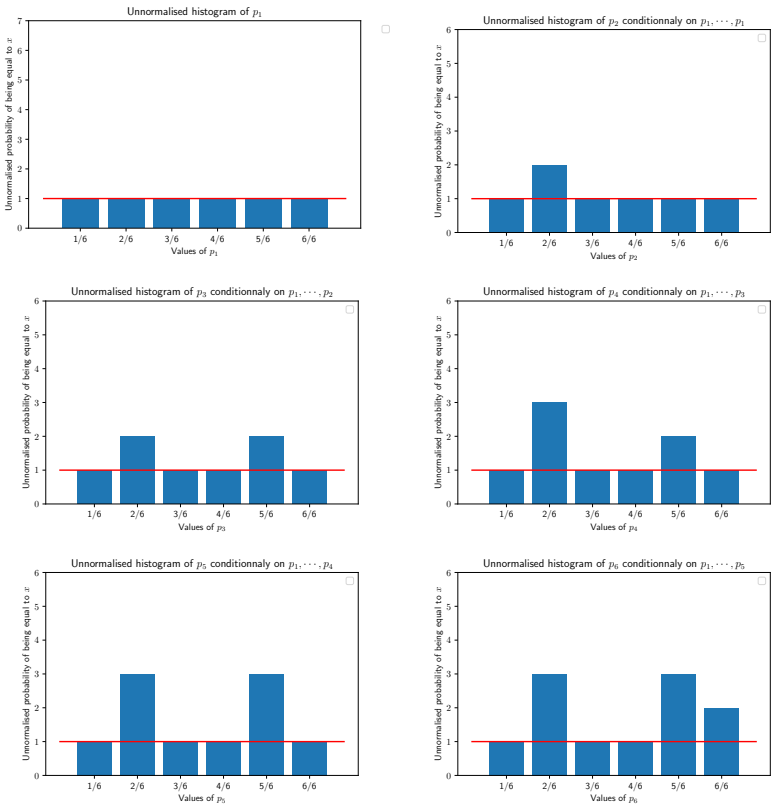

Figure 4: Illustration of the sequential realization of $P_{n,m}$ as proved in Theorem A.1 (ii) for $n = 5$ and $m = 6$.

Relation to Marques F. (2023); Huang et al. (2023). As a consequence of (27), given any $I \subset \left\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\right\}$, we have

$$\mathbb{P}(p_{i+1} \in I \,|\, p_1, \ldots, p_i) = \frac{|I| + N_i(I)}{n+1+i} = \mathbb{P}\Big(p_{i+1} \in I \,\Big|\, N_i(I)\Big),$$

where $N_i(I) = |\{k \in [\![i]\!] : p_k \in I)\}|$. In words, it means that the Pólya urn model continues to hold if we group (or "re-paint") the initial $(n+1)$ colors into only two colors, determined by whether the original color label belongs to I or not.

In particular, we recover the Pólya urn model put forward by Marques F. (2023): letting $Z_i = \mathbf{1}\{p_i > \alpha\}$, we have that for all $i \in [\![0, m-1]\!]$, the distribution of Z_{i+1} conditionally on $Z_1, \ldots, Z_i$ does not depend on m and is given by

$$\mathcal{D}(Z_{i+1} \,|\, Z_1, \ldots, Z_i) = \frac{\lfloor \alpha(n+1) \rfloor + \sum_{k=1}^{i} \mathbf{1}\{Z_k = j\}}{n+1+i} \delta_0 + \frac{\lceil (1-\alpha)(n+1) \rceil + \sum_{k=1}^{i} \mathbf{1}\{Z_k = j\}}{n+1+i} \delta_1. \tag{31}$$

Hence, the distribution of $(Z_1, \ldots, Z_m)$ corresponds to the distribution of the colors of m successive draws in a standard Pólya urn model with 2 colors labeled as $\{0, 1\}$ (with an urn starting with $\lfloor \alpha(n+1) \rfloor$ balls 0 and $\lceil (1-\alpha)(n+1) \rceil$ balls 1).

In particular, we recover Theorem 1 of Marques F. (2023) and Theorem 3 in Huang et al. (2023).

Corollary A.3 (Theorem 1 in Marques F. (2023) and Theorem 3 in Huang et al. (2023))**.** *In the*

setting of Theorem A.1, we have for all $\alpha \in (0,1)$ *and* $k \in [\![m]\!]$*, by denoting* $k_0 = \lceil \alpha(n+1) \rceil$*,*

$$\mathbb{P}\left(\widehat{F}_m(\alpha) = \frac{k}{m}\right) = \binom{m}{k} \frac{(n-k_0+1)\dots(n-k_0+m-k) \times k_0 \dots (k_0+k-1)}{(n+1)\dots(n+m)}. \tag{32}$$

Proof. By Proposition 2.2, (6) and the notation of (4), we have

$$\begin{aligned}
\mathbb{P}\left(\widehat{F}_m(\alpha) = \frac{k}{m}\right) &= \binom{m}{k} \mathbb{E}\big[(U_{(k_0)})^k (1-U_{(k_0)})^{m-k}\big] \\
&= \binom{m}{k} \frac{n!}{(k_0-1)!(n-k_0)!} \int_0^1 u^{k+k_0-1}(1-u)^{m-k+n-k_0} du \\
&= \binom{m}{k} \frac{n!}{(k_0-1)!(n-k_0)!} \frac{(k+k_0-1)!(m+n-k-k_0)!}{(m+n)!},
\end{aligned}$$

by using that $U_{(k_0)}$ follows a beta distribution with parameter $(k_0, n+1-k_0)$ and by using the beta distribution with parameter $(k+k_0, m+n+1-k-k_0)$. This shows the result. □

B Numerical bounds and templates

The bound proposed in Theorem 2.3 are explicit and elegant, but can be conservative in some cases and we develop here the numerical approach mentioned in Remark 2.6.

We rely on showing (11), which immediately implies a confidence envelope on $\widehat{F}_m$ because

$$\begin{aligned}
\left\{\forall k \in \mathcal{K} \,:\, \widehat{F}_m(t_k) \le \frac{k}{m}\right\} &= \left\{\forall k \in \mathcal{K} \,:\, \widehat{F}_m(t_k) < \frac{k+1}{m}\right\} \\
&= \{\forall k \in \mathcal{K} \,:\, p_{(k+1)} > t_k\}.
\end{aligned}$$

To establish (11), we use the notion of template introduced by Blanchard et al. (2020), see also Li et al. (2022). A template is a one-parameter family $t_k(\lambda)$, $\lambda \in [0,1]$, $k \in \mathcal{K} \subset [\![m]\!]$, such that $t_k(0) = 0$ and $t_k(\cdot)$ is continuous increasing on $[0,1]$. From above, we have for all λ,

$$\begin{aligned}
\{\forall k \in \mathcal{K} \,:\, \widehat{F}_m(t_k(\lambda)) \le k/m\} &= \{\forall k \in \mathcal{K} \,:\, p_{(k+1)} > t_k(\lambda)\} \\
&= \Big\{ \min_{k \in \mathcal{K}} \{t_k^{-1}(p_{(k+1)})\} > \lambda \Big\}.
\end{aligned}$$

Hence, let us consider

$$\lambda(\delta, n, m) = \max\left\{\lambda \in \Lambda \,:\, \mathbb{P}_{\boldsymbol{p} \sim P_{n,m}}\Big(\min_{k \in \mathcal{K}} \{t_k^{-1}(p_{(k)})\} > \lambda \Big) \ge 1 - \delta \right\}, \tag{33}$$

where Λ is the finite set $\{t_k^{-1}(\ell/(n+1)), k \in \mathcal{K}, \ell \in [\![n+1]\!]\}$. Then by Proposition 2.2 we have the following result.

Theorem B.1. *Let us consider the process* $\widehat{F}_m$ *defined by* (2)*, the distribution* $P_{n,m}$ *given by* (5)*, a template* $t_k(\lambda)$*,* $\lambda \in [0,1]$*,* $k \in \mathcal{K}$ *as above, and assume* (Exch) *and* (NoTies)*. Then we have for all* $\delta \in (0,1)$*,* $n, m \ge 1$*,*

$$\mathbb{P}\left(\forall k \in \mathcal{K} \,:\, \widehat{F}_m\Big(t_k\big(\lambda(\delta, n, m)\big)\Big) \le \frac{k}{m}\right) \ge 1 - \delta, \tag{34}$$

for $\lambda(\delta, n, m)$ *given by* (33)*.*

Here are two template choices:

- The linear template $t_k(\lambda) = k\lambda/m$, $\mathcal{K} = [\![m]\!]$, which leads to the inequality

$$\mathbb{P}\bigg(\exists t \in (0,1) \,:\, \widehat{F}_m(t) > \frac{\lceil tm/\lambda(\delta,n,m)\rceil}{m}\bigg) \leq \delta,$$

  which recovers the Simes inequality (44) with an adjusted scaling parameter.

- The "beta template" Blanchard et al. (2020), for which $t_k(\lambda)$ is the λ-quantile of the distribution $\mathrm{Beta}(k, m+1-k)$ and thus $\Lambda = \{F_{\mathrm{Beta}(k,m+1-k)}(\ell/(n+1)), k \in \mathcal{K}, \ell \in [\![n+1]\!]\}$. For instance, it can be used with $\mathcal{K} = \{1 + k\lceil \log(m)\rceil, k \in [\![K]\!]\}$.

C Proofs

C.1 Proof of Proposition 2.1

Assumption (NoTies) implies that marginal score distribution is atomless, so that F is continuous and $1 - F(S_i)$ has Unif$[0,1]$ distribution. Therefore, $(U_1, \ldots, U_{n+m}) = (1 - F(S_1), \ldots, 1 - F(S_{n+m}))$ are i.i.d. $\sim$ Unif$[0,1]$. Recall

$$p_i = (n+1)^{-1}\bigg(1 + \sum_{j=1}^{n} \mathbf{1}\{S_j \geq S_{n+i}\}\bigg), \;\; i \in [\![m]\!],$$

since p_i is a function of S_{n+i} and $\mathcal{D}_{\mathrm{cal}}$ only, it follows that conditionally on $\mathcal{D}_{\mathrm{cal}}$, the variables $p_1, \ldots, p_m$ are independent (and identically distributed).

Since F is continuous, it holds $F^\dagger(F(S_i)) = S_i$ almost surely, where $F^\dagger$ is the generalized inverse of F. Therefore $\mathbf{1}\{S_j \geq S_{n+i}\} = \mathbf{1}\{U_j \leq U_{n+i}\}$ almost surely. Hence, p_1 is distributed as

$$(n+1)^{-1}\bigg(1 + \sum_{j=1}^{n} \mathbf{1}\{U_j \leq U_{n+1}\}\bigg) = (n+1)^{-1}\bigg(1 + \sum_{j=1}^{n} \mathbf{1}\big\{U_{(j)} \leq U_{n+1}\big\}\bigg),$$

where $U_{(1)} \leq \cdots \leq U_{(n)}$ denotes the order statistics of $(U_1, \ldots, U_n)$. Therefore, we have for all $x \in [0,1]$,

$$\begin{aligned}
\mathbb{P}(p_1 \leq x \,|\, \mathcal{D}_{\mathrm{cal}}) &= \mathbb{P}\bigg(1 + \sum_{j=1}^{n} \mathbf{1}\big\{U_{(j)} \leq U_{n+1}\big\} \leq x(n+1)\Big|\mathcal{D}_{\mathrm{cal}}\bigg) \\
&= \mathbb{P}\bigg(1 + \sum_{j=1}^{n} \mathbf{1}\big\{U_{(j)} \leq U_{n+1}\big\} \leq \lfloor x(n+1)\rfloor\Big|\mathcal{D}_{\mathrm{cal}}\bigg) \\
&= \mathbb{P}(U_{n+1} < U_{(\lfloor x(n+1)\rfloor)}|\mathcal{D}_{\mathrm{cal}}) = U_{(\lfloor x(n+1)\rfloor)},
\end{aligned}$$

which finishes the proof.

C.2 Proof of Proposition 2.2

If there are no tied scores, which by assumption (NoTies) happens with probability 1, the ranks R_i of the ordered scores are well-defined and the vector $(p_1, \ldots, p_m)$ is only a function of the rank vector $(R_1, \ldots, R_{n+m})$. Namely, $R_i \leq R_j$ if and only if $S_i \leq S_j$, and the conformal p-values (1) can be written as

$$p_i = (n+1)^{-1}\bigg(1 + \sum_{j=1}^{n} \mathbf{1}\{R_j \geq R_{n+i}\}\bigg), \;\; i \in [\![m]\!].$$

Now, by (Exch), the vector $(R_1, \ldots, R_{n+m})$ is uniformly distributed on the permutations of $[\![n+m]\!]$. Any score distribution satisfying (NoTies) and (Exch) therefore gives rise to the same rank distribution, and thus the same joint p-value distribution. This joint distribution has been identified as (5)-(6) from the result of Proposition 2.1 in the particular case of i.i.d. scores. (Thus the i.i.d. assumption turns out to be unnecessary for what concerns the joint, unconditional distribution of the p-values, but provides a convenient representation.)

C.3 Proof of Theorem A.1

Proof of (ii) By (Exch),(NoTies) the permutation that orders the scores $(S_1, \dots, S_{n+m})$ that is σ such that

$$S_{()} = (S_{\sigma(1)} > \cdots > S_{\sigma(n+m)}),$$

is uniformly distributed in the set of permutations of $[\![n+m]\!]$. In addition, σ is independent of the order statistics $S_{()}$ and we seek for identifying the distribution of $(p_1, \dots, p_m)$ conditionally on $S_{()}$. Next, using again (Exch), we can assume without loss of generality that $j_1 \le \cdots \le j_m$ when computing the probability in (28). Now, due to the definition (1), the event $\{(p_1, \dots, p_m) = (j_1/(n+1), \dots, j_m/(n+1)\}$ corresponds to a specific event on σ. Namely, by denoting $(a_1, \dots, a_\ell)$ the vector of unique values of the set $\{j_1, \dots, j_m\}$ with $1 \le a_1 < \cdots < a_\ell \le n$, and $M_k = \sum_{i=1}^m \mathbf{1}\{j_i = a_k\}$, $1 \le k \le \ell$, the corresponding multiplicities, the above event corresponds to the situation

$$\underbrace{S_{\sigma(1)} > \cdots > S_{\sigma(a_1-1)}}_{a_1-1 \text{ null scores}} > \underbrace{S_{\sigma(a_1)} > \cdots > S_{\sigma(a_1+M_1-1)}}_{M_1 \text{ test scores in } \{S_{n+1}, \dots, S_{n+M_1}\}} >$$

$$\underbrace{S_{\sigma(a_1+M_1)} > \cdots > S_{\sigma(a_2+M_1-1)}}_{a_2 - a_1 \text{ null scores}} > \underbrace{S_{\sigma(a_2+M_1)} > \cdots > S_{\sigma_1(a_2+M_1+M_2-1)}}_{M_2 \text{ test scores in } \{S_{n+M_1+1}, \dots, S_{n+M_1+M_2}\}} > \cdots$$

$$\underbrace{S_{\sigma(a_{\ell-1}+M_1+\cdots+M_{\ell-1})} > \cdots > S_{\sigma(a_\ell+M_1+\cdots+M_{\ell-1}-1)}}_{a_\ell - a_{\ell-1} \text{ null scores}} > \underbrace{S_{\sigma(a_\ell+M_1+\cdots+M_{\ell-1})} > \cdots > S_{\sigma(a_\ell+m-1)}}_{M_\ell \text{ test scores in } \{S_{n+M_1+\cdots+M_{\ell-1}+1}, \dots, S_{n+m}\}} >$$

$$\underbrace{S_{(a_\ell+m)} > \cdots > S_{(n+m)}}_{n-a_\ell+1 \text{ null scores}}.$$

This event can be formally described as follows:

$$\Big\{\forall k \in [\![\ell]\!] \ : \sigma\big(\{a_\ell + M_1 + \cdots + M_{k-1}, \dots, a_\ell + M_1 + \cdots + M_k - 1\}\big) = \{n + M_1 + \cdots + M_{k-1} + 1, \dots, n + M_1 + \cdots + M_k\}\Big\}.$$

Since σ is uniformly distributed in the set of permutations of $[\![n+m]\!]$, the probability of this event (conditionally on $S_{()}$) is equal to $n!\big(\prod_{k=1}^\ell (M_k!)\big)/(n+m)!$, which yields (28).

Proof of (i) By using (28) of (ii), we have

$$\mathbb{P}(p_{i+1} = j_{i+1}/(n+1) \mid (p_1, \dots, p_i) = (j_1/(n+1), \dots, j_i/(n+1))) = \frac{\boldsymbol{M}(j_1, \dots, j_{i+1})! \frac{n!}{(n+i+1)!}}{\boldsymbol{M}(j_1, \dots, j_i)! \frac{n!}{(n+i)!}}.$$

Now, we have

$$\begin{aligned}
\boldsymbol{M}(j_1, \dots, j_{i+1})! &= \prod_{j=1}^{n+1} \left[\left(\sum_{k=1}^{i+1} \mathbf{1}\{j_k = j\} \right)! \right] = \prod_{j=1}^{n+1} \left[\left(\sum_{k=1}^{i} \mathbf{1}\{j_k = j\} + \mathbf{1}\{j_{i+1} = j\} \right)! \right] \\
&= \prod_{j=1}^{n+1} \left(\sum_{k=1}^{i} \mathbf{1}\{j_k = j\} \right)! \left[1 + \mathbf{1}\{j_{i+1} = j\} \sum_{k=1}^{i} \mathbf{1}\{j_k = j\} \right] \\
&= \boldsymbol{M}(j_1, \dots, j_i)! \left[1 + \mathbf{1}\{j_{i+1} = j\} \sum_{k=1}^{i} \mathbf{1}\{j_k = j\} \right].
\end{aligned}$$

This proves (27).

Proof of (iii) For all $\boldsymbol{m} = (m_1, \dots, m_{n+1}) \in [\![0, m]\!]^{n+1}$ with $m_1 + \cdots + m_{n+1} = m$, we have

$$\begin{aligned}
\mathbb{P}\big(\boldsymbol{M}((n+1)\boldsymbol{p}) = \boldsymbol{m}\big) &= \sum_{\boldsymbol{j} \in [\![n+1]\!]^m} \mathbf{1}\{\boldsymbol{M}(\boldsymbol{j}) = \boldsymbol{m}\} \mathbb{P}\big((n+1)\boldsymbol{p} = \boldsymbol{j}\big) \\
&= \boldsymbol{m}! \frac{n!}{(n+m)!} \sum_{\boldsymbol{j} \in [\![n+1]\!]^m} \mathbf{1}\{\boldsymbol{M}(\boldsymbol{j}) = \boldsymbol{m}\} \\
&= \boldsymbol{m}! \frac{n!}{(n+m)!} \frac{m!}{\boldsymbol{m}!} = \frac{n! m!}{(n+m)!},
\end{aligned}$$

where we have used (ii) and the multinomial coefficient.

C.4 Proof of Theorem 2.4

First observe that the LHS of (9) is 0 if $\lambda \geq 1$ so that we can assume $\lambda < 1$.

Let us prove (9) with the more complex bound

$$B^{\mathrm{DKWfull}}(\lambda, n, m) := \frac{n}{n+m} e^{-2m\lambda^2} + \frac{m}{n+m} e^{-2n\lambda^2} + C_{\lambda,n,m} \frac{2\sqrt{2\pi}\lambda nm}{(n+m)^{3/2}} e^{-\frac{2nm}{n+m}\lambda^2}, \tag{35}$$

where $C_{\lambda,n,m} = \mathbb{P}(\mathcal{N}(\lambda\mu, \sigma^2) \in [0, \lambda]) < 1$, for $\sigma^2 = (4(n+m))^{-1}$ and $\mu = n(n+m)^{-1}$. Let us comment the expression (35) of $B^{\mathrm{DKWfull}}(\lambda, n, m)$. As we can see, the role of n and m are symmetric (except in $C_{\lambda,n,m}$, that we can always further upper-bound by 1), and the two first terms are a convex combination of the usual DKW bounds for m and n i.i.d. variables, respectively. The third term is a "crossed" term between n and m, which becomes negligible if $n \gg m$ or $n \ll m$ but should be taken into account otherwise.

Below, we establish

$$\mathbb{P}\bigg(\sup_{t \in [0,1]} \big(\widehat{F}_m(t) - I_n(t)\big) > \lambda\bigg) \leq B^{\mathrm{DKWfull}}(\lambda, n, m); \tag{36}$$

$$\mathbb{P}\bigg(\sup_{t \in [0,1]} \big(-\widehat{F}_m(t) + I_n(t)\big) > \lambda\bigg) \leq B^{\mathrm{DKWfull}}(\lambda, n, m); \tag{37}$$

$$\mathbb{P}\Big(\big\|\widehat{F}_m - I_n\big\|_\infty > \lambda\Big) \leq 2B^{\mathrm{DKW}}(\lambda, n, m). \tag{38}$$

The result will be proved from (36) because $B^{\mathrm{DKWfull}}(\lambda, n, m) \leq B^{\mathrm{DKW}}(\lambda, n, m)$ since $n \vee m \geq nm/(n+m)$ and $C_{\lambda,n,m} \leq 1$.

The proof relies on Proposition 2.2 and the representation (6). Let $U = (U_1, \dots, U_n)$ i.i.d. $\sim U(0,1)$, and denote $F^U(x) = U_{(\lfloor (n+1)x \rfloor)}$, $x \in [0,1]$. Conditionally on U, draw $(q_i(U), i \in [\![m]\!])$ i.i.d. of common c.d.f. F^U and let

$$\widehat{G}_m(t) = m^{-1} \sum_{i=1}^m \mathbf{1}\{q_i(U) \leq t\}, \quad t \in [0,1],$$

the empirical c.d.f. of $(q_i(U), i \in [\![m]\!])$. By Proposition 2.2, we have that $\widehat{F}_m$ has the same distribution as $\widehat{G}_m$ (unconditionally on U), so that for any fixed $n, m \geq 1$ and $\lambda > 0$,

$$\mathbb{P}\bigg(\sup_{t \in [0,1]} \big(\widehat{F}_m(t) - I_n(t)\big) > \lambda\bigg) = \mathbb{E}\bigg[\mathbb{P}\bigg(\sup_{t \in [0,1]} \big(\widehat{G}_m(t) - I_n(t)\big) > \lambda \,\Big|\, U\bigg)\bigg]. \tag{39}$$

We now prove the bound (36) (the proof for (37) is analogous). Denote $Z = \sup_{t \in [0,1]} \big(F^U(t) -$

$I_n(t)\big) \in [0,1]$. We write by (39) and the triangle inequality

$$\begin{aligned}
\mathbb{P}\bigg(\sup_{t\in[0,1]}\big(\widehat{F}_m(t) - I_n(t)\big) > \lambda\bigg) &\le \mathbb{E}\bigg[\mathbb{P}\Big(\sup_{t\in[0,1]}(\widehat{F}_m(t) - F^U(t)) + Z > \lambda \,\Big|\, U\Big)\bigg] \\
&\le \mathbb{E}\bigg[\mathbb{P}\Big(\sup_{t\in[0,1]}(\widehat{F}_m(t) - F^U(t)) \ge \big(\lambda - Z\big)_+ \,\Big|\, U\Big)\bigg] \\
&\le \mathbb{E}\Big[e^{-2m(\lambda-Z)_+^2}\Big].
\end{aligned}$$

The last inequality above is the DKW inequality (Massart, 1990) applied to control the inner conditional probability, since conditionally to U, $\widehat{F}_m$ is the e.c.d.f. of $(q_i(U), \in [\![m]\!])$, which are i.i.d. $\sim F^U$; and Z conditional to U is a constant. Now the last bound can be rewritten as

$$\begin{aligned}
\int_0^1 \mathbb{P}\Big(e^{-2m(\lambda-Z)_+^2} > v\Big)dv &= e^{-2m\lambda^2} + \int_{e^{-2m\lambda^2}}^1 \mathbb{P}\Big((\lambda - Z)_+ < \sqrt{\log(1/v)/(2m)}\Big)dv \\
&= e^{-2m\lambda^2} + \int_{e^{-2m\lambda^2}}^1 \mathbb{P}\Big(\lambda - Z < \sqrt{\log(1/v)/(2m)}\Big)dv \\
&= e^{-2m\lambda^2} + \int_{e^{-2m\lambda^2}}^1 \mathbb{P}\Big(Z > \big(\lambda - \sqrt{\log(1/v)/(2m)}\big)\Big)dv. \qquad (40)
\end{aligned}$$

To upper bound the integrand above, denote $\widehat{H}_n$ the ecdf of $(U_1, \dots, U_n)$; it holds for any $x \in [0,1]$:

$$\begin{aligned}
\mathbb{P}(Z > x) &= \mathbb{P}\bigg(\sup_{t\in[0,1]}\big(U_{(\lfloor(n+1)t\rfloor)} - \lfloor(n+1)t\rfloor/(n+1)\big) > x\bigg) \\
&= \mathbb{P}\big(\exists k \in [\![n]\!] \,:\, U_{(k)} > x + k/(n+1)\big) \\
&= \mathbb{P}\bigg(\exists k \in [\![n]\!] \,:\, \sum_{i=1}^n \mathbf{1}\{U_i \le x + k/(n+1)\} \le k - 1\bigg) \\
&= \mathbb{P}\Big(\exists k \in [\![n]\!] \,:\, \widehat{H}_n\big(x + k/(n+1)\big) - [x + k/(n+1)] \le (k-1)/n - [x + k/(n+1)]\Big). \\
&\le P\Big(\exists k \in [\![n]\!] \,:\, \widehat{H}_n\big(x + k/(n+1)\big) - [x + k/(n+1)] \le -x\Big) \\
&\le e^{-2nx^2},
\end{aligned}$$

where we used $(k-1)/n \le k/(n+1)$ in the first inequality, and the left-tail DKW inequality for the last one. Plugging this into (40) yields

$$\int_0^1 \mathbb{P}(e^{-2m(\lambda-Z)_+^2} > v)dv \le e^{-2m\lambda^2} + \int_{e^{-2m\lambda^2}}^1 e^{-2n(\lambda - \sqrt{\log(1/v)/(2m)})^2}dv.$$

Now letting $u = \sqrt{\log(1/v)/(2m)}$ (hence $v = e^{-2mu^2}$, $dv = -4mue^{-2mu^2}du$), we obtain

$$\mathbb{P}\bigg(\sup_{t\in[0,1]}\big(\widehat{F}_m(t) - I_n(t)\big) > \lambda\bigg) \le e^{-2m\lambda^2} + 4m\int_0^\lambda ue^{-2n(\lambda-u)^2}e^{-2mu^2}du.$$

Now, by denoting $\sigma^2 = \big(4(n+m)\big)^{-1}$ and $\mu = n(n+m)^{-1}$, we get

$$\begin{aligned}
e^{\frac{2nm}{n+m}\lambda^2}\int_0^\lambda ue^{-2n(\lambda-u)^2}e^{-2mu^2}du &= \int_0^\lambda ue^{-2(n+m)\left(u - \frac{n\lambda}{n+m}\right)^2}du \\
&= \int_0^\lambda ue^{-\frac{1}{2\sigma^2}(u-\lambda\mu)^2}du \\
&= \int_0^\lambda (u-\lambda\mu)e^{-\frac{1}{2\sigma^2}(u-\lambda\mu)^2}du + \int_0^\lambda \lambda\mu e^{-\frac{1}{2\sigma^2}(u-\lambda\mu)^2}du \\
&= \sigma^2 e^{-2\lambda^2\frac{n^2}{m+n}} - \sigma^2 e^{-2\lambda^2\frac{m^2}{m+n}} + \lambda\mu\sqrt{2\pi}\sigma C_{\lambda,n,m}.
\end{aligned}$$

where $C_{\lambda,n,m} = \mathbb{P}\big(\mathcal{N}(\lambda\mu, \sigma^2) \in [0, \lambda]\big)$. Hence,

$$\begin{aligned}\int_0^\lambda u e^{-2n(\lambda-u)^2} e^{-2mu^2} du &= e^{-\frac{2nm}{n+m}\lambda^2}\Big(\sigma^2 e^{-2\lambda^2\frac{n^2}{m+n}} - \sigma^2 e^{-2\lambda^2\frac{m^2}{m+n}} + \lambda\mu\sqrt{2\pi}\sigma C_{\lambda,n,m}\Big) \\ &= \sigma^2 e^{-2n\lambda^2} - \sigma^2 e^{-2m\lambda^2} + \lambda\mu\sqrt{2\pi}\sigma C_{\lambda,n,m} e^{-\frac{2nm}{n+m}\lambda^2}.\end{aligned}$$

This leads to

$$\begin{aligned}&e^{-2m\lambda^2} + 4m\int_0^\lambda u e^{-2n(\lambda-u)^2} e^{-2mu^2} du \\ &= \frac{n}{n+m} e^{-2m\lambda^2} + \frac{m}{n+m} e^{-2n\lambda^2} + \lambda\sqrt{2\pi}\frac{nm}{(n+m)^{3/2}} 2C_{\lambda,n,m} e^{-\frac{2nm}{n+m}\lambda^2},\end{aligned}$$

which finishes the proof of (36).

Finally, let us prove $B^{\mathrm{DKW}}(\lambda^{\mathrm{DKW}}_{\delta,n,m}, n, m) \leq \delta$ for $\lambda^{\mathrm{DKW}}_{\delta,n,m} = \Psi^{(r)}(1)$ where $\Psi^{(r)}$ denotes the function Ψ iterated r times (for an arbitrary integer $r \geq 1$), where

$$\Psi(x) = 1 \wedge \widetilde{\Psi}(x); \quad \widetilde{\Psi}(x) := \left(\frac{\log(1/\delta) + \log\left(1 + \sqrt{2\pi}\frac{2\tau_{n,m}x}{(n+m)^{1/2}}\right)}{2\tau_{n,m}}\right)^{1/2}.$$

If $\Psi(1) = 1$, then $\Psi^{(r)}(1) = 1$ for all r and the announced claim holds since $B^{\mathrm{DKW}}(1, n, m) = 0$ by definition. We therefore assume $\Psi(1) < 1$ from now on. Since Ψ is non-decreasing, by an immediate recursion we have $\Psi^{(r+1)}(1) \leq \Psi^{(r)}(1) < 1$, for all integers r.

On the other hand, note that for any $x \in (0, 1)$ satisfying $\Psi(x) \leq x < 1$, it holds $\Psi(x) = \widetilde{\Psi}(x)$ and thus

$$B^{\mathrm{DKW}}(\Psi(x), n, m) = \left[1 + \frac{2\sqrt{2\pi}\Psi(x)\tau_{n,m}}{(n+m)^{1/2}}\right]\left[1 + \frac{2\sqrt{2\pi}x\tau_{n,m}}{(n+m)^{1/2}}\right]^{-1}\delta \leq \delta.$$

Since we established that $x = \Psi^{(r)}(1)$ satisfies $\Psi(x) \leq x$ for any integer r the claim follows.

D Explicit control of (16)

By applying (32) with $k = 0$, the control (16) for $\overline{\alpha} = 0$ is satisfied by choosing

$$t_{0,\delta} = \max\left\{k/(n+1) \ : \ \frac{(n-k+1)\dots(n-k+m)}{(n+1)\dots(n+m)} \geq 1 - \delta, k \in [\![n+1]\!]\right\}.$$

We can also obtain an implicit formula for $t_{\overline{\alpha},\delta}$ when $\overline{\alpha} > 0$ as follows. By definition, $t_{\overline{\alpha},\delta}$ is the maximum of the $t \in [\![n+1]\!]/(n+1)$ such that $\mathbb{P}_{\boldsymbol{p}\sim P_{n,m}}(p_{(\lfloor\overline{\alpha}m\rfloor+1)} \leq t) \leq \delta$, or equivalently $\mathbb{P}_{\boldsymbol{p}\sim P_{n,m}}(\mathrm{FCP}(\boldsymbol{\mathcal{C}}(t)) > \overline{\alpha}) \leq \delta$. The latter probability can be obtained explicitly from (32) with the formula

$$\mathbb{P}_{\boldsymbol{p}\sim P_{n,m}}(\mathrm{FCP}(\boldsymbol{\mathcal{C}}(t)) > \overline{\alpha}) = \sum_{k=\lfloor\overline{\alpha}m\rfloor+1}^{m}\binom{m}{k}\frac{(n-k_0+1)\dots(n-k_0+m-k)\times k_0\dots(k_0+k-1)}{(n+1)\dots(n+m)},$$

where $k_0 = \lceil t(n+1)\rceil$. Of course whenever this formula is too computationally complex for a practical use (e.g., when m is large), we can alternative use a Monte-Carlo scheme to simulate draws from $P_{n,m}$ and thus approximate $t_{\alpha,\delta}$ as an empirical δ-quantile of $p_{(\lfloor\overline{\alpha}m\rfloor+1)}$ with $\mathbf{p} \sim P_{n,m}$.

E On the tightness of the new DKW bound

The bound $B^{\mathrm{DKW}}(\lambda, n, m)$ (8) is simple and explicit and we comment here briefly about its sharpness:

- First, for a fixed m we have $B^{\mathrm{DKW}}(\lambda, n, m) \to \mathbf{1}_{\{\lambda<1\}} e^{-2m\lambda^2}$ when n tends to infinity. This bound hence recovers the usual DKW inequality Massart (1990) for $\widehat{F}_m$, which is well expected because $n = \infty$ corresponds to the case of i.i.d. uniform p-values ('theoretical' p-values rather than 'conformal' p-values). In addition, note that the usual DKW bound (in n) can be also recovered when n is fixed and m tends to infinity.
- This bound provides the correct variance term. Indeed, we can deduce from Theorem A.1 the following equality: for all $t, s \in \mathbb{R}$,

$$\mathrm{Cov}_{\mathbf{p}\sim P_{n,m}}\left(\widehat{F}_m(t), \widehat{F}_m(s)\right) = \frac{m+n+1}{m(n+2)}(I_n(t) \wedge I_n(s) - I_n(t)I_n(s)).$$

  Clearly, we have $\frac{m+n+1}{m(n+2)} \sim \tau_{n,m}$ when $m \wedge n \to +\infty$.
- The bound is compared to the true probability by using simulations in Figure 5. We observe that the bound is fairly close to the target when λ is large enough or/and $m \wedge n$ is large.

As mentioned in Remark 2.6, recall that this bound can be made sharper by using (non-explicit) numerical approximations.

F Proof of Corollary 4.1

Let $m_0 = |\mathcal{H}_0|$. We establish the following more general result.

Lemma F.1. *With probability at least $1-\delta$, we have both*

$$\forall t \in (0,1),\ \mathrm{FDP}(\mathcal{R}(t)) \leq \frac{m_0 I_n(t) + m_0 \lambda^{DKW}_{\delta,n,m_0}}{1 \vee |\mathcal{R}(t)|}; \tag{41}$$

$$m_0 \leq \max\left\{ r \in [\![m]\!] \,:\, \inf_t \left(\frac{\sum_{i=1}^m \mathbf{1}\{p_i > t\} + \max_{u\in[\![r]\!]}\left(u\lambda^{DKW}_{\delta,n,u}\right)}{1 - I_n(t)} \right) \geq r \right\}. \tag{42}$$

Lemma F.1 implies Corollary 4.1 because if $\hat{m}_0$ is as in (26), with probability at least $1-\delta$, $\hat{m}_0 \geq m_0$ by (42), and by (41)

$$\forall t \in (0,1),\ \mathrm{FDP}(\mathcal{R}(t)) \leq \frac{m_0 I_n(t) + m_0 \lambda^{\mathrm{DKW}}_{\delta,n,m_0}}{1 \vee |\mathcal{R}(t)|} \leq \frac{\hat{m}_0 I_n(t) + \max_{r\in[\![\hat{m}_0]\!]}\left(r\lambda^{\mathrm{DKW}}_{\delta,n,r}\right)}{1 \vee |\mathcal{R}(t)|}.$$

Now, let us prove Lemma F.1.

First, in the work of Marandon et al. (2022), it is proved that $(S_1, \ldots, S_n, S_{n+i}, i \in \mathcal{H}_0)$ is exchangeable conditionally on $(S_{n+i}, i \in \mathcal{H}_1)$ (see Lemma 3.2 therein). Hence, the vector $(S_1, \ldots, S_n, S_{n+i}, i \in \mathcal{H}_0)$, of size $n + m_0$, and the p-value vector $(p_i, i \in \mathcal{H}_0)$, of size m_0, fall into the setting described in Section 2.1 with calibration scores being $(S_i)_{i\in[\![n]\!]}$ and test scores being $(S_{n+i})_{i\in\mathcal{H}_0}$. By Proposition 2.2, this means $(p_i, i \in \mathcal{H}_0) \sim P_{n,m_0}$.

Second, consider the event

$$\Omega = \left\{ \sup_{t\in[0,1]} (\widehat{F}_{m_0}(t) - I_n(t)) \leq \lambda^{\mathrm{DKW}}_{\delta,n,m_0} \right\}. \tag{43}$$

By applying Theorem 2.4 and the explicit bound (10), we have $\mathbb{P}(\Omega) \geq 1-\delta$. Next, $|\mathcal{R}(t) \cap \mathcal{H}_0| = m_0 \widehat{F}_{m_0}(t) \leq m_0 I_n(t) + m_0 \lambda^{\mathrm{DKW}}_{\delta,n,m_0}$ on Ω. This gives (41).

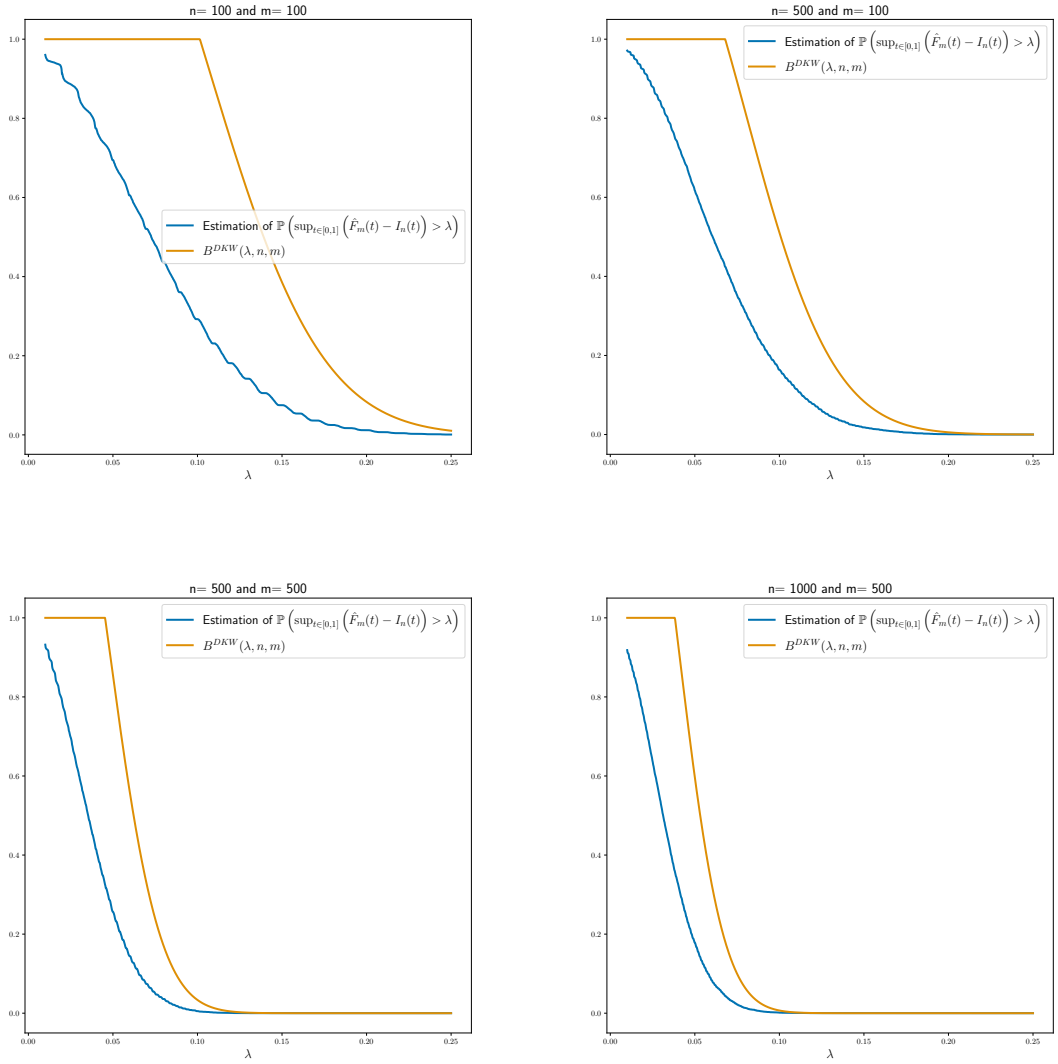

Figure 5: Plot of $\lambda \mapsto \mathbb{P}(\sup_{t\in[0,1]}(\widehat{F}_m(t)-I_n(t))>\lambda)$ (Blue) and of $\lambda \mapsto B^{\mathrm{DKW}}(\lambda,n,m)$ (Orange) for different values of n and m. These probabilities are estimated with 10^4 Monte-Carlo iterations.

Let us now turn to prove (42) on Ω. For this, let us observe that on this event, we have for all $t \in (0,1)$,

$$\begin{aligned}\sum_{i=1}^{m} \mathbf{1}\{p_i > t\} \geq \sum_{i\in\mathcal{H}_0} \mathbf{1}\{p_i > t\} &= m_0(1-\widehat{F}_{m_0}(t)) \\ &\geq m_0(1-I_n(t)) - \max_{r\in[\![m_0]\!]}\left(r\lambda^{\mathrm{DKW}}_{\delta,n,r}\right)\end{aligned}$$

Hence, m_0 is an integer $r \in [\![m]\!]$ such that $\inf_t\left(\frac{\sum_{i=1}^m \mathbf{1}\{p_i>t\}+\max_{u\in[\![r]\!]}\{u\lambda^{\mathrm{DKW}}_{\delta,n,u}\}}{1-I_n(t)}\right) \geq r$, which gives (42).

G Confidence envelope and bounds derived from the Simes inequality

As proved in Bates et al. (2023) in the i.i.d. case, and since the joint distribution of the conformal p-values is the same under exchangeability of the scores (Proposition 2.2), the conformal p-values are positively regressively dependent on each one of a subset (PRDS) under (Exch) and (NoTies), see Benjamini and Yekutieli (2001) for a formal definition of the latter.

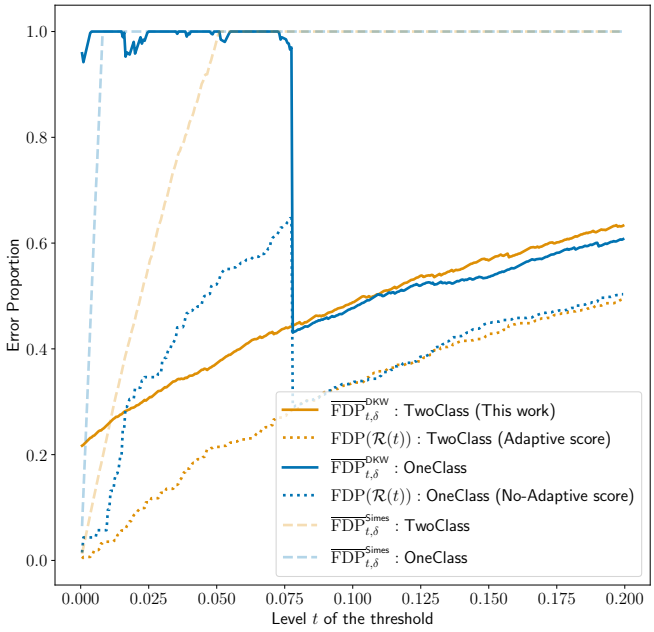

Figure 6: Same as Figure 3 with in addition Simes bound $\overline{\mathrm{FDP}}^{\mathrm{Simes}}_{t,\delta}$ (47) (transparent dashed, $\delta = 0.2$).

Hence, by Benjamini and Yekutieli (2001), the Simes inequality (Simes, 1986) is valid, that is, for all $\lambda > 0$, we have

$$\mathbb{P}\left(\sup_{t\in(0,1]} (\widehat{F}_m(t)/t) \geq \lambda \right) \leq 1/\lambda. \tag{44}$$

This envelope can be applied in the two applications of the paper as follows:

(PI) Under the condition of Corollary 3.1, the bound

$$\overline{\mathrm{FCP}}^{\mathrm{Simes}}_{\alpha,\delta} = (\alpha/\delta)\mathbf{1}\{\alpha \geq 1/(n+1)\} \tag{45}$$

is valid for (17).

(ND) Under the condition of Corollary 4.1 the following control is valid

$$\mathbb{P}\Big(\forall t \in (0,1), \mathrm{FDP}(\mathcal{R}(t)) \leq \overline{\mathrm{FDP}}^{\mathrm{Simes}}_{t,\delta}\Big) \geq 1-\delta, \tag{46}$$

for

$$\overline{\mathrm{FDP}}^{\mathrm{Simes}}_{t,\delta} := \frac{\hat{m}_0 t/\delta}{1 \vee |\mathcal{R}(t)|}, \tag{47}$$

for any estimator

$$\hat{m}_0 \geq m \wedge \inf_{t \in (0,\delta)} \frac{\sum_{i=1}^m \mathbf{1}\{p_i > t\}}{1 - t/\delta}. \tag{48}$$

A comparison between the Simes bound and the DKW bound is presented in Figure 6 for the (ND) task. While the Simes bound is better for extremely small t, the DKW bound is in general sharper.

H Uniform FDP bound for AdaDetect

AdaDetect (Marandon et al., 2022) is obtained by applying the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) to the conformal p-values, that is, $\mathrm{AD}_\alpha := \mathcal{R}(\alpha \hat{k}_\alpha / m)$, where

$$\hat{k}_\alpha := \max\left\{ k \in [\![0, m]\!] \,:\, \sum_{i=1}^m \mathbf{1}\{p_i \leq \alpha k/m\} \geq k \right\}. \tag{49}$$

It is proved there to control the false discovery rate (FDR), defined as the mean of the FDP:

$$\mathrm{FDR}(\mathrm{AD}_\alpha) := \mathbb{E}[\mathrm{FDP}(\mathrm{AD}_\alpha)] \leq \alpha m_0 / m. \tag{50}$$

Applying Corollary 25, we obtain on the top of the in-expectation guarantee (50) the following uniform FDP bound for AD_α: with probability at least $1 - \delta$, we have

$$\begin{aligned} &\forall \alpha \in (0,1),\ \ \mathrm{FDP}(\mathrm{AD}_\alpha) \leq \overline{\mathrm{FDP}}^{\mathrm{DKW}}_{\alpha,\delta} \\ &\overline{\mathrm{FDP}}^{\mathrm{DKW}}_{\alpha,\delta} := \left(\alpha \frac{\hat{m}_0}{m} + \frac{\hat{m}_0 \lambda^{\mathrm{DKW}}_{\delta,n,\hat{m}_0}}{\hat{k}_\alpha \vee 1} \right) \mathbf{1}\Big\{ \hat{k}_\alpha > 0 \Big\}, \end{aligned} \tag{51}$$

where $\hat{k}_\alpha$ is the rejection number (49) of AD_α and $\hat{m}_0$ satisfies (26).

In addition, we consider

$$\overline{\mathrm{FDP}}^{\mathrm{Simes}}_{\alpha,\delta} := \frac{\hat{m}_0 \alpha}{m\delta} \mathbf{1}\Big\{ \hat{k}_\alpha > 0 \Big\}, \tag{52}$$

for any estimator $\hat{m}_0$ given by (48).

I Additional experiments

In this section, we provide experiments to illustrate the FDP confidence bounds for AdaDetect, as mentioned in Remark 4.2 and Section H.

The two procedures used are of the AdaDetect type (49) but with two different score functions: the Random Forest classifier from Marandon et al. (2022) (adaptive score), and the one class classifier Isolation Forest as in Bates et al., 2023 (non adaptive score). The hyperparameters of these two machine learning algorithms are those given by Marandon (2022).

The FDP and the corresponding bounds are computed for the two procedures. The true discovery proportion is defined by

$$\mathrm{TDP}(R) = \frac{|R \cap \mathcal{H}_1|}{|\mathcal{H}_1| \vee 1}, \tag{53}$$

Table 1: Summary of datasets. "Shuttle" is originally from UCI depository. "Credit card" is from Dal Pozzolo et al. (2015). "Mammography" is from Woods et al. (1993).

	Shuttle	Credit card	Mammography
Dimension d	9	30	6
Feature type	Real	Real	Real
$\lvert\mathcal{D}_{\text{train}}\rvert$	3000	2000	2000
n calibration sample size	2000	1000	1000
m_0 (test) inlier number	1500	500	500
m_1 (test) novelty number	300	260	260
$m = m_0 + m_1$ total test sample size	1800	760	760

where $\mathcal{H}_1 = [\![m]\!] \setminus \mathcal{H}_0$; this criterion will be considered in addition to the FDP to evaluate the detection power of the procedures.

Following the numerical experiments of Marandon et al. (2022) and Bates et al. (2023), we consider the three different real data from OpenML dataset (CC-BY license)(Vanschoren et al., 2013) given in Table 1.

The results are displayed in Figure 7 for comparison of adaptive versus non-adaptive scores for the different FDP confidence bounds and the TDP. On Figure 8, we focus on the adaptive scores and corresponding FDP bounds only; we compare the effect (on the bounds) of demanding a more conservative error guarantee ($\delta = 0.05$ versus $\delta = 0.2$), as well as the effect of estimating m_0 via (26) instead of just using the inequality (25) with $\hat{m}_0 = m$.

The high-level conclusions are the following:

- using adaptive scores rather that non-adaptive ones results in a performance improvement (better true discovery proportion for the same target FDR level)
- for small target FDR level α, the Simes upper bounds $\overline{\text{FDP}}^{\text{Simes}}_{\alpha,\delta}$ are sharper than the DKW bound, elsewhere the new DKW bound is sharper than Simes. Furthermore, the relevant region for the Simes bound having the advantage becomes all the more tenuous as the error guarantee for the bound becomes more stringent (smaller δ). The reason is that the Simes upper bound is linear in δ^{-1}, while the DKW is only (square root) logarithmic.
- estimating the estimator $\hat{m}_0$ from (26) yields sharper bounds on the FDP and is therefore advantageous.

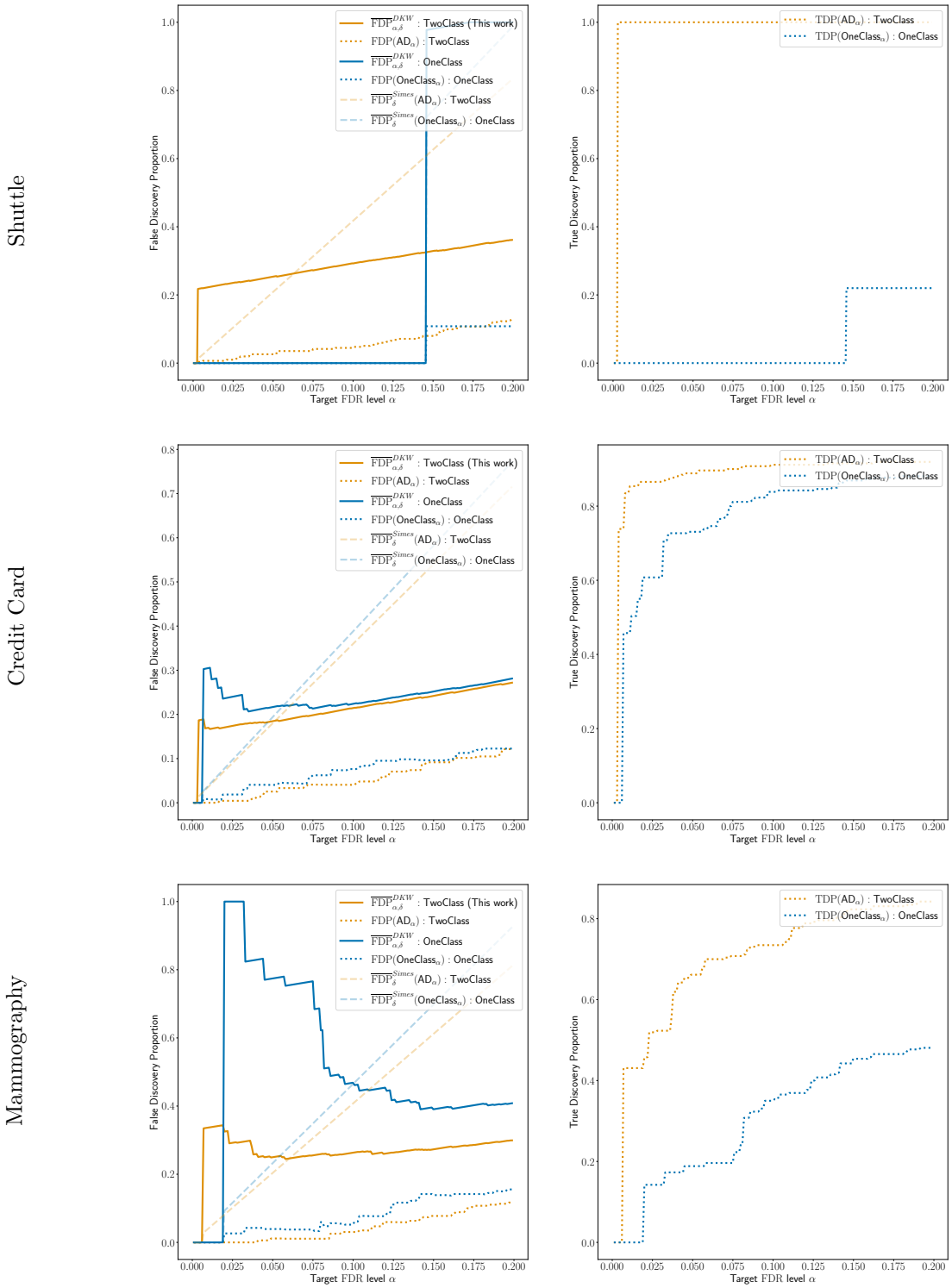

Figure 7: Left: $\text{FDP}(\text{AD}_\alpha)$ (22)(49) (dotted) and bounds $\overline{\text{FDP}}^{\text{DKW}}_{\alpha,\delta}$ (51) (solid) $\overline{\text{FDP}}^{\text{Simes}}_{\alpha,\delta}$ (52) (dashed) ($\delta = 0.2$) in function of the nominal FDR-level α. Right: corresponding TDP (53). In AdaDetect, the score is obtained either with a one-class classification (non-adaptive, blue) or a two-class classification (adaptive, orange); higher is better.

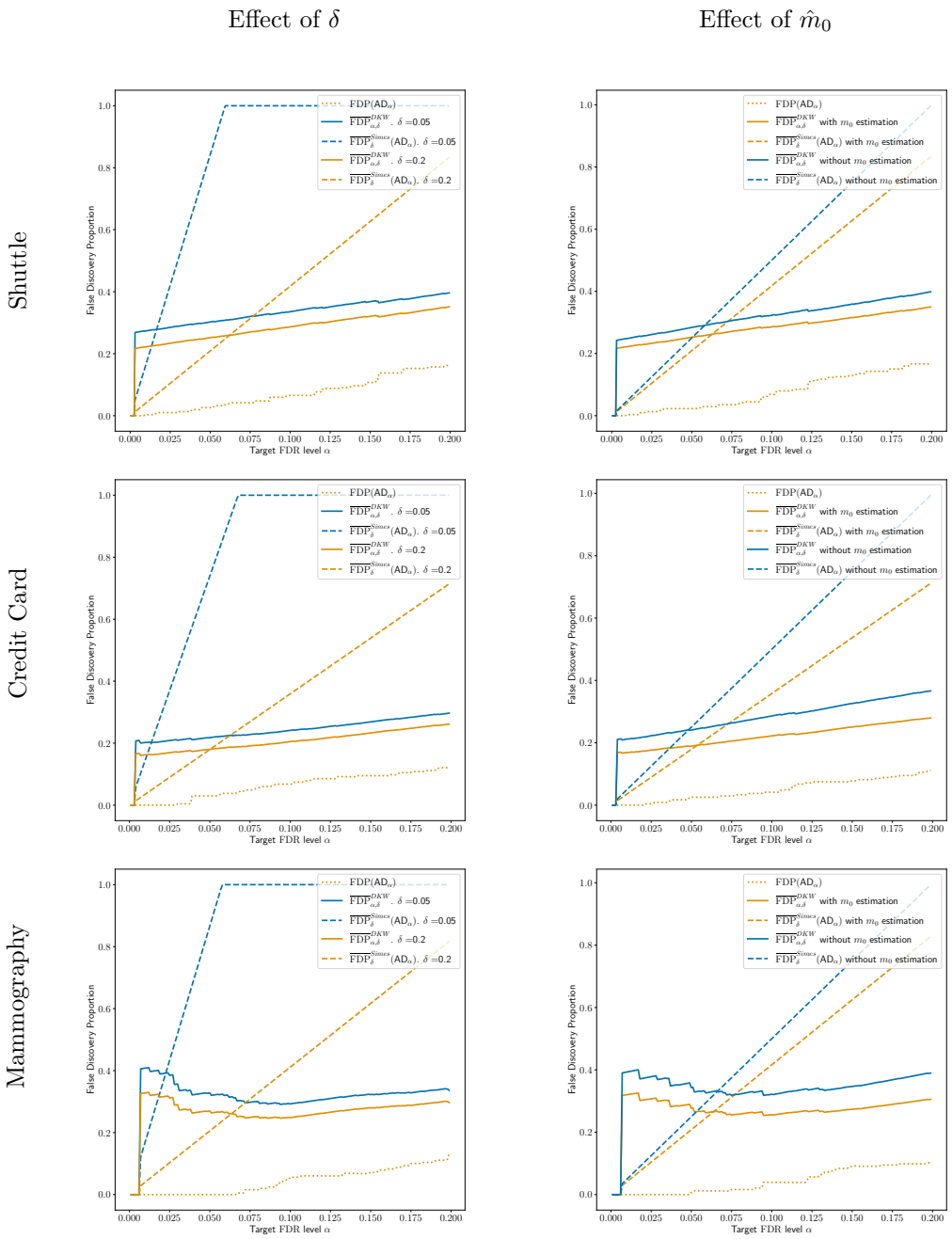

Figure 8: Same curves as Figure 7 (left), but only for two-class classification (adaptive, orange). Left: for comparison, the bounds $\overline{\text{FDP}}$ were also plotted for a smaller $\delta = 0.05$ value (blue). Right: for comparison, bounds $\overline{\text{FDP}}$ also plotted without an estimator of m_0 (taking m instead of $\hat{m}_0$).