Abstract

Covariate shift is a common and important assumption in transfer learning and domain adaptation to treat the distributional difference between the training and testing data. We propose a nonparametric test of covariate shift using the conformal prediction framework. The construction of our test statistic combines recent developments in conformal prediction with a novel choice of conformity score, resulting in a valid and powerful test statistic under very general settings. To our knowledge, this is the first successful attempt of using conformal prediction for testing statistical hypotheses. Our method is suitable for modern machine learning scenarios where the data has high dimensionality and large sample sizes, and can be effectively combined with existing classification algorithms to find good conformity score functions. The performance of the proposed method is demonstrated in synthetic and real data examples.

Keywords: Conformal prediction; Covariate shift; Distribution-free; Model misspecification; Hypothesis testing.
1. INTRODUCTION

In traditional machine learning, a common assumption is that the training data and testing data are independent samples from the same distribution. Building on such an independent, identically distributed (iid) assumption, many classical methods and results such as empirical risk minimization, regularization, and generalization have been developed. In recent years, as machine learning methods become more widely applied, it is often of interest to learn good prediction algorithms from related but different distributions. When the training data and testing data have different distributions, classical methods and theory developed for iid data often break down and need to be modified to account for such distribution difference. Sub-fields known as domain adaptation and transfer learning have emerged to deal with scenarios in which the training data and testing data may come from different but related distributions (Pan and Yang, 2009; Csurka, 2017; Kouw and Loog, 2018).

To avoid arbitrary changes in the distribution, one situation of particular practical and theoretical interest assumes that the conditional distribution of the response given the covariates remains the same between the training and testing data while the covariates may follow different marginal distributions. This assumption, called covariate shift, enables us to obtain improved predictive performance by weighting the training data according to the marginal density ratio. This approach has been widely studied in the machine learning literature. For examples, see Shimodaira (2000), Sugiyama et al. (2007), Sugiyama et al. (2008), Bickel et al. (2009), Gretton et al. (2009), and Sugiyama and Kawanabe (2012).

The purpose of this paper is to provide a statistical test for the hypothesis of covariate shift. Such a test is fundamentally important for the validity of the methods and theory developed under the covariate shift assumption. Given two independent samples of a pair of random variables \((X, Y)\) from two distributions, we are interested in testing whether the conditional distributions of \(Y\) given \(X\) are the same under these two distributions, which can be referred to as the two-sample test of equality of conditional distributions. If we do not reject the null hypothesis that the two conditional distributions are the same, then the covariate shift based methods and theory may be plausible. Otherwise, those methods shall be used with caution.

Despite its fundamental importance in transfer learning and domain adaptation, relatively less
is known about such two-sample conditional distribution tests. A related line of work studies the problem of testing the equality of conditional moments, including semiparametric methods (Hardle and Marron, 1990), nonparametric methods (Hall and Hart, 1990; Kulasekera, 1995; Kulasekera and Wang, 1997; Neumeyer and Dette, 2003), and second order moments (Pardo-Fernández et al., 2015). However, in many applications such as risk management and insurance, it is not enough to just consider mean and variance terms, and it is necessary to consider the whole conditional distribution of the response given the covariates. Another direction is to extend testing methods for unconditional distributions to conditional distributions. Andrews (1997) extended the Kolmogorov-Smirnov test to the conditional distribution case. Zheng (2000) proposed a test statistic based on the first-order linear expansion of Kullback-Leibler divergence. Fan et al. (2006) proposed a bootstrap test. Bai (2003) and Corradi and Swanson (2006) studied the problem of testing conditional distributions in a dynamic model. Most of the aforementioned methods rely on some assumptions that are hard to verify in a data-driven manner, such as smoothness of density and regression functions, and/or correctly specified parametric models. In addition, existing nonparametric methods usually involve nonparametric density estimation as an intermediate step, making them less feasible when the dimensionality is moderately high.

Our new method for testing covariate shift has three remarkable features. First, our test statistic is constructed using the conformal prediction technique (Vovk et al., 2005; Lei et al., 2013, 2018), a framework of converting point estimators to prediction sets by exploiting the symmetry of data. The type I error control is guaranteed by only assuming exchangeability of the training and testing data, and symmetry of the fitting procedure. To our knowledge, this is the first work applying conformal prediction to statistical hypothesis testing problems. Second, our method does not require estimating the density functions. Instead, it uses a classification algorithm to estimate the density ratio, and can incorporate almost any existing classification algorithms ranging from classical parametric estimators to modern black-box neural nets. In some cases, the validity of the p-value under the null hypothesis depends on the accuracy of the classifier, which can be empirically validated. This makes our method particularly useful in modern machine learning scenarios with high dimensionality and large sample sizes. Third, the asymptotic null distribution of our test statistic and its universal power guarantee are rigorously established under certain
moment conditions on the density ratios and the accuracy of classification algorithms. These theoretical results are supported by our simulation and real data examples.

**Related work in conformal prediction** We provide a general review of conformal prediction in Section 2.2. Roughly speaking, conformal prediction determines a sample point’s agreement, or conformity, with the current data set and fitting procedure using a conformity score measure, and includes points in the sample space with high conformity scores in the prediction set. Conformal prediction in regression has been studied by Lei and Wasserman (2014) in a nonparametric setting and Lei et al. (2018) in the high dimensional setting, where the conformity scores are chosen to be either the conditional density of $Y$ given $X$, or the absolute fitted residual. However, these conformity scores do not guarantee power against general alternatives when covariate shift is tested. A main methodological contribution of this paper is to show that using the conditional likelihood ratio as the conformity score can provide universal power guarantee against any alternatives. Such a choice of conformity score is partially inspired by a recent work of conformal prediction in classification by Guan and Tibshirani (2019). In the context of transfer learning and covariate shift, the data points are often exchangeable within the training set or the testing set alone, but not exchangeable when the training and testing datasets are merged together. This non-exchangeability issue can be treated using the weighted conformal prediction method developed in Tibshirani et al. (2019). Our theoretical study combines these methods, with further theoretical development to allow the weights and conformity scores to carry estimation errors. Moreover, these existing methods only consider prediction, and our method successfully transforms conformal prediction from a prediction tool to a hypothesis testing tool.

## 2. PROBLEM FORMULATION AND PRELIMINARIES

### 2.1 The two-sample conditional distribution testing problem

Consider two independent random samples

$$
\{(X_{1i}, Y_{1i})\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} P_1 \quad \text{and} \quad \{(X_{2i}, Y_{2i})\}_{i=1}^{n_2} \overset{i.i.d.}{\sim} P_2,
$$

where $P_1$, $P_2$ are distributions on the product space $X \times Y$. Here we consider a regression or classification setting where $X$ is the free variable (covariate), and $Y$ is the response variable. We
allow the spaces $\mathcal{X}$, $\mathcal{Y}$ to be general, and do not assume specific forms such as smoothness or parametric forms of $P_1$, $P_2$. For example, $\mathcal{X}$ and $\mathcal{Y}$ can be multi-dimensional Euclidean spaces, manifolds, or discrete sets. In the context of transfer learning and domain adaptation, $P_1$ represents the training sample, for which we have a large amount of labeled independent samples, and $P_2$ represents the testing sample, where we typically have unlabeled data. But to make testing covariate shift possible, we assume that we have a sample from $P_2$, whose size may be much smaller than the training sample.

For $j = 1, 2$, let $P_j(\cdot|\mathbf{x})$ be the conditional distribution of $Y$ given $X = \mathbf{x}$ under $P_j$, and $P_j, X(\cdot)$ be the corresponding marginal distribution of $X$. We are interested in testing whether these two conditional distributions are the same: $P_1(\cdot|\mathbf{x}) = P_2(\cdot|\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. If an $\mathbf{x} \in \mathcal{X}$ is not in the support of $P_{1, X}$ or $P_{2, X}$, then the testing problem becomes trivial for this $\mathbf{x}$ since we can define a conditional distribution at this $\mathbf{x}$ arbitrarily. Therefore, to facilitate discussion we assume that $P_{1, X}$ and $P_{2, X}$ are equivalent to each other in the following sense,

$$P_{1, X} \ll P_{2, X} \quad \text{and} \quad P_{2, X} \ll P_{1, X},$$

where “$\ll$” stands for absolute continuity. We further assume $P_{1, X}$, $P_{2, X}$ have density functions $f_{1, X}$, $f_{2, X}$ respectively, under a common base measure.

We consider the following two-sample conditional distribution testing problem.

$$H_0 : P_{1, X} \{P_1(\cdot|\mathbf{x}) = P_2(\cdot|\mathbf{x})\} = 1, \quad \text{versus} \quad H_1 : P_{1, X} \{P_1(\cdot|\mathbf{x}) = P_2(\cdot|\mathbf{x})\} < 1.$$  \hspace{1cm} (1)

Due to the assumed equivalence between $P_{1, X}$ and $P_{2, X}$, the hypotheses in (1) can be equivalently stated by replacing $P_{1, X}$ with $P_{2, X}$. For a similar consideration of avoiding triviality and the ease of discussion, we also assume that $P_1(\cdot|\mathbf{x})$ and $P_2(\cdot|\mathbf{x})$ are equivalent, with density functions $f_1(y|x)$ and $f_2(y|x)$ under a common base measure.

2.2 Background on conformal prediction

Here we briefly introduce conformal prediction in a regression setting. For conformal prediction in other contexts, such as unsupervised learning, see (Vovk et al., 2005; Lei et al., 2013, 2015).

Given iid data $\{(X_i, Y_i)\}_{i=1}^m$, conformal prediction converts a point estimator of the regression function $\hat{f} : \mathcal{X} \mapsto \mathcal{Y}$ to a prediction set $\hat{C} \in \mathcal{X} \times \mathcal{Y}$, with guaranteed finite-sample expected
prediction coverage:

\[ P \left\{ Y_{m+1} \in \hat{C}(X_{m+1}) \right\} \geq 1 - \alpha , \]  

(2)

where \( \hat{C}(x) = \{ y \in \mathcal{Y} : (x, y) \in \hat{C} \} \), and the probability is taken over the \((m+1)\)-tuple of iid data \( \{(X_i, Y_i) : 1 \leq i \leq m+1\} \).

Let \( \mathcal{D}_m \) denote the sample \( \{(X_i, Y_i) : 1 \leq i \leq m\} \), and \( \mathcal{D}_{m,-i} \) the sample obtained by removing \((X_i, Y_i)\) from \( \mathcal{D}_m \). A conformal prediction set \( \hat{C} \) is constructed using a conformity score function \( V : (\mathcal{X} \times \mathcal{Y})^{m+1} \rightarrow \mathbb{R} \) that is symmetric in its first \( m \) inputs. For a given data set \( \mathcal{D}_m \), a new \( X_{m+1} \) for which a prediction of the corresponding \( Y_{m+1} \) is wanted, and a \( y \in \mathcal{Y} \), let \( \mathcal{D}_{m+1}(y) \) be the augmented data set with the \((m+1)\)th data point being \((X_{m+1}, y)\). Let

\[ V_i(y) = V(\mathcal{D}_{m+1,-i}(y), (X_i, Y_i)), \quad i = 1, \ldots, m+1, \]  

(3)

be the conformity score for the \( i \)th sample point in the augmented data with \( Y_{m+1} = y \). The conformal prediction set using conformity score function \( V \) is

\[ \hat{C}(X_{m+1}) = \left\{ y \in \mathcal{Y} : \sum_{i=1}^{m+1} 1[V_i(y) \leq V_{n+1}(y)] \geq \lfloor (m+1)\alpha \rfloor \right\} . \]  

(4)

The finite sample coverage (2) can be easily derived from the iid assumption, the symmetry of \( V \), and the construction of \( \hat{C} \) in (4). To see this, if we replace \( y \) by \( Y_{m+1} \), then the iid assumption and symmetry of \( V \) implies exchangeability of \((V_i(Y_{m+1}) : 1 \leq i \leq m+1)\). Thus the rank of \( V_{m+1} \) being lower than \( \lfloor (m+1)\alpha \rfloor \) has probability no more than \( \alpha \).

Although the finite sample coverage guarantee only requires \( V \) to satisfy a symmetry condition, its choice will have a crucial impact on the quality of the resulting prediction set. A good choice of \( V \) needs to reflect the structure of the underlying distribution of \((X,Y)\) and be able to tell whether a sample point is likely drawn from this distribution. Such a function \( V \) is often constructed from a point estimate \( \hat{f} \) of the regression function. For example, in nonparametric regression, one can choose \( V(\mathcal{D}, (x, y)) = \hat{p}(y|x) \), where \( \hat{p}(\cdot|x) \) is an estimated conditional density function of \( y \) given \( x \) using the sample \( \mathcal{D} \cup \{(x, y)\} \) (Lei and Wasserman, 2014). In high dimensional regression, one can use \( V(\mathcal{D}, (x, y)) = |y_{m+1} - \hat{f}(x_{m+1})| \) where \( \hat{f} \) is an estimated regression function using \( \mathcal{D} \cup \{(x, y)\} \).

In general, the conformity score function \( V \) consists of a fitting part, which aims at recovering the relationship between \( Y \) and \( X \), and a conformity assessment part, which evaluates how well
a particular data point conforms to the estimated model. If both parts accurately reflect the underlying joint distribution, then the resulting prediction set will not only have coverage guarantee, but will also be nearly optimal.

The definition of \( \hat{C} \) in (4) is only conceptual and not practical if \( \mathcal{Y} \) is infinite, as it requires to evaluate \( V_i(y) \) for all \( y \) and all \( 1 \leq i \leq m+1 \). For practical implementation of conformal prediction, we refer to Lei et al. (2018) and Barber et al. (2019). However, in our hypothesis testing problem, we do not need to actually construct a prediction set. Instead, we only need to compute a single \( p \)-value for each observed data, and evaluate its deviance from the null distribution. The details are given in the next section as we develop our covariate shift testing method.

3. A CONFORMAL TEST OF COVARIATE SHIFT

3.1 The conformal \( p \)-value

The starting point of developing our conformal test for covariate shift is a \( p \)-value thresholding perspective to the definition of the conformal prediction set \( \hat{C} \) in (4).

Given data set \( D_m \), and \( X_{m+1} \), the construction of conformal prediction set \( \hat{C} \) at \( X_m+1 \) in (4) can be viewed as inverting a hypothesis test \((D_m, (X_{m+1}, Y_{m+1})) \overset{iid}{\sim} P \) with observed value \( Y_{m+1} = y \), using the test statistic

\[
U(y) = \frac{1}{m+1} \sum_{i=1}^{m} 1(V_i(y) \leq V_{m+1}(y))
\]

which, under the null hypothesis that \((D_m, (X_{m+1}, Y_{m+1})) \) is a realized iid sample from \( P \), behaves like an \( U(0,1) \) random variable. This provides a way to generate a valid \( p \)-value. In the following we show how to modify it so that \( U(Y_{m+1}) \) have the exact \( U(0,1) \) distribution.

We will first treat the simpler case where \( P_{1,X} = P_{2,X} \). Suppose we have data \( D_{(1)} = \{(X_{1i}, Y_{1i}) : 1 \leq i \leq m\} \) iid from \( P_1 \), and just a single pair \((X_2, Y_2) \sim P_2 \). We will use simplified notation \((X_i, Y_i) = (X_{1i}, Y_{1i}) \) for \( 1 \leq i \leq m \) and \((X_{m+1}, Y_{m+1}) = (X_2, Y_2) \).

If the covariate shift hypothesis \( H_0 \) in (1) is true, then \((X_1, Y_1), ... (X_m, Y_m), (X_{m+1}, Y_{m+1}) \) are iid from \( P_1 \) because for now the \( X \)-marginals are assumed to be equal. Then following the conformal prediction framework, we use a conformity score \( V \) as in (3) and consider the rank of \( V_{m+1}(Y_{m+1}) \) among all \( \{V_i(Y_{m+1}) : 1 \leq i \leq m+1\} \), with random tiebreak: Let \( R_- = 1 + \sum_{i=1}^{m+1} 1(V_i(Y_{m+1}) < V_{m+1}(Y_{m+1})) \) and \( R_+ = \sum_{i=1}^{m+1} 1(V_i(Y_{m+1}) \leq V_{m+1}(Y_{m+1})) \). Let \( R \) be uniformly and independently
sampled from the integers in \([R_-, R_+].\) Now we can construct a uniform random variable

\[ U = \frac{R - 1 + \zeta}{m + 1} \]  

(6)

where \( \zeta \sim U(0, 1) \) is independent of everything else. Because \( R \) has a uniform distribution on \( \{1, \ldots, m + 1\} \) by exchangeability, it follows that \( U \) has a uniform distribution on \([0, 1].\) The construction of \( U \) can be viewed as a continuous version of \( U(Y_{m+1}) \) given in (5).

The derivation above indicates that if \( P_{1,X} = P_{2,X} \) the statistic \( U \) in (6) can be used as an exact \( p \)-value to test the covariate shift hypothesis. Therefore, in the rest of this paper we call such a statistic \( U \) a conformal \( p \)-value.

In order to develop a test procedure, we need to resolve the following three issues.

1. How to choose a conformity score function \( V \) so that the test can separate \( H_0 \) and \( H_1? \)

2. How to make use of multiple data points from \( P_2 \) to have increased power under \( H_1? \)

3. How to allow for \( P_{1,X} \neq P_{2,X}? \)

These three issues are addressed in the next three subsections, respectively.

3.2 A choice of \( V \) that separates \( H_0 \) and \( H_1 \)

For now we still make the assumption \( P_{1,X} = P_{2,X}. \) A good choice of \( V \) will be such that the conformal \( p \)-value \( U \) constructed in (6) has a non-uniform distribution under the alternative hypothesis \( H_1. \) Common existing choices such as conditional density and absolute residual do not satisfy this. Our choice of \( V \) is the conditional Radon-Nikodym derivative between \( P_1 \) and \( P_2: \)

\[ V(x, y) = \frac{dP_1(y|x)}{dP_2(y|x)} = \frac{f_1(y|x)}{f_2(y|x)}. \]

This \( V \) function is different from the conformity score functions introduced in (3), as it only involves a single data pair \((x, y).\) This is not a real problem as we can let \( V \) be independent of the first \( m \) arguments and treat the input \((x, y)\) as the last argument.

The ability of \( V \) separating \( H_0 \) and \( H_1 \) is established by the following lemma.

**Lemma 1** (Separation of \( H_0 \) and \( H_1 \) by \( V \) under equal \( X \)-marginal). If \( P_{1,X} = P_{2,X}, \) then there exist \( \delta > 0 \) and \( m_0, \) depending only on \((P_1, P_2),\) such that the statistic \( U \) constructed in (6) with \( V(D_{m+1,-i}, (x, y)) = \frac{f_1(y|x)}{f_2(y|x)} \) satisfies \( \mathbb{E}(U) \leq 1/2 - \delta \) under \( H_1 \) when \( m \geq m_0. \)
Lemma 1 is a special case of Lemma 2(c), for which a complete proof is provided in Appendix C.

**Remark 1.** The choice of $V$ is motivated from an information theoretic perspective. Define $\gamma(x) = \mathbb{E}_{Y \sim f_1(\cdot|x)} \frac{f_1(Y|x)}{f_2(Y|x)} = \mathbb{E}_{Y \sim f_1(\cdot|x)} V(Y, x)$. Then by Jensen’s inequality and non-negativity of the Kullback-Leibler divergence, we have $\log \gamma(x) \geq \mathbb{E}_{Y \sim f_1(\cdot|x)} \log \frac{f_1(Y|x)}{f_2(Y|x)} \geq 0$. Thus, $\gamma(x) \geq 1$ with equality holds if and only if $f_1(\cdot|x) = f_2(\cdot|x)$. Now $\mathbb{E} V \sim P_1 V - \mathbb{E} V \sim P_2 V = \mathbb{E} X \sim P_1 (\gamma(X) - 1) \geq 0$ with equality holds only if $\gamma(x) = 1$ a.e. $P_1$. A more involved argument is needed in order to carry over this intuition rigorously to analyze the rank of $V_{m+1}$ and the continuous version in (6).

**Remark 2.** The function $V$ involves the unknown density ratio $f_1(y|x)/f_2(y|x)$. Our method will need to use an empirical version of $V$:

$$\hat{V}(x, y) = \frac{\hat{f}_1(\cdot|\cdot)}{\hat{f}_2(\cdot|\cdot)}(x, y),$$

where $\hat{f}_1(\cdot|\cdot)/\hat{f}_2(\cdot|\cdot)$ is an estimate of the conditional density ratio, independent of $\{(X_{1i}, Y_{1i}) : 1 \leq i \leq m\}$ and $(X_2, Y_2)$. A remarkable advantage of our choice of $V$ and $\hat{V}$ is that the density ratio $f_1(\cdot)/f_2(\cdot)$ can be conveniently estimated using classification algorithms, which is both theoretically and practically much easier than estimating the density functions themselves. These is a rich literature on classification and density ratio estimation, with many powerful algorithms even in high dimensional settings. Further discussion of estimating the conditional density ratio is provided in Section 3.5 when we summarize our algorithm.

### 3.3 Incorporating multiple testing sample points for better power

So far we have only focused on obtaining a single conformal $p$-value from a single testing sample point. Such a single $p$-value often has limited power in distinguishing $H_1$ from $H_0$. To have a test with asymptotically full power, we must consider multiple testing sample points: $((X_{2k}, Y_{2k}) : 1 \leq k \leq K)$, where $K$ is a positive integer whose relationship with the original testing sample size $n_2$ will be specified later.

Given $(X_{1i}, Y_{1i})_{i=1}^m$ from $P_1$, we can repeat the procedure used to obtain $U$ in (6) for each testing sample point $(X_{2k}, Y_{2k})$ for $1 \leq k \leq K$, resulting in $(U_k : 1 \leq k \leq K)$. Then each $U_k \sim U(0, 1)$ under $H_0$ and $U_k \sim U(0, 1)$ with $\mathbb{E} U_k < 1/2$ under $H_1$. However, we cannot simply construct a one-sided mean test over the $U_k$’s, because they are dependent as their constructions all involve
the same set of training data \((X_{1i}, Y_{1i})_{i=1}^m\). To make a one-sided mean test over the \(U_k\)’s work, we construct \(U_j\’s using separate batches of training data. That is, for each \(k\), \(U_k\) is constructed as in (6) using \((X_{2k}, Y_{2k})\) and \((X_{1i_k,l}, Y_{1i_k,l})_{l=1}^m\), where each \((X_{1i_k,l}, Y_{1i_k,l})_{l=1}^m\) is a disjoint batch of training data. By construction, the \(U_k\)’s are iid and the standard normal approximation theory would be valid in testing the mean of \(U_k\).

3.4 Allowing for \(P_{1,X} \neq P_{2,X}\) using weighted conformalization

Now we drop the assumption of equal marginal distribution of \(X\) under \(P_{1}\) and \(P_{2}\). Again we use the simplified notation, where the data points \((X_i, Y_i)_{i=1}^{m+1}\) are independent with \((X_i, Y_i) \sim P_1\) for \(1 \leq i \leq m\) and \((X_{m+1}, Y_{m+1}) \sim P_2\). Now the \((m+1)\)-tuple used to construct \(U\) in (6) are no longer exchangeable and the distribution of \(U\) will in general not be uniform. Here we use the “weighted conformal prediction” idea developed in Tibshirani et al. (2019) to obtain a modified version of \(U\) with valid uniform sampling distribution under \(H_0\). The key idea is to condition on a randomly permuted data sequence.

In the subsequent discussion, we will focus on conformity score functions \(V\) that only depends on the last argument, and write \(V_i = V(X_i, Y_i)\).

We begin by assuming the data \(Z = (X_i, Y_i)_{i=1}^{m+1}\) is stored in two parts: a randomly permuted sequence \(\tilde{Z} = (\tilde{X}_i, \tilde{Y}_i)_{i=1}^{m+1}\), and the permutation \(\sigma: [m+1] \mapsto [m+1]\) with the correspondence \((X_i, Y_i) \leftrightarrow (\tilde{X}_{\sigma(i)}, \tilde{Y}_{\sigma(i)})\). By construction, the vector \((V_i: 1 \leq i \leq m+1)\) is a deterministic function of \((\tilde{Z}, \sigma)\). Given \(\tilde{Z}\), \(V_{m+1}\) may take \(m+1\) possible values, and \(V_{m+1} = V(\tilde{X}_i, \tilde{Y}_i)\) if \(\sigma(m+1) = i\).

Now we are ready to derive the conditional distribution of \(V_{m+1}\) given \(\tilde{Z}\), construct the uniformly distributed weighted conformal p-value, and establish its ability to separate \(H_0\) and \(H_1\).

**Lemma 2.** 1. Under \(H_0\), for any choice of \(V(x,y)\) we have

\[
(V_{m+1}|\tilde{Z}) \sim \sum_{i=1}^{m+1} p_i(\tilde{Z}) \delta_{V_i}
\]

with

\[
p_i(\tilde{Z}) = \frac{f_{2,X}(X_i) f_{1,X}(X_i)}{\sum_{l=1}^{m+1} f_{2,X}(X_l) f_{1,X}(X_l)}, \quad i = 1, ..., m+1,
\]

and \(\delta_v\) denotes the point mass at \(v\).
2. For any choice of \( V(x, y) \), the randomized statistic

\[
U = \sum_{i=1}^{m+1} p_i(Z) \mathbb{1}(V_i < V_{m+1}) + \zeta \sum_{i=1}^{m+1} p_i(Z) \mathbb{1}(V_i = V_{m+1})
\]  

has a uniform distribution under \( H_0 \), where \( \zeta \) is an \( U(0, 1) \) random variable independent of everything else.

3. Under \( H_1 \), if \( V(x, y) = \frac{f_1(y|x)}{f_2(y|x)} \), there exist \( \delta > 0 \) and \( m_0 \), depending only on \( P_1, P_2 \), such that

\[ \mathbb{E}U \leq 1/2 - \delta \text{ when } m \geq m_0. \]

The definitions of \( U \) in (7) and (6) are compatible, as the construction with random tie-breaking in (6) can be viewed as a special case of (7) with \( p_i(Z) = (m + 1)^{-1} \).

Part (a) of Lemma 2 is due to Tibshirani et al. (2019), who first considered weighted conformal prediction under covariate shift. Part (b) is a simple consequence of part (a), which can be viewed as a discrete version of the CDF transform. The most non-obvious part of the proof is that of part (c), which exploits the form of \( V(x, y) = \frac{f_1(y|x)}{f_2(y|x)} \). The detailed proofs are given in Appendix C.

3.5 The conformal covariate shift test algorithm

Given the ideas and methods presented in the previous subsections, we can now describe the full testing procedure in Algorithm 1 below. The algorithm assumes availability of two subroutines for estimating the density ratios \( V(x, y) = \frac{f_1(y|x)}{f_2(y|x)} \), and \( g(x) = \frac{f_{2,X}(x)}{f_{1,X}(x)} \). In particular, we assume \( A_1 \) is an algorithm that takes input of two labeled samples \( \{ (X_{1i}, Y_{1i}) : 1 \leq i \leq n_{11} \} \) and \( \{ (X_{2j}, Y_{2j}) : 1 \leq j \leq n_{21} \} \), and outputs an estimate of the function \( V \). On the other hand, \( A_2 \) is a marginal density ratio estimator that takes input of two labeled samples \( \{ X_{1i} : 1 \leq i \leq n_{11} \} \), \( \{ X_{2j} : 1 \leq j \leq n_{21} \} \), and outputs an estimate of the marginal density ratio \( g \).

Given \( A_1, A_2 \), the testing procedure first splits the sample, applying the density ratio estimation subroutines \( A_1, A_2 \) on the first part to obtain approximate versions of the density ratios. Then the second part is further split into mini batches to construct a group of conformal \( p \)-values \( \hat{U}_k \) according to (7) by plugging-in the estimated density ratios. The final test statistic is then obtained from a standard one-sample one-sided mean test applied to the collection of \( \hat{U}_k \)'s.

Theoretical results of the asymptotic behavior of the test statistic \( T \) under \( H_0 \) and \( H_1 \), along with sufficient conditions guaranteeing the control of type I and type II errors, are presented in
**Algorithm 1** Conformal test of covariate shift

**Require:** Training data \((X_{1i}, Y_{1i})_{i=1}^{n_1}\); testing data \((X_{2j}, Y_{2j})_{j=1}^{n_2}\); number of minibatches \(K\), mini-batch size \(m\); density ratio estimation subroutines \(A_1, A_2\)

Randomly split \(\{1, \ldots, n_1\}\) into subsets \(I_{11}, I_{12}\) such that \(|I_{12}| = mK\)

Randomly split \(\{1, \ldots, n_2\}\) into subsets \(I_{21}, I_{22}\) such that \(|I_{22}| = K\)

Randomly split \(I_{12}\) into \(K\) equal-sized batches \(J_k, k = 1, \ldots, K\)

\(\hat{V}(\cdot, \cdot) = A_1[\{(X_{1i}, Y_{1i}), i \in I_{11}, (X_{2j}, Y_{2j}), j \in I_{21}\}]\)

\(\hat{g}(\cdot) = A_2[\{(X_{1i}, i \in I_{11}, X_{2j}, j \in I_{21}\}]

for \(k \in I_{22}\) and \(J_k = \{i_{k,1}, \ldots, i_{k,m}\}\) do

for \(l \in \{1, \ldots, m\}\) do

\(\hat{V}_{k,l} = \hat{V}(X_{1i_{k,l}}, Y_{1i_{k,l}})\) and \(\hat{p}_{k,l} = \hat{g}(X_{1i_{k,l}})/\{\sum_{l=1}^{m} \hat{g}(X_{1i_{k,l}}) + \hat{g}(X_{2k})\}\)

end for

\(\hat{V}_{k,m+1} = \hat{V}(X_{2k}, Y_{2k})\) and \(\hat{p}_{k,m+1} = \hat{g}(X_{2k})/\{\sum_{l=1}^{m} \hat{g}(X_{1i_{k,l}}) + \hat{g}(X_{2k})\}\)

Generate \(\zeta_k \sim U(0,1)\), independent of everything else

\(\hat{U}_k = \sum_{l=1}^{m} \hat{p}_{k,l} I(\hat{V}_{k,l} < \hat{V}_{k,m+1}) + \zeta_k \{\hat{p}_{k,m+1} + \sum_{l=1}^{m} \hat{p}_{k,l} I(\hat{V}_{k,l} = \hat{V}_{k,m+1})\}\)

end for

\(T = \sqrt{12K} \left(1/2 - \sum_{k=1}^{K} \hat{U}_k/K\right)\)

Reject \(H_0\) if \(T \geq \Phi^{-1}(1 - \alpha)\) (\(\Phi\) is the CDF of \(N(0,1)\), \(\alpha\) is the nominal type I error level)
Section 4. Empirical performance in simulation and a real data example will be presented in Section 5 and Section 6 respectively. Here we discuss some practical aspects of the algorithm.

Simplification when the marginals are equal. Sometimes it is plausible to assume that the marginal distributions $P_{1,X}, P_{2,X}$ are equal. This can happen, for example, when the sampling schemes and environments of $X$ are known to be the same, or they come from the same experimental design. In this scenario, the algorithm becomes much simpler, as we know that $f_{1,X}/f_{2,X} \equiv 1$. As a result, the algorithm does not need to use the marginal density ratio subroutine $A_2$ and can assign all $\hat{p}_{k,l}$ the theoretical value $1/(m + 1)$.

Choice of the number of minibatches and batch size. The algorithm requires $K$ minibatches to construct independent conformal $p$-values, which costs $K$ sample points from the testing data and $mK$ from the training data, and leaves $n_2 - K$ testing sample points and $n_1 - mK$ training sample points for estimating the density ratios. Our theory suggests that the test will be more powerful in detecting $H_1$ when $m$ is large. In practice we found the power to be reasonable for moderate $m$ values such as $m = 10$. But $m = 50$ could also be used if permitted by a large value of $n_1$. The conformal $p$-value sample size $K$ needs to be large enough for the normal approximation of the final test statistic $T$ to be accurate, but cannot be too large to be sensitive to the estimation error carried in the density ratios. In practice we recommend setting $K$ to be a small fraction of $n_2$ such as $n_2/\log(n_2)$ while not exceeding $n_1/(2m)$.

Choice of classification algorithms. As mentioned in Remark 2, our method does not require estimating the densities $f_{j,X}(\cdot)$ or conditional densities $f_j(\cdot|\cdot)$ for $j = 1, 2$. Instead, it only requires estimating, up to an arbitrary constant multiplier, the conditional density ratio $f_1(y|x)/f_2(y|x)$, and in the case $P_{1,X} \neq P_{2,X}$, the marginal density ratio $f_{1,X}(x)/f_{2,X}(x)$. The problem of estimating density ratios is often easier than estimating the density functions themselves, and has been well studied in the statistics and machine learning literature, including moment matching approach (Gretton et al., 2009), the ratio matching approach (Sugiyama et al., 2008; Kanamori et al., 2009; Tsuboi et al., 2009), and probabilistic classification approach (Qin, 1998; Cheng and Chu, 2004; Bickel et al., 2007).
Our algorithm can be implemented with any available density ratio estimators. Here we provide some further detail about the probabilistic classification estimator due to its simplicity. In the case of $P_{1,X} = P_{2,X}$, we only need to consider a single classification problem over the joint distribution $(X,Y)$, where class “1” represents the training data, and class “2” represents the testing data. Let $\eta(x,y)$ be the true conditional probability $P(1|x,y)$, then $f_1(x,y)/f_2(x,y) \propto \eta(x,y)/(1 - \eta(x,y))$, which also equals $f_1(y|x)/f_2(y|x)$ since $f_{1,X} = f_{2,X}$. When $P_{1,X} \neq P_{2,X}$, we can consider an additional classification problem using only $X$. Let $\eta(x) = P(1|x)$, then $f_{2,X}(x)/f_{1,X}(x) = (1 - \eta(x))/\eta(x)$. With probabilistic classifiers providing $\hat{\eta}(x,y)$ and $\hat{\eta}(x)$, the corresponding joint and marginal density ratios can be estimated by plugging in $\hat{\eta}(x,y)$ and $\hat{\eta}(x)$. The conditional density ratio can be obtained by taking a further ratio between the joint and marginal density ratios. Many commonly used classification methods offer a probability output, including the classical linear and quadratic discriminant analysis, logistic regression, popular machine learning algorithms such as random forest and support vector machines (Sollich, 2000), and modern deep neural nets.

4. ASYMPTOTIC PROPERTIES

In this section, we investigate the theoretical properties of the testing procedure described in Algorithm 1 under an asymptotic framework. We consider three quantities that play different roles in determining different sample sizes: the fitting subsample size $n \equiv \min(n_1 - mK, n_2 - K)$; the ranking batch size $m$; and the number of minibatches, $K$. We will show that the asymptotic null distribution is valid under no additional assumptions when $P_{1,X} = P_{2,X}$, and requires a consistent estimator of the marginal density ratio $f_{2,X}(x)/f_{1,X}(x)$ when $P_{1,X} \neq P_{2,X}$. On the other hand, the asymptotic power guarantee additionally requires a consistent estimator of the conditional density ratio. Designing a high quality density ratio estimator is a rich and context-dependent topic, and is beyond the scope of this paper. Therefore, some of our assumptions will be the asymptotic accuracy of such density ratio estimators.

Recall that we use the notation $g(x) = f_{2,X}(x)/f_{1,X}(x)$ and $V(x,y) = f_1(y|x)/f_2(y|x)$. We denote the subsample used to fit $\hat{V}$, $\hat{g}$ as the fitting subsample:

$$D_{\text{fit}} = \{(X_{1i},Y_{1i}) : i \in I_{11}\}, \{(X_{2j},Y_{2j}) : j \in I_{21}\}.$$ 

For $q \geq 1$, $j = 1,2$, denote the $\ell_q$ norm of a function $\gamma$ under distribution $P_j$ by $||\gamma||_{q,P_j}^q \equiv$
Our first assumption puts some moment conditions on the marginal density ratio $g(\cdot)$, which enables us to control the behavior of the weights $g(X_i)/\sum_{l=1}^{m+1} g(X_i)$, whose empirical version corresponds to the probability weight $\hat{p}_{k,l}$ in Algorithm 1.

**Assumption 1.** The marginal likelihood ratio $g(x) = f_{2,X}(x)/f_{1,X}(x)$ satisfies $\|g\|_{2,P_j} < \infty$ and $\|g^{-1}\|_{2,P_j} < \infty$ for $j = 1, 2$.

Our next assumption quantifies the required accuracy of the density ratio estimators.

**Assumption 2.** When $n = \min\{n_1 - mK, n_2 - K\} \to \infty, K \to \infty$ we have,

(a) \[ \|\hat{V} - V\|_{1,P_2} = o_P(1). \] \hspace{1cm} (8)

(b) \[ r_n \equiv \max\{\|\hat{g} - g\|_{2,P_1}, \|\hat{g} - g\|_{2,P_2}\} = o_P(K^{-1/2}). \] \hspace{1cm} (9)

Assumption 2(a) is necessary to guarantee the power under $H_1$ since it ensures the estimated conformity score function is asymptotically close to optimal. Assumption 2(b) is only needed to establish the asymptotic null distribution when $P_{1,X} \neq P_{2,X}$, as it ensures the estimated weights $\hat{p}_{k,l}$ behave similarly as the true weights. A weaker version of Assumption 2(b) is needed for the power guarantee, where $r_n$ only needs to be $o_P(1)$.

**Asymptotic null distribution.** Next we present our main theorem on the asymptotic null distribution of the test statistic output by Algorithm 1.

**Theorem 3.** Suppose that Assumptions 1 and 2(b) hold. For any estimate $\hat{V}$ and any value of $m$, the statistic $T$ output by Algorithm 1 converges in distribution to the standard normal as $n, K \to \infty$ under $H_0$.

As a consequence, given a nominal type I error level $\alpha$, the test rule which rejects $H_0$ when $T \geq \Phi^{-1}(1 - \alpha)$ has type I error controlled asymptotically at the nominal level.

Theorem 3 only requires the marginal density ratio $g$ to be accurately estimated, while having no requirement on the estimated conditional density ratio $\hat{V}$, and ranking minibatch size $m$. This
is a feature of conformal prediction, where the validity of the conformal p-values only requires weighted exchangeability. In the case $P_{1,X} = P_{2,X}$, we have $\hat{g} \equiv 1$ and both Assumption 1 and Assumption 2(b) hold trivially. When $P_{1,X} \neq P_{2,X}$ the marginal density ratio $g$ needs to be estimated by classification, one can use the out-of-sample classification error as a proxy to the accuracy of $\hat{g}$. If the out-of-sample classification error rate is high, then the type I error control should be treated with caution. This is illustrated in our synthetic and real data examples in Sections 5 and 6.

**Asymptotic power guarantee.** If the density ratios $V, g$ are accurately estimated, then the test is powerful against any alternative.

**Theorem 4.** Suppose that Assumptions 1, 2 hold. There exists constants $C > 0, \delta > 0$ and $m_0$ depending only on $P_1, P_2$, such that under $H_1$, the empirical conformal p-values $\hat{U}_k$ in Algorithm 1 satisfies $E(\hat{U}_k | D_{\text{fit}}) \leq 1/2 - \delta + Cr_n$ when $m \geq m_0$. As a consequence, when $m \geq m_0$ the test statistic $T$ output by Algorithm 1 satisfies $T \overset{p}{\to} +\infty$ as $n, K \to \infty$.

As a consequence, for any nominal level $\alpha \in (0, 1)$, the probability of rejecting the null hypothesis $H_0$ tends to one under the alternative hypothesis $H_1$.

Unlike the null case, Theorem 4 requires $m$ to be large enough in order to detect the alternative. This is needed in our proof to provide a lower bound of the density ratio sum $\sum_{i=1}^m g(X_i)$. It can be removed at the cost of more stringent assumptions such as $g(x)$ being bounded away from 0 for all $x$. Moreover, Theorem 4 only requires Assumption 2(b) to hold with $r_n = o_P(1)$ instead of $r_n = o_P(K^{-1/2})$. The proof of Theorem 4 is non-trivial due to the need to control the differences in indicator functions under perturbation (Lemma 8(b)). The details are given in Appendices B and C.

5. SIMULATION STUDY

In this section, we illustrate the performance of our method in several simulation settings. For brevity, we focus on the more challenging and interesting case where the $X$-marginals are different. Denote $x_i = (x_i(1), \ldots, x_i(p))^T$, and $I_p$ is a $p \times p$ identity matrix. We consider four prototypical regression models with $p = 5$ that are similar to those in Lei et al. (2018) and Zheng (2000). A higher dimensional case is presented in Section 5.2.
Model A (Gaussian, linear). Let \( y_{\ell i} = \alpha_\ell + \beta^T x_{\ell i} + \epsilon_{\ell i}, i = 1, \ldots, n_\ell, \ell = 1, 2 \), where \( x_{\ell i} \overset{iid}{\sim} N(0, I_p) \), \( x_{2i} \overset{iid}{\sim} N(\mu, I_p) \) where \( \mu = (1, 1, -1, -1, 0)^T \), and \( \epsilon_{1i}, \epsilon_{2i} \overset{iid}{\sim} N(0, 1) \), independent of the features. Set \( \alpha_1 = \alpha_2 = 0 \) under the null and \( \alpha_1 = 0, \alpha_2 = 0.5 \) under the alternative.

Model B (nonlinear, heavy-tailed). Let \( y_{\ell i} = \alpha_\ell + \beta_1 x_{\ell i}(1) + \beta_2 x_{\ell i}(2) + \beta_3 x_{\ell i}^2(3) + \beta_4 x_{\ell i}^2(4) + \beta_5 x_{\ell i}^3(5) + \epsilon_{\ell i}, i = 1, \ldots, n_\ell, \ell = 1, 2 \), where \( x_{1i} \overset{iid}{\sim} N(0, I_p) \), \( x_{2i} \overset{iid}{\sim} N(\mu, I_p) \) where \( \mu = (1, 1, -1, -1, 0)^T \), and \( \epsilon_{1i}, \epsilon_{2i} \overset{iid}{\sim} t(5) \), the student’s t-distribution with 5 degrees of freedom, independent of the features. Set \( \alpha_1 = \alpha_2 = 0 \) under the null and \( \alpha_1 = 0, \alpha_2 = 1 \) under the alternative.

Model C (linear, heteroskedastic, correlated features). Let \( y_{\ell i} = \beta^T x_{\ell i} + \epsilon_{\ell i}, i = 1, \ldots, n_\ell, \ell = 1, 2 \), where \( x_{\ell i} \overset{iid}{\sim} N(\mu, \Sigma) \), \( x_{2i} \overset{iid}{\sim} N(\mu, \Sigma) \) where \( \mu = (1, 1, -1, -1, 0)^T \) and \( \Sigma \) is a \( p \times p \) positive definite matrix with \( \Sigma_{ii} = 1 \) and \( \Sigma_{ij} = 1/\max\{i, j\} \) if \( \neq j \). Set \( \epsilon_{\ell i} \sim N(0, 4/(1 + x_{\ell i}^2(1))) \), \( \ell = 1, 2 \) under the null and \( \epsilon_{1i} \sim N(0, 4/(1 + x_{1i}^2(1))) \), \( \epsilon_{2i} \sim N(0, 1/(1 + x_{2i}^2(1))) \) under the alternative. Here the noises are not independent of the covariates.

Model D (linear, quadratic shift) Let \( y_{\ell i} = \beta^T x_{\ell i} + \epsilon_{\ell i}, i = 1, \ldots, n_\ell, \ell = 1, 2 \), where \( x_{1i(j)} \overset{iid}{\sim} N(0, 1) \), \( x_{2i(j)} \overset{iid}{\sim} N(0, 2) \). Set \( \epsilon_{\ell i} \overset{iid}{\sim} N(0, 1) \) under the null, and \( \epsilon_{1i} \overset{iid}{\sim} N(0, 1) \), \( \epsilon_{2i} \overset{iid}{\sim} N(0, 2) \) under the alternative.

In order to make density ratio estimation stable, we remove sample points whose marginal density ratio \( f_{1,X}(x)/f_{2,X}(x) \) or the joint density ratio \( f_1(x, y)/f_2(x, y) \) are outside of the interval \([1/100, 100]\).

5.1 The low dimensional case

We first consider low-dimensional cases with \( p = 5 \), setting the regression coefficients values to \( \pm 1 \) with random signs. We consider different combinations of sample size \( n_2 = 200, 500, 1000, 5000 \) and ranking batch size \( m = 5, 10, 50 \) for each model. Since estimating the density ratios requires larger sample size than the number of minibatches as noted previously in Theorem 3, we set the number of minibatches \( K = \lceil n_2 / \log n_2 \rceil \), and set the fitting subsample size \( n = n_{11} = n_{21} = n_2 - K \). So overall \( n_1 = n_2 + (m - 1)K \). In addition, we use four different probabilistic classification methods including linear logistic (LL), quadratic logistic (QL), neural network (NN) and random forest (RF) to estimate the conditional density ratio \( V \) and marginal density ratio \( g \). For the neural network method, we use the sigmoid activation function and the stochastic gradient descent algorithm. In
simpler settings such as models A, B, and C under the null hypothesis, we use one hidden layer with ten nodes. In more complicated settings such as model D and joint density ratio estimation in models B, C under the alternative hypothesis, two hidden layers are used. The code for the simulation is available to ensure reproducibility.

All the simulation results in the following are computed over 1000 repetitions with nominal type I error level $\alpha = 0.05$. To evaluate the accuracy of estimated probability weights, we calculate the probability weight estimation error

$$E_{rr\hat{p}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{l=1}^{m+1} |\hat{p}_{k,l} - p_{k,l}|,$$

where $\hat{p}_{k,l}$ is defined in Algorithm 1, and $p_{k,l}$ is the corresponding population version.

The results for Models A and B are summarized in Tables 1-4. The results for Models C and D are qualitatively similar and are deferred to Appendix A. As predicted by and Lemma 2(b) and Theorem 3, regardless of the choices of $\hat{V}$ and $m$, the empirical sizes are always close to the nominal level $\alpha$ as long as the true marginal density ratio $g$ is used or a good estimate is plugged in. The LL estimator is the correctly specified parametric method, and has very good performance even with small sample sizes. The NN and QL estimators requires larger sample sizes to achieve comparable estimation accuracy, but still yield satisfying control of the type I error even under moderate sample sizes. In contrast, the RF estimator fails to control the type I error in all models considered, which is mainly due to the substantially higher error in estimating the probability weights $p_{k,l}$.

When the alternative hypothesis is true, the power increases as the sample size increases. In all settings $m = 5$ is large enough to distinguish between the null and alternative hypotheses. Moreover, a larger $m$ tends to improve the power, especially for small sample sizes. So when the sample size permits, larger $m$ should be used for better power. In Model A, the LL and NN methods yield comparable power against the oracle approach with theoretical weights even when the sample size is small, thanks to accurately estimated density ratios $V$ and $g$. In Model B, the density ratios are much more difficult to estimate. However, the NN method is flexible to capture the complicated relationships and yields superior power, especially when the sample size is moderately large. Although the LL and QL estimators are misspecified, they still have non-trivial power, as these estimators can still partially capture the difference between these two conditional distributions. As noted previously, the RF estimator fails to provide valid type I error control, so
its power has limited use.

5.2 The high dimensional case

The flexibility of choosing probabilistic classification algorithms makes our method applicable in high-dimensional problems. Here we illustrate its performance in a high-dimensional scenario, which is similar to Model A in the low dimensional case but with ambient dimensionality \( p = 500 \) and signal dimensionality \( s = 5 \). The additional coordinates are filled with zeros in \( \mu \) and \( \beta \). As discussed earlier, the random forest method does not give accurate probability estimate and hence not suitable for our method. The neural net method is expected to work but needs some careful construction and tuning in the high dimensional case, which is out of the scope of this paper. Here we focus on a sparse linear classifier and investigate the effect of tuning and regularization. Letting \( L \) be the class label with \( L = 1 \) for training data and \( L = 0 \) for testing data, we learn a sparse logistic regression model by minimizing

\[
- \sum_{i \in I} \left[ L_{ki} \log \eta(x_{ki}; \beta) + (1 - L_{ki}) \log \{1 - \eta(x_{ki}; \beta)\} \right] + \lambda \| \beta \|_1,
\]

for \( k = 1, 2 \), where \( \beta = (\beta_0, \beta_1, \ldots, \beta_p) \) and \( \eta(x; \beta) = P(L = 1|x) = 1/[1 + \exp\{\beta_0 + \sum_{j=1}^p \beta_j x(j)\}] \).

Then we obtain the marginal ratio estimator up to a constant with \( \hat{g}(x) \propto \{1 - \eta(x; \hat{\beta})\} / \eta(x; \hat{\beta}) \).

In a similar manner, we can estimate the joint density ratio and hence the conditional density ratio.

The empirical rejection frequency and estimation errors of the probability weights \( p_{k,l} \) are shown in Figure 1 and Figure 2. Since the estimation errors are not observable in practice, we plot the out-of-sample marginal classification error (MCE) in the classification problem involved in estimating \( \hat{g} \) (the solid lines with star-shaped marks). Moreover, under the alternative, we also report the empirical out-of-sample estimation error of the conditional density ratio \( V \), defined by

\[
Err_{\hat{V}} = \{(m + 1)K\}^{-1} \sum_{k=1}^K \sum_{l=1}^{m+1} (\hat{V}_{k,l} - V_{k,l})^2.
\]

Again, when the true marginal density ratios are used, the empirical sizes are close to the nominal level \( \alpha = 0.05 \) as expected. When the probability weights are estimated, the type I error is well controlled for a wide range of tuning parameter values, indicating good robustness of validity. The plot of out-of-sample marginal classification error (MCE) suggests that in practice one can choose the tuning parameter value near the elbow of the error plot. Under the alternative
Table 1: Percentage of rejections and probability weight estimation error in Model A under the null over 1000 repetitions.

|                      | Percentage of rejections |                      |                      | Err ñ̂ |                      |                      |
|----------------------|---------------------------|----------------------|----------------------|--------|----------------------|----------------------|
|                      | m=5 m=10 m=50             | m=5 m=10 m=50        | m=5 m=10 m=50        |        |                      |                      |
|                      | ñ̂_LL + g                 | ñ̂_LL + ñ̂_LL        |                      | ñ̂_LL  |                      |                      |
| n_2 = 200            | 0.058 0.041 0.054         | 0.044 0.055 0.054    | 0.134 0.174 0.241    |        |                      |                      |
| n_2 = 500            | 0.045 0.060 0.055         | 0.055 0.055 0.065    | 0.081 0.104 0.147    |        |                      |                      |
| n_2 = 1000           | 0.047 0.054 0.053         | 0.047 0.053 0.049    | 0.057 0.074 0.104    |        |                      |                      |
| n_2 = 5000           | 0.046 0.052 0.055         | 0.056 0.061 0.057    | 0.025 0.033 0.046    |        |                      |                      |
|                      | ñ̂_QL + g                 | ñ̂_QL + ñ̂_QL        |                      | ñ̂_QL  |                      |                      |
| n_2 = 200            | 0.043 0.051 0.058         | 0.056 0.057 0.060    | 0.194 0.254 0.367    |        |                      |                      |
| n_2 = 500            | 0.045 0.050 0.060         | 0.045 0.055 0.065    | 0.118 0.152 0.220    |        |                      |                      |
| n_2 = 1000           | 0.047 0.053 0.062         | 0.047 0.051 0.062    | 0.081 0.106 0.155    |        |                      |                      |
| n_2 = 5000           | 0.051 0.060 0.056         | 0.052 0.060 0.056    | 0.034 0.044 0.063    |        |                      |                      |
|                      | ñ̂_NN + g                 | ñ̂_NN + ñ̂_NN        |                      | ñ̂_NN  |                      |                      |
| n_2 = 200            | 0.058 0.050 0.056         | 0.049 0.056 0.077    | 0.153 0.200 0.284    |        |                      |                      |
| n_2 = 500            | 0.040 0.051 0.055         | 0.055 0.042 0.060    | 0.109 0.145 0.215    |        |                      |                      |
| n_2 = 1000           | 0.047 0.054 0.049         | 0.062 0.052 0.061    | 0.093 0.123 0.187    |        |                      |                      |
| n_2 = 5000           | 0.050 0.056 0.053         | 0.048 0.052 0.056    | 0.063 0.084 0.130    |        |                      |                      |
|                      | ñ̂_RF + g                 | ñ̂_RF + ñ̂_RF        |                      | ñ̂_RF  |                      |                      |
| n_2 = 200            | 0.050 0.052 0.039         | 0.104 0.139 0.215    | 0.429 0.534 0.669    |        |                      |                      |
| n_2 = 500            | 0.045 0.035 0.055         | 0.161 0.256 0.417    | 0.365 0.462 0.600    |        |                      |                      |
| n_2 = 1000           | 0.049 0.051 0.047         | 0.302 0.357 0.690    | 0.333 0.422 0.570    |        |                      |                      |
| n_2 = 5000           | 0.060 0.041 0.056         | 0.778 0.833 1.000    | 0.270 0.352 0.494    |        |                      |                      |
Table 2: Percentage of rejections and probability weight estimation error in Model A under the alternative over 1000 repetitions.

|                  | Percentage of rejections | $Err_{\hat{p}}$ |
|------------------|--------------------------|-----------------|
|                  | $m=5$ | $m=10$ | $m=50$ | $m=5$ | $m=10$ | $m=50$ | $m=5$ | $m=10$ | $m=50$ |
| $V_{LL} + g$     |       |        |        |       |        |        |       |        |        |
| $n_2 = 200$      | 0.320 | 0.466  | 0.759  | 0.320 | 0.478  | 0.750  | 0.138 | 0.178  | 0.247  |
| $n_2 = 500$      | 0.469 | 0.743  | 0.975  | 0.508 | 0.741  | 0.978  | 0.081 | 0.104  | 0.147  |
| $n_2 = 1000$     | 0.747 | 0.923  | 0.998  | 0.685 | 0.930  | 1.000  | 0.057 | 0.074  | 0.104  |
| $n_2 = 5000$     | 1.000 | 1.000  | 1.000  | 1.000 | 1.000  | 1.000  | 0.024 | 0.032  | 0.045  |
| $V_{QL} + g$     |       |        |        |       |        |        |       |        |        |
| $n_2 = 200$      | 0.149 | 0.205  | 0.334  | 0.152 | 0.180  | 0.314  | 0.203 | 0.261  | 0.377  |
| $n_2 = 500$      | 0.370 | 0.516  | 0.801  | 0.363 | 0.499  | 0.769  | 0.116 | 0.152  | 0.220  |
| $n_2 = 1000$     | 0.668 | 0.816  | 0.987  | 0.610 | 0.806  | 0.978  | 0.081 | 0.106  | 0.153  |
| $n_2 = 5000$     | 0.974 | 1.000  | 1.000  | 1.000 | 1.000  | 1.000  | 0.035 | 0.047  | 0.068  |
| $V_{NN} + g$     |       |        |        |       |        |        |       |        |        |
| $n_2 = 200$      | 0.297 | 0.458  | 0.714  | 0.332 | 0.498  | 0.778  | 0.155 | 0.201  | 0.286  |
| $n_2 = 500$      | 0.496 | 0.678  | 0.966  | 0.508 | 0.740  | 0.981  | 0.107 | 0.142  | 0.213  |
| $n_2 = 1000$     | 0.720 | 0.903  | 0.997  | 0.759 | 0.910  | 0.998  | 0.089 | 0.120  | 0.182  |
| $n_2 = 5000$     | 1.000 | 1.000  | 1.000  | 1.000 | 1.000  | 1.000  | 0.059 | 0.079  | 0.122  |
| $V_{RF} + g$     |       |        |        |       |        |        |       |        |        |
| $n_2 = 200$      | 0.092 | 0.116  | 0.165  | 0.186 | 0.281  | 0.452  | 0.434 | 0.531  | 0.669  |
| $n_2 = 500$      | 0.132 | 0.190  | 0.306  | 0.356 | 0.498  | 0.826  | 0.365 | 0.459  | 0.605  |
| $n_2 = 1000$     | 0.226 | 0.302  | 0.492  | 0.603 | 0.794  | 0.963  | 0.329 | 0.421  | 0.566  |
| $n_2 = 5000$     | 0.662 | 0.883  | 1.000  | 0.987 | 1.000  | 1.000  | 0.270 | 0.351  | 0.492  |
Table 3: Percentage of rejections and probability weight estimation error in Model B under the null over 1000 repetitions.

|                  | Percentage of rejections |                  |                  |
|------------------|--------------------------|------------------|------------------|
|                  | m=5 m=10 m=50           | m=5 m=10 m=50    | m=5 m=10 m=50    |
|                  | \( \hat{V}_{LL} + g \) | \( \hat{V}_{LL} + \hat{g}_{LL} \) | \( Err_{\hat{\rho},LL} \) |
| \( n_2 = 200 \) | 0.067 0.053 0.044       | 0.056 0.044 0.056 | 0.124 0.165 0.228 |
| \( n_2 = 500 \) | 0.046 0.056 0.057       | 0.041 0.047 0.051 | 0.082 0.106 0.149 |
| \( n_2 = 1000 \)| 0.051 0.050 0.059       | 0.058 0.050 0.056 | 0.057 0.074 0.104 |
| \( n_2 = 5000 \)| 0.053 0.055 0.053       | 0.056 0.049 0.052 | 0.025 0.033 0.046 |
|                  | \( \hat{V}_{QL} + g \) | \( \hat{V}_{QL} + \hat{g}_{QL} \) | \( Err_{\hat{\rho},QL} \) |
| \( n_2 = 200 \) | 0.047 0.044 0.058       | 0.052 0.052 0.069 | 0.186 0.248 0.372 |
| \( n_2 = 500 \) | 0.041 0.052 0.062       | 0.062 0.052 0.057 | 0.118 0.156 0.222 |
| \( n_2 = 1000 \)| 0.049 0.059 0.058       | 0.054 0.048 0.059 | 0.081 0.105 0.153 |
| \( n_2 = 5000 \)| 0.053 0.048 0.053       | 0.048 0.053 0.054 | 0.038 0.049 0.071 |
|                  | \( \hat{V}_{NN} + g \) | \( \hat{V}_{NN} + \hat{g}_{NN} \) | \( Err_{\hat{\rho},NN} \) |
| \( n_2 = 200 \) | 0.044 0.053 0.047       | 0.056 0.064 0.067 | 0.145 0.193 0.273 |
| \( n_2 = 500 \) | 0.051 0.047 0.052       | 0.067 0.052 0.051 | 0.110 0.147 0.217 |
| \( n_2 = 1000 \)| 0.045 0.053 0.054       | 0.059 0.070 0.071 | 0.086 0.116 0.179 |
| \( n_2 = 5000 \)| 0.052 0.051 0.053       | 0.051 0.053 0.050 | 0.061 0.083 0.128 |
|                  | \( \hat{V}_{RF} + g \) | \( \hat{V}_{RF} + \hat{g}_{RF} \) | \( Err_{\hat{\rho},RF} \) |
| \( n_2 = 200 \) | 0.067 0.052 0.069       | 0.111 0.089 0.322 | 0.423 0.518 0.660 |
| \( n_2 = 500 \) | 0.041 0.067 0.047       | 0.181 0.311 0.420 | 0.364 0.460 0.608 |
| \( n_2 = 1000 \)| 0.039 0.054 0.051       | 0.259 0.396 0.683 | 0.329 0.421 0.565 |
| \( n_2 = 5000 \)| 0.060 0.053 0.044       | 0.632 0.842 1.000 | 0.269 0.354 0.494 |
Table 4: Percentage of rejections and probability weight estimation error in Model B under the alternative over 1000 repetitions.

|                | Percentage of rejections | $Err_{\hat{\beta}}$ |
|----------------|--------------------------|---------------------|
|               | m=5 m=10 m=50           | m=5 m=10 m=50       |
|                | $\hat{V}_{LL} + g$      | $\hat{V}_{LL} + \hat{g}_{LL}$ | $Err_{\hat{\beta},LL}$ |
| $n_2 = 200$   | 0.211 0.314 0.501       | 0.233 0.328 0.492   | 0.131 0.171 0.243       |
| $n_2 = 500$   | 0.340 0.496 0.787       | 0.340 0.511 0.774   | 0.082 0.106 0.150       |
| $n_2 = 1000$  | 0.537 0.759 0.952       | 0.523 0.747 0.961   | 0.058 0.075 0.105       |
| $n_2 = 5000$  | 0.980 1.000 1.000       | 0.980 1.000 1.000   | 0.027 0.035 0.050       |
|                | $\hat{V}_{QL} + g$      | $\hat{V}_{QL} + \hat{g}_{QL}$ | $Err_{\hat{\beta},QL}$ |
| $n_2 = 200$   | 0.147 0.236 0.319       | 0.141 0.193 0.260   | 0.191 0.251 0.370       |
| $n_2 = 500$   | 0.355 0.464 0.685       | 0.338 0.451 0.672   | 0.116 0.151 0.221       |
| $n_2 = 1000$  | 0.552 0.786 0.944       | 0.561 0.748 0.948   | 0.082 0.108 0.156       |
| $n_2 = 5000$  | 1.000 1.000 1.000       | 0.980 1.000 1.000   | 0.036 0.047 0.068       |
|                | $\hat{V}_{NN} + g$      | $\hat{V}_{NN} + \hat{g}_{NN}$ | $Err_{\hat{\beta},NN}$ |
| $n_2 = 200$   | 0.157 0.229 0.336       | 0.165 0.185 0.278   | 0.149 0.196 0.277       |
| $n_2 = 500$   | 0.442 0.605 0.827       | 0.394 0.555 0.797   | 0.110 0.146 0.215       |
| $n_2 = 1000$  | 0.825 0.959 1.000       | 0.787 0.959 1.000   | 0.090 0.121 0.183       |
| $n_2 = 5000$  | 1.000 1.000 1.000       | 1.000 1.000 1.000   | 0.060 0.080 0.126       |
|                | $\hat{V}_{RF} + g$      | $\hat{V}_{RF} + \hat{g}_{RF}$ | $Err_{\hat{\beta},RF}$ |
| $n_2 = 200$   | 0.130 0.196 0.268       | 0.252 0.366 0.605   | 0.420 0.523 0.665       |
| $n_2 = 500$   | 0.241 0.317 0.519       | 0.497 0.697 0.924   | 0.362 0.457 0.603       |
| $n_2 = 1000$  | 0.456 0.600 0.827       | 0.784 0.955 0.995   | 0.329 0.421 0.566       |
| $n_2 = 5000$  | 0.939 1.000 1.000       | 1.000 1.000 1.000   | 0.269 0.349 0.491        |
hypotheses, the power is maximized at tuning parameter values corresponding to the smallest estimation error in $\hat{V}$, which can also be chosen using its out-of-sample classification error plot (not shown in the plots).

6. REAL DATA

We consider the airfoil data set from the UCI Machine Learning Repository (Dua and Graff, 2019), which has $n = 1503$ observations of a response $Y$ (scaled sound pressure level of NASA airfoils), and a covariate $X$ with $p = 5$ dimensions (log frequency, angle of attack, chord length, free-stream velocity, and suction side log displacement thickness). This data set has been used by Tibshirani et al. (2019), who first studied conformalization under covariate shift.

The original data does not have a training and testing sample separation. We consider five experiments based on different ways to generate the training and testing datasets.

(i) Random partition. We randomly partition the dataset with $n_1 = 1127$ and $n_2 = 376$.

(ii) Random partition and exponential tilting. We first randomly partition the data into two sets $D_1$, $\tilde{D}_2$. Then following (Tibshirani et al., 2019), we construct $D_2$ by sampling 25% of the points from $\tilde{D}_2$ with replacement, with probabilities proportional to $w(x) = \exp(x^T\alpha)$, where $\alpha = (-1, 0, 0, 0, 1)$.

The final sample sizes are $n_1 = 601$ and $n_2 = 226$.

(iii) Chord-based partition. We split the dataset into two subsets where the values of the “chord” variable in $D_1$ are smaller than the 75% quantile and exclude the “chord” variable in subsequent analyses, resulting in $n_1 = 1049$ and $n_2 = 454$. To avoid singularity between the two populations, we randomly select $0.05n_1$ samples in $D_1$ and the same amount of samples in $D_2$ to flip their groups.

(iv) Velocity-based partition. We partition the dataset into two subsets where the values of the “velocity” variable in $D_1$ are smaller than the 75% quantile and remove the covariate “velocity” from subsequent analyses, with $n_1 = 1038$ and $n_2 = 465$. We randomly select $0.05n_1$ samples in $D_1$ and the same amount of samples in $D_2$ to flip their groups.
Figure 1: Comparison of empirical sizes in Model A under the null with $p = 500$ and $s = 5$, across a variety of base estimators and combinations of $m$ and $n_2$ over 1000 repetitions.
Figure 2: Comparison of empirical powers in Model A under the alternative with $p = 500$ and $s = 5$, across a variety of base estimators and combinations of $m$ and $n_2$ over 1000 repetitions.
Table 5: Percentage of rejections and test error in Airfoil data set for cases (i-ii) over 1000 repetitions.

|        | $\hat{V}_{LL} + \hat{g}_{LL}$ | $\hat{V}_{NN} + \hat{g}_{NN}$ | $\text{Err}_{\hat{p},LL}$ | $\text{Err}_{\hat{p},NN}$ |
|--------|-------------------------------|-------------------------------|-----------------------------|-----------------------------|
| case (i) |                               |                               |                             |                             |
| $m = 2$ | 0.054                         | 0.040                         | 0.091                       | 0.093                       |
| $m = 5$ | 0.056                         | 0.046                         | 0.105                       | 0.106                       |
| $m = 10$ | 0.038                        | 0.054                         | 0.121                       | 0.120                       |
| case (ii) |                              |                               |                             |                             |
| $m = 2$ | 0.048                         | 0.059                         | 0.081                       | 0.092                       |
| $m = 5$ | 0.069                         | 0.074                         | 0.129                       | 0.143                       |
| $m = 10$ | 0.067                         | 0.068                         | 0.183                       | 0.187                       |

(v) Response-based partition. We split the data according to the value of the response variable, where the first group contains the sample points with smaller response values, while $D_2$ contains the rest, with $n_1 = 1052$ and $n_2 = 451$. A similar label flipping is applied to avoid singularity.

As in the simulation study, we use $K = \lceil n_2 / \log n_2 \rceil$, $n_{21} = n_2 - K$, $n_{12} = mK$, and significance level $\alpha = 0.05$. Consider $m = 2, 5$ and 10, respectively. Linear logistic regression (LL) and neural network (NN) are used to estimate the density ratios. The neural network uses one hidden layer in cases (i-ii), and two hidden layers in cases (iii-v). Each experiment is repeated for 1000 trials.

For experiments (i) and (ii), which are clearly under the null hypothesis, we can re-generate the data by repeating the random generation of the training and testing data. With these repeatedly generated data sets, we can compute the empirical frequency of rejections. In cases (i-ii), the true probability weights are actually available, and we use them to compute the probability weights estimation error $\text{Err}_{\hat{p}}$ defined in (10). As shown in Table 5, the type I error is close to the nominal level.

For experiments (iii-v), we can only have a single deterministic generation of the training and testing data, except a small fraction of group flipping, and only a single $p$-value can be computed. We use these experiments to illustrate the effect of multiple realizations of auxiliary randomizations. Recall that our method uses auxiliary randomizations. One such randomization is the use of $\zeta$ in generating $U$ from the weighted ranking, and another randomization is splitting the data sets in to
Table 6: Median p-values and average classification error of estimating $g$ in Airfoil data set for cases (iii-v) over 1000 random splits.

|           | $\hat{V}_{LL} + \hat{g}_{LL}$ | $\hat{V}_{NN} + \hat{g}_{NN}$ | $MCE_{LL}$ | $MCE_{NN}$ |
|-----------|-------------------------------|-------------------------------|------------|------------|
| case (iii)|                               |                               |            |            |
| $m = 2$   | 0.269                         | 0.506                         | 0.195      | 0.071      |
| $m = 5$   | 0.167                         | 0.472                         | 0.151      | 0.061      |
| $m = 10$  | 0.087                         | 0.487                         | 0.287      | 0.064      |
| case (iv) |                               |                               |            |            |
| $m = 2$   | 0.003                         | 0.495                         | 0.333      | 0.091      |
| $m = 5$   | 0.000                         | 0.413                         | 0.198      | 0.081      |
| $m = 10$  | 0.000                         | 0.293                         | 0.744      | 0.171      |
| case (v)  |                               |                               |            |            |
| $m = 2$   | 0.000                         | 0.009                         | 0.329      | 0.155      |
| $m = 5$   | 0.000                         | 0.000                         | 0.234      | 0.127      |
| $m = 10$  | 0.000                         | 0.000                         | 0.371      | 0.196      |

fitting subsamples and ranking minibatches. Such auxiliary randomizations may lead to different results on the same dataset if the inference is carried out independently by different researchers. To mitigate this effect, one can obtain multiple p-values using multiple realizations of auxiliary randomizations. Although each single p-value has asymptotically valid null distribution and power, their dependence requires a careful aggregation of these p-values. Here we use a median p-value approach in DiCiccio et al. (2020). Formally, suppose we repeat the auxiliary randomization $B$ times, obtaining p-values $\hat{p}_1, ..., \hat{p}_B$. Then $\hat{p} = 1 \wedge [2 \times \text{Median}(\hat{p}_1, ..., \hat{p}_B)]$ is a valid p-value.

In experiments (iii-v), the true probability weights are not available and we use out-of-sample marginal classification error (MCE) as a proxy of the accuracy of probability weight estimation. In experiment (iii), both LL and NN methods give large p-values, and the NN method also has small marginal classification errors. Thus there is no strong evidence against the covariate shift assumption. In experiment (iv), the LL method gives small p-values while the NN method suggests otherwise. But the marginal classification errors indicate that the NN method is likely more accurate in estimating the marginal density ratios, and hence providing more trustworthy p-values. In experiment (v), both methods agree to reject the null hypothesis, which is the correct decision by the construction of training and testing samples. The neural net method also gives marginal
classification errors comparable to those in the null cases, further confirming the validity of $p$-value.

7. DISCUSSION

In applications it is often the case that the training data $(X_{1i}, Y_{1i})_{i=1}^{n_1}$ has a large sample size, whereas the testing data $(X_{2j}, Y_{2j})_{j=1}^{n_2}$ has a limited sample size. As our theory and experiments have demonstrated, a valid type I error control of the proposed test only depends on the accuracy of marginal density ratio estimation. Our method would be particularly useful in the semi-supervised scenario, where unlabeled testing sample points $X_{2j}$ are easy to obtain. In this case we can use these unlabeled testing sample points to estimate the marginal density ratio, and save the scarce labeled testing sample to estimate the joint density ratio.

The covariate shift hypothesis can be viewed as a special case of conditional independence given a binary random variable. Let $(X,Y,A)$ be a triplet of random variables, where $A$ is binary. Then the conditional distributions $(Y|X,A=0)$ and $(Y|X,A=1)$ are the same if and only if $Y$ and $A$ are conditionally independent given $X$. In some applications one may think of $Y$ as a response variable, $X$ the covariate, and $A$ a binary treatment. Then the method developed in this paper can be used to test the dependence between the treatment and response, conditioning on the covariate.

There are several potential directions to follow up. One is to consider specific directions of the alternative hypothesis, such as that the treatment is expected to increase the mean response by a certain amount. Another direction is to extend the method for binary $A$ to cover more general random variables, leading to new nonparametric conditional independence tests.

The use of sample splitting and auxiliary randomization for valid and efficient statistical inference has been studied by many authors in the high dimensional regression literature (Wasserman and Roeder, 2009; Meinshausen et al., 2009; Rinaldo et al., 2019), and more recently in the conformal inference literature (Kuchibhotla and Ramdas, 2019; Kim et al., 2020). The theory in Kim et al. (2020) also required an inflated non-coverage by a factor of two after aggregating multiple subsamples. It is unclear whether such a loss of coverage is unavoidable. It would be interesting and important to better understand the dependence between the $p$-values from multiple splits, and improve the current conservative inflation method when combining multiple $p$-values.
REFERENCES

Andrews, D. W. (1997), “A conditional Kolmogorov test,” *Econometrica: Journal of the Econometric Society*, 1097–1128.

Bai, J. (2003), “Testing parametric conditional distributions of dynamic models,” *Review of Economics and Statistics*, 85, 531–549.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2019), “Predictive inference with the jackknife+,” *arXiv preprint arXiv:1905.02928*.

Bickel, S., Brückner, M., and Scheffer, T. (2007), “Discriminative learning for differing training and test distributions,” in *Proceedings of the 24th international conference on Machine learning*, pp. 81–88.

— (2009), “Discriminative learning under covariate shift,” *Journal of Machine Learning Research*, 10, 2137–2155.

Cheng, K. F. and Chu, C.-K. (2004), “Semiparametric density estimation under a two-sample density ratio model,” *Bernoulli*, 10, 583–604.

Corradi, V. and Swanson, N. R. (2006), “Bootstrap conditional distribution tests in the presence of dynamic misspecification,” *Journal of Econometrics*, 133, 779–806.

Csurka, G. (2017), *Domain adaptation in computer vision applications*, vol. 2, Springer.

DiCiccio, C. J., DiCiccio, T. J., and Romano, J. P. (2020), “Exact tests via multiple data splitting,” *Statistics & Probability Letters*, 166, 108865.

Dua, D. and Graff, C. (2019), “UCI Machine Learning Repository;” .

Fan, Y., Li, Q., and Min, I. (2006), “A nonparametric bootstrap test of conditional distributions,” *Econometric Theory*, 22, 587–613.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009), “Covariate shift by kernel mean matching,” *Dataset shift in machine learning*, 3, 5.

Guan, L. and Tibshirani, R. (2019), “Prediction and outlier detection in classification problems,” *arXiv preprint arXiv:1905.04396*.

Hall, P. and Hart, J. D. (1990), “Bootstrap test for difference between means in nonparametric regression,” *Journal of the American Statistical Association*, 85, 1039–1049.

Hardle, W. and Marron, J. S. (1990), “Semiparametric comparison of regression curves,” *The
Annals of Statistics, 63–89.

Kanamori, T., Hido, S., and Sugiyama, M. (2009), “A least-squares approach to direct importance estimation,” Journal of Machine Learning Research, 10, 1391–1445.

Kim, B., Xu, C., and Barber, R. F. (2020), “Predictive Inference Is Free with the Jackknife+-after-Bootstrap,” arXiv preprint arXiv:2002.09025.

Kouw, W. M. and Loog, M. (2018), “An introduction to domain adaptation and transfer learning,” arXiv preprint arXiv:1812.11806.

Kuchibhotla, A. K. and Ramdas, A. K. (2019), “Nested Conformal Prediction and the Generalized Jackknife+,” arXiv preprint arXiv:1910.10562.

Kulasekera, K. (1995), “Comparison of regression curves using quasi-residuals,” Journal of the American Statistical Association, 90, 1085–1093.

Kulasekera, K. and Wang, J. (1997), “Smoothing parameter selection for power optimality in testing of regression curves,” Journal of the American Statistical Association, 92, 500–511.

Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018), “Distribution-free predictive inference for regression,” Journal of the American Statistical Association, 113, 1094–1111.

Lei, J., Rinaldo, A., and Wasserman, L. (2015), “A conformal prediction approach to explore functional data,” Annals of Mathematics and Artificial Intelligence, 74, 29–43.

Lei, J., Robins, J., and Wasserman, L. (2013), “Distribution-free prediction sets,” Journal of the American Statistical Association, 108, 278–287.

Lei, J. and Wasserman, L. (2014), “Distribution-free prediction bands for non-parametric regression,” Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76, 71–96.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), “P-values for high-dimensional regression,” Journal of the American Statistical Association, 104, 1671–1681.

Neumeyer, N. and Dette, H. (2003), “Nonparametric comparison of regression curves: an empirical process approach,” The Annals of Statistics, 31, 880–920.

Pan, S. J. and Yang, Q. (2009), “A survey on transfer learning,” IEEE Transactions on knowledge and data engineering, 22, 1345–1359.

Pardo-Fernández, J. C., Jiménez-Gamero, M. D., and El Ghouch, A. (2015), “Tests for the equality
of conditional variance functions in nonparametric regression,” *Electronic Journal of Statistics*, 9, 1826–1851.

Qin, J. (1998), “Inferences for case-control and semiparametric two-sample density ratio models,” *Biometrika*, 85, 619–630.

Rinaldo, A., Wasserman, L., G’Sell, M., et al. (2019), “Bootstrapping and sample splitting for high-dimensional, assumption-lean inference,” *The Annals of Statistics*, 47, 3438–3469.

Shimodaira, H. (2000), “Improving predictive inference under covariate shift by weighting the log-likelihood function,” *Journal of statistical planning and inference*, 90, 227–244.

Sollich, P. (2000), “Probabilistic methods for support vector machines,” in *Advances in neural information processing systems*, pp. 349–355.

Sugiyama, M. and Kawanabe, M. (2012), *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*, MIT press.

Sugiyama, M., Krauledat, M., and MĂžller, K.-R. (2007), “Covariate shift adaptation by importance weighted cross validation,” *Journal of Machine Learning Research*, 8, 985–1005.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P. V., and Kawanabe, M. (2008), “Direct importance estimation with model selection and its application to covariate shift adaptation,” in *Advances in neural information processing systems*, pp. 1433–1440.

Tibshirani, R. J., Barber, R. F., Candes, E., and Ramdas, A. (2019), “Conformal Prediction Under Covariate Shift,” in *Advances in Neural Information Processing Systems*, pp. 2526–2536.

Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., and Sugiyama, M. (2009), “Direct density ratio estimation for large-scale covariate shift adaptation,” *Journal of Information Processing*, 17, 138–155.

Vovk, V., Gammerman, A., and Shafer, G. (2005), *Algorithmic learning in a random world*, Springer Science & Business Media.

Wasserman, L. and Roeder, K. (2009), “High dimensional variable selection,” *Annals of statistics*, 37, 2178.

Zheng, J. X. (2000), “A consistent test of conditional parametric distributions,” *Econometric Theory*, 16, 667–691.
A. MORE SIMULATION RESULTS

We provide results for Models C and D in Tables 7-10. The observation for Model C is very similar to those in Models A and B, where all methods except RF obtain trustworthy estimates of marginal ratios and thus favorable empirical sizes under both the null and alternative.

In Model D, the QL is the correct model and works well, while NN requires a relatively large sample size to perform well. This is because in this case, the marginal density ratio $g$ is more difficult to estimate, as reflected by $Err_{g}$. When the sample size is moderately large, the estimators $\hat{g}_{NN}$ are reasonably accurate and the empirical sizes of NN are close to $\alpha$. The LL method gives correct empirical sizes, because $\hat{V}_{LL}$ and $\hat{g}_{LL}$ are approximately equal to 1, and the resulting conformal $p$-values $\hat{U}_{k}$ are trivially uniform. However, such a trivial estimate $\hat{V}_{LL}$ leads to poor power as it completely fails to capture the difference between the two distributions.

B. AUXILIARY LEMMAS

In the rest of the Appendix we provide proofs of main results and related auxiliary lemmas. The notation will mostly follow the main text, although some proofs involves their own notation. Throughout the proofs, we use $C$ to denote a constant whose value only depends on $(P_{1},P_{2})$ but not the sample sizes $(n_{1},n_{2},K,m)$. The value of $C$ may also change from line to line.

Our first auxiliary lemma provides an analogous CDF transformation for discrete random variable.

Lemma 5. Let $V$ be a discrete random variable with finite support. Let $F(\cdot)$ be the CDF of $V$, with $F_{-}(\cdot)$ being its left limit. Let $p(\cdot)$ be the corresponding probability mass function. Let $\zeta$ be an independent $U(0,1)$ random variable. Then $U = F_{-}(V) + \zeta p(V) \sim U(0,1)$.

Proof of Lemma 5. Let $v_{1} < ... < v_{s}$ be the support of $V$. Let $q_{0} = 0$, $q_{j} = p(v_{1}) + ... + p(v_{j})$ for $j = 1,...,s$. Then $q_{s} = 1$. For $1 \leq j \leq s$, it is direct to verify that density of $U$ on $(q_{j-1},q_{j})$ is a constant. Moreover, the length of this interval is $q_{j} - q_{j-1} = p(v_{j})$, which is the same as the probability of $U$ falling in $(q_{j-1},q_{j})$. So the density of $U$ on this interval is 1. Since this holds for each $j$, we conclude that $U \sim U(0,1)$.

The next two lemmas are useful in establishing separation between $H_{0}$ and $H_{1}$ when the correct $V$ function is used. Lemma 6 provides a “change-of-variable” trick to simplify an integral involved
Table 7: Percentage of rejections and probability weight estimation error in Model C under the null over 1000 repetitions.

| Percentage of rejections | $Err_{\hat{\beta}}$ |
|--------------------------|----------------------|
|                          | $m=5$ | $m=10$ | $m=50$ | $m=5$ | $m=10$ | $m=50$ | $m=5$ | $m=10$ | $m=50$ |
| $\hat{V}_{LL} + g$       |       |        |       |       |        |       |
| $n_2 = 200$              | 0.053 | 0.045  | 0.045 | 0.045 | 0.046  | 0.049  | 0.142 | 0.178  | 0.238  |
| $n_2 = 500$              | 0.055 | 0.059  | 0.053 | 0.053 | 0.046  | 0.050  | 0.093 | 0.118  | 0.159  |
| $n_2 = 1000$             | 0.057 | 0.043  | 0.050 | 0.051 | 0.048  | 0.057  | 0.059 | 0.075  | 0.101  |
| $n_2 = 5000$             | 0.060 | 0.058  | 0.053 | 0.043 | 0.058  | 0.056  | 0.025 | 0.032  | 0.043  |
| $\hat{V}_{QL} + g$       |       |        |       |       |        |       |
| $n_2 = 200$              | 0.045 | 0.056  | 0.043 | 0.035 | 0.049  | 0.054  | 0.205 | 0.261  | 0.361  |
| $n_2 = 500$              | 0.053 | 0.036  | 0.054 | 0.056 | 0.047  | 0.055  | 0.129 | 0.167  | 0.230  |
| $n_2 = 1000$             | 0.055 | 0.057  | 0.041 | 0.045 | 0.050  | 0.059  | 0.084 | 0.108  | 0.147  |
| $n_2 = 5000$             | 0.043 | 0.060  | 0.052 | 0.060 | 0.053  | 0.057  | 0.036 | 0.046  | 0.062  |
| $\hat{V}_{NN} + g$       |       |        |       |       |        |       |
| $n_2 = 200$              | 0.053 | 0.062  | 0.049 | 0.060 | 0.062  | 0.066  | 0.159 | 0.202  | 0.273  |
| $n_2 = 500$              | 0.036 | 0.052  | 0.066 | 0.053 | 0.055  | 0.062  | 0.126 | 0.161  | 0.221  |
| $n_2 = 1000$             | 0.049 | 0.060  | 0.046 | 0.043 | 0.066  | 0.059  | 0.089 | 0.116  | 0.168  |
| $n_2 = 5000$             | 0.053 | 0.052  | 0.043 | 0.057 | 0.058  | 0.049  | 0.064 | 0.085  | 0.122  |
| $\hat{V}_{RF} + g$       |       |        |       |       |        |       |
| $n_2 = 200$              | 0.060 | 0.047  | 0.045 | 0.128 | 0.169  | 0.253  | 0.443 | 0.542  | 0.656  |
| $n_2 = 500$              | 0.036 | 0.053  | 0.051 | 0.105 | 0.368  | 0.526  | 0.382 | 0.468  | 0.597  |
| $n_2 = 1000$             | 0.060 | 0.049  | 0.052 | 0.345 | 0.464  | 0.735  | 0.347 | 0.431  | 0.554  |
| $n_2 = 5000$             | 0.050 | 0.054  | 0.045 | 0.842 | 0.895  | 1.000  | 0.289 | 0.363  | 0.486  |
Table 8: Percentage of rejections and probability weight estimation error in Model C under the alternative over 1000 repetitions.

|                  | Percentage of rejections | $Err_{\hat{p}}$ |
|------------------|--------------------------|-----------------|
|                  | $m=5$  | $m=10$  | $m=50$  | $m=5$  | $m=10$  | $m=50$  |
| $V_{LL} + g$     |        |         |         |        |         |         |
| $n_2 = 200$      | 0.046  | 0.040   | 0.030   | 0.038  | 0.052   | 0.036   |
| $n_2 = 500$      | 0.056  | 0.036   | 0.032   | 0.048  | 0.046   | 0.028   |
| $n_2 = 1000$     | 0.028  | 0.040   | 0.020   | 0.032  | 0.052   | 0.014   |
| $n_2 = 5000$     | 0.057  | 0.029   | 0.000   | 0.114  | 0.000   | 0.000   |
| $V_{QL} + g$     |        |         |         |        |         |         |
| $n_2 = 200$      | 0.318  | 0.402   | 0.668   | 0.286  | 0.392   | 0.602   |
| $n_2 = 500$      | 0.602  | 0.838   | 0.992   | 0.644  | 0.824   | 0.984   |
| $n_2 = 1000$     | 0.900  | 0.990   | 1.000   | 0.890  | 0.994   | 1.000   |
| $n_2 = 5000$     | 1.000  | 1.000   | 1.000   | 1.000  | 1.000   | 1.000   |
| $V_{NN} + g$     |        |         |         |        |         |         |
| $n_2 = 200$      | 0.112  | 0.130   | 0.174   | 0.094  | 0.104   | 0.132   |
| $n_2 = 500$      | 0.580  | 0.740   | 0.924   | 0.530  | 0.708   | 0.900   |
| $n_2 = 1000$     | 0.856  | 0.972   | 0.998   | 0.834  | 0.962   | 0.998   |
| $n_2 = 5000$     | 1.000  | 1.000   | 1.000   | 1.000  | 1.000   | 1.000   |
| $V_{RF} + g$     |        |         |         |        |         |         |
| $n_2 = 200$      | 0.082  | 0.078   | 0.106   | 0.174  | 0.242   | 0.384   |
| $n_2 = 500$      | 0.106  | 0.170   | 0.270   | 0.348  | 0.520   | 0.792   |
| $n_2 = 1000$     | 0.166  | 0.296   | 0.542   | 0.596  | 0.782   | 0.980   |
| $n_2 = 5000$     | 0.714  | 0.943   | 1.000   | 1.000  | 1.000   | 1.000   |

35
Table 9: Percentage of rejections and test error in Model D under the null over 1000 repetitions.

| Null | Percentage of rejections |    |    |    |    |    |    |    |
|------|--------------------------|----|----|----|----|----|----|----|
|      |                          | $\hat{V}_{LL} + g$ | $\hat{V}_{LL} + \hat{g}_{LL}$ | $\hat{E}_{V_{,LL}}$ | $\hat{E}_{V_{,QL}}$ | $\hat{E}_{V_{,NN}}$ | $\hat{E}_{V_{,RF}}$ |
| $n_2 = 200$ | 0.038 | 0.066 | 0.040 | 0.064 | 0.064 | 0.044 | 0.756 | 0.789 | 0.802 |
| $n_2 = 500$ | 0.042 | 0.064 | 0.038 | 0.046 | 0.044 | 0.048 | 0.751 | 0.784 | 0.795 |
| $n_2 = 1000$ | 0.038 | 0.044 | 0.056 | 0.056 | 0.052 | 0.040 | 0.750 | 0.783 | 0.794 |
| $n_2 = 5000$ | 0.060 | 0.043 | 0.053 | 0.057 | 0.043 | 0.053 | 0.749 | 0.780 | 0.792 |
| $n_2 = 200$ | 0.022 | 0.046 | 0.048 | 0.052 | 0.058 | 0.052 | 0.266 | 0.313 | 0.401 |
| $n_2 = 500$ | 0.046 | 0.064 | 0.052 | 0.060 | 0.074 | 0.046 | 0.162 | 0.191 | 0.236 |
| $n_2 = 1000$ | 0.042 | 0.050 | 0.038 | 0.058 | 0.032 | 0.038 | 0.111 | 0.131 | 0.162 |
| $n_2 = 5000$ | 0.032 | 0.035 | 0.057 | 0.050 | 0.028 | 0.050 | 0.051 | 0.060 | 0.073 |
| $n_2 = 200$ | 0.044 | 0.042 | 0.042 | 0.176 | 0.232 | 0.260 | 0.555 | 0.622 | 0.688 |
| $n_2 = 500$ | 0.028 | 0.042 | 0.054 | 0.028 | 0.044 | 0.040 | 0.383 | 0.450 | 0.534 |
| $n_2 = 1000$ | 0.058 | 0.048 | 0.042 | 0.046 | 0.046 | 0.060 | 0.274 | 0.323 | 0.382 |
| $n_2 = 5000$ | 0.043 | 0.021 | 0.035 | 0.074 | 0.046 | 0.057 | 0.124 | 0.151 | 0.186 |
| $n_2 = 200$ | 0.040 | 0.058 | 0.050 | 0.248 | 0.314 | 0.334 | 0.574 | 0.639 | 0.687 |
| $n_2 = 500$ | 0.044 | 0.038 | 0.046 | 0.364 | 0.438 | 0.592 | 0.519 | 0.582 | 0.634 |
| $n_2 = 1000$ | 0.044 | 0.046 | 0.054 | 0.516 | 0.662 | 0.744 | 0.476 | 0.541 | 0.593 |
| $n_2 = 5000$ | 0.060 | 0.064 | 0.046 | 0.950 | 0.989 | 1.000 | 0.395 | 0.456 | 0.512 |
Table 10: Percentage of rejections and test error in Model D under the alternative over 1000 repetitions.

| Alternative | Percentage of rejections | $E_{\text{Err}_{\hat{p}}}$ |
|-------------|--------------------------|-----------------------------|
|             | $\hat{V}_{LL} + g$      | $\hat{V}_{LL} + \hat{g}_{LL}$ | $E_{\text{Err}_{\hat{p},LL}}$ |
| $n_2 = 200$ | 0.038 0.062 0.102 | 0.064 0.070 0.096 | 0.755 0.788 0.800 |
| $n_2 = 500$ | 0.054 0.044 0.088 | 0.070 0.068 0.100 | 0.750 0.785 0.797 |
| $n_2 = 1000$ | 0.040 0.072 0.083 | 0.051 0.076 0.083 | 0.752 0.782 0.793 |
| $n_2 = 5000$ | 0.053 0.058 0.077 | 0.066 0.063 0.074 | 0.749 0.780 0.792 |

|             | $\hat{V}_{QL} + g$      | $\hat{V}_{QL} + \hat{g}_{QL}$ | $E_{\text{Err}_{\hat{p},QL}}$ |
| $n_2 = 200$ | 0.238 0.298 0.344 | 0.242 0.276 0.282 | 0.265 0.317 0.401 |
| $n_2 = 500$ | 0.500 0.620 0.688 | 0.500 0.594 0.658 | 0.160 0.188 0.234 |
| $n_2 = 1000$ | 0.799 0.888 0.926 | 0.799 0.879 0.930 | 0.111 0.132 0.162 |
| $n_2 = 5000$ | 1.000 1.000 1.000 | 1.000 1.000 1.000 | 0.049 0.058 0.071 |

|             | $\hat{V}_{NN} + g$      | $\hat{V}_{NN} + \hat{g}_{NN}$ | $E_{\text{Err}_{\hat{p},NN}}$ |
| $n_2 = 200$ | 0.056 0.098 0.090 | 0.214 0.318 0.384 | 0.551 0.619 0.676 |
| $n_2 = 500$ | 0.230 0.302 0.308 | 0.440 0.560 0.634 | 0.382 0.450 0.536 |
| $n_2 = 1000$ | 0.754 0.786 0.875 | 0.911 0.949 0.989 | 0.300 0.354 0.428 |
| $n_2 = 5000$ | 1.000 1.000 1.000 | 1.000 1.000 1.000 | 0.124 0.150 0.186 |

|             | $\hat{V}_{RF} + g$      | $\hat{V}_{RF} + \hat{g}_{RF}$ | $E_{\text{Err}_{\hat{p},RF}}$ |
| $n_2 = 200$ | 0.064 0.056 0.078 | 0.290 0.376 0.462 | 0.573 0.641 0.685 |
| $n_2 = 500$ | 0.140 0.112 0.108 | 0.528 0.620 0.696 | 0.515 0.581 0.632 |
| $n_2 = 1000$ | 0.121 0.180 0.180 | 0.725 0.833 0.911 | 0.477 0.538 0.592 |
| $n_2 = 5000$ | 0.470 0.562 0.578 | 1.000 1.000 1.000 | 0.394 0.454 0.512 |
in calculating $\mathbb{E}U$, the expected value of the conformal $p$-value given in (7). Lemma 7 will be used to prove that $\mathbb{E}U$ tends to be smaller when the testing sample point is generated from $P_2$ as compared to $P_1$.

**Lemma 6.** Let $(X_j, Y_j) \sim P_j$ be independent for $j = 1, 2$, and $(X'_2, Y'_2)$ be another independent draw from $P_2$. Let $V(x, y) = \frac{f_1(y|x)}{f_2(y|x)}$ and $g(x) = f_2(X(x)/f_1(x)$. Define $V_1 = V(X_1, Y_1)$, $V_2 = V(X_2, Y_2)$, and $V'_2 = V(X'_2, Y'_2)$. Let $\hat{V}: \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ be an arbitrary non-random function and define $\hat{V}_1, \hat{V}_2, \hat{V}'_2$ similarly as $V_1, V_2, V'_2$ using $\hat{V}$. We have

$$\mathbb{E}g(X_1) \mathbb{I}(\hat{V}_1 < \hat{V}_2) = \mathbb{E}V'_2 \mathbb{I}(\hat{V}'_2 < \hat{V}_2)$$

and

$$\mathbb{E}g(X_1) \mathbb{I}(\hat{V}_1 \leq \hat{V}_2) = \mathbb{E}V'_2 \mathbb{I}(\hat{V}'_2 \leq \hat{V}_2).$$

**Proof.** By definition,

$$\mathbb{E}g(X_1) \mathbb{I}(\hat{V}_1 < \hat{V}_2) = \mathbb{E} \left\{ f_2(X_1, Y_1) f_1(Y_1|x_1) f_2(Y_1|x_2) \mathbb{I}(\hat{V}_1 < \hat{V}_2) \right\}$$

$$= \int f_2(x_2, y_2) \int_{\hat{V}(x_1, y_1) < \hat{V}(x_2, y_2)} f_2(x_1, y_1) V(x_1, y_1) dx_1 dy_1 dx_2 dy_2$$

$$= \mathbb{E} \{ V'_2 \mathbb{I}(\hat{V}'_2 < \hat{V}_2) \}.$$

The proof of the second equation is identical.

**Lemma 7.** If $V$ is a non-degenerate nonnegative random variable with $\mathbb{E}V = 1$. Let $V'$ be an iid copy of $V$. Then $\mathbb{E}\{ V \mathbb{I}(V < V') \} + \mathbb{E}\{ V \mathbb{I}(V \leq V') \} < 1$.

**Proof.** By non-degeneracy of $V$, $P(V < V') > 0$, and hence $\mathbb{E}\{ V \mathbb{I}(V < V') \} < \mathbb{E}\{ V' \mathbb{I}(V < V') \}$, and $\mathbb{E}\{ V \mathbb{I}(V \leq V') \} < \mathbb{E}\{ V' \mathbb{I}(V \leq V') \}$. The claims follows by observing that

$$\mathbb{E}\{ V \mathbb{I}(V < V') \} + \mathbb{E}\{ V' \mathbb{I}(V < V') \} + \mathbb{E}\{ V \mathbb{I}(V \leq V') \} + \mathbb{E}\{ V' \mathbb{I}(V \leq V') \}$$

$$= \mathbb{E}\{ V \mathbb{I}(V < V') \} + \mathbb{E}\{ V \mathbb{I}(V > V') \} + \mathbb{E}\{ V \mathbb{I}(V \leq V') \} + \mathbb{E}\{ V \mathbb{I}(V \geq V') \}$$

$$= 2.$$
Lemma 8. Let \((X_{1l}, Y_{1l})\) iid \(P_1\), \((X_2, Y_2) \sim P_2\) be independent. Let \(\hat{V}, \hat{g}\) be non-random approximate versions of \(V, g\). Define \(g_{1l} = g(X_{1l})\) for \(l = 1, \ldots, m\) and \(g_2 = g(X_2)\). Define \(\hat{g}_{1l}, \hat{g}_2, V_{1l}, V_2, \hat{V}_{1l}, \hat{V}_2\) correspondingly. Let \(p_{1l} = \sum_{l=1}^m g_{1l}/g_{1l} + g_2\), \(p_2 = \sum_{l=1}^m g_2/g_{1l} + g_2\), and define \(\hat{p}_{1l}, \hat{p}_2\) similarly using \(\hat{g}\). Let \(\zeta \sim U(0, 1)\) be independent of everything else. The following holds.

(a) Under Assumption 1, we have, for all values of \(m\),

\[
\mathbb{E} \left( \sum_{l=1}^m \left| \hat{p}_{1l} - p_{1l} \right| + \left| \hat{p}_2 - p_2 \right| \right) \leq C \max \left\{ \|\hat{g} - g\|_2, \|\hat{g} - g\|_2, p_2 \right\},
\]

(b)

\[
\left| \mathbb{E} g_{1l} \left( \mathbb{I} (\hat{V}_{1l} < \hat{V}_2) + \mathbb{I} (\hat{V}_{1l} \leq \hat{V}_2) \right) \right| - \mathbb{E} g_{1l} \left( \mathbb{I} (V_{1l} < V_2) + \mathbb{I} (V_{1l} \leq V_2) \right) \leq 4 \|\hat{V} - V\|_{1, p_2}
\]

Proof of Lemma 8. For part (a) we use shorthand notation \(g_l = g(X_{1l})\) for \(1 \leq l \leq m\), \(g_{m+1} = g(X_2)\), and \(\hat{g}_l\) correspondingly.

\[
\sum_{l=1}^{m+1} \left| \hat{p}_l - p_l \right| = \sum_{l=1}^{m+1} \left| \frac{g_l}{\sum_{l=1}^{m+1} g_l} \right| \leq \frac{\sum_{l=1}^{m+1} |\hat{g}_l - g_l|}{\sum_{l=1}^{m+1} g_l} \leq 2 \frac{\sum_{l=1}^{m+1} \left| \hat{g}_l - g_l \right|}{\sum_{l=1}^{m+1} g_l} \leq 2 \left( \frac{1}{m+1} \sum_{l=1}^{m+1} |\hat{g}_l - g_l| \right) \left( \frac{1}{m+1} \sum_{l=1}^{m+1} \frac{1}{g_l} \right)
\]

where the last step follows from Jensen's inequality. Using Cauchy-Schwarz inequality,

\[
\mathbb{E} \sum_{l=1}^{m+1} |\hat{p}_l - p_l| \leq 2 \mathbb{E} \left( \frac{1}{m+1} \sum_{l=1}^{m+1} |\hat{g}_l - g_l| \right) \left( \frac{1}{m+1} \sum_{l=1}^{m+1} \frac{1}{g_l} \right) \leq 2 \left[ \mathbb{E} \left( \frac{1}{m+1} \sum_{l=1}^{m+1} |\hat{g}_l - g_l|^2 \right) \right]^{1/2} \left[ \mathbb{E} \left( \frac{1}{m+1} \sum_{l=1}^{m+1} \frac{1}{g_l} \right) \right]^{1/2} \leq C \max \{ \|\hat{g} - g\|_2, p_1, \|\hat{g} - g\|_2, p_2 \},
\]

where \(C\) is some positive constant such that \(\mathbb{E} \frac{1}{g^2(X)} \leq (C/2)^2\) for both \(X \sim F_{1,X}\) and \(X \sim F_{2,X}\).
For part (b), first applying Lemma 6 we have

\[ E_g(X_{1t}) \mathbb{I}(\hat{V}_{1t} < \hat{V}_2) = E V_2 \mathbb{I}(\hat{V}_2 < \hat{V}_2'), \]
\[ E_g(X_{1t}) \mathbb{I}(\hat{V}_{1t} \leq \hat{V}_2) = E V_2 \mathbb{I}(\hat{V}_2 \leq \hat{V}_2'), \]

where \( \hat{V}_2' = \hat{V}(X_2', Y_2') \) with \((X_2', Y_2')\) an independent copy of \((X_2, Y_2)\).

Let \( \xi = V_2 - \hat{V}_2 + \hat{V}_2' - V_2' \). Then

\[ EV_2 \mathbb{I}(\hat{V}_2 < \hat{V}_2') \]
\[ = EV_2 \mathbb{I}(V_2 < V_2') + EV_2 \mathbb{I}(V_2' \leq V_2 < V_2' + \xi, \xi > 0) - EV_2 \mathbb{I}(V_2' + \xi \leq V_2 < V_2', \xi < 0) \]
\[ = EV_2 \mathbb{I}(V_2 < V_2') + EV_2 \mathbb{I}(V_2' \leq V_2 < V_2' + \xi, \xi > 0) - EV_2 \mathbb{I}(V_2' < V_2 \leq V_2' + \xi, \xi > 0) \]

the last step holds by swapping the roles of \((X_2, Y_2)\) and \((X_2', Y_2')\) in the expectation.

Similarly

\[ EV_2 \mathbb{I}(\hat{V}_2 \leq \hat{V}_2') \]
\[ = EV_2 \mathbb{I}(V_2 \leq V_2') + EV_2 \mathbb{I}(V_2' < V_2 \leq V_2' + \xi, \xi > 0) - EV_2 \mathbb{I}(V_2' + \xi \leq V_2 < V_2', \xi < 0) \]
\[ = EV_2 \mathbb{I}(V_2 \leq V_2') + EV_2 \mathbb{I}(V_2' < V_2 \leq V_2' + \xi, \xi > 0) - EV_2 \mathbb{I}(V_2' \leq V_2 < V_2' + \xi, \xi > 0) \]

Adding these two equations, we obtain

\[ \left| EV_2 \mathbb{I}(\hat{V}_2 < \hat{V}_2') + EV_2 \mathbb{I}(\hat{V}_2 \leq \hat{V}_2') \right| - \left| EV_2 \mathbb{I}(V_2 < V_2') + EV_2 \mathbb{I}(V_2 \leq V_2') \right| \]
\[ = \left| E(V_2 - V_2') \mathbb{I}(V_2' \leq V_2 < V_2' + \xi, \xi > 0) + E(V_2 - V_2') \mathbb{I}(V_2' < V_2 \leq V_2' + \xi, \xi > 0) \right| \]
\[ \leq 2E|V_2 - V_2'| \mathbb{I}(|V_2 - V_2'| \leq |\xi|) \]
\[ \leq 2E|\xi| \leq 4\|
\hat{V} - V\|_{1,p_2}. \]

C. PROOFS OF MAIN RESULTS

Proof of Lemma 2. For part (a), the proof is a standard application of the Bayes theorem, and is implicitly given in (Tibshirani et al., 2019, equation (15) and Section 3.4). Here we provide the
calculation for the readers’ convenience.

Recall that given \( \tilde{Z} \), we have \( V_{m+1} = V(\tilde{X}_1, \tilde{Y}_i) \) if \( \sigma(m + 1) = i \). Thus

\[
(V_{m+1|\tilde{Z}}) \sim \sum_{\sigma} P(\sigma|\tilde{Z})\delta_{V(\tilde{X}_{\sigma(m+1)}, \tilde{Y}_{\sigma(m+1)})} = \sum_{i=1}^{m+1} P[\sigma(m + 1) = i|\tilde{Z}] \delta_{V(\tilde{X}_i, \tilde{Y}_i)}. \tag{A.1}
\]

Using Bayes rule,

\[
P[\sigma(m + 1) = i|\tilde{Z}] = \frac{\sum_{\sigma(m+1)=i} \Pi_{l=1}^{m+1} f_1(x_l) f(Y_l|X_l) f_{l,2,1}(X_l) f_{1,2}(X_l)}{\sum_{j=1}^{m+1} \sum_{\sigma(m+1)=j} \Pi_{l=1}^{m+1} f_1(x_l) f(Y_l|X_l) f_{l,2,1}(X_l) f_{1,2}(X_l)} = p_i(\tilde{Z}).
\]

As a result, (A.1) becomes

\[
(V_{m+1|\tilde{Z}}) \sim \sum_{i=1}^{m+1} p_i(\tilde{Z})\delta_{V(\tilde{X}_i, \tilde{Y}_i)} = \sum_{i=1}^{m+1} p_i(\tilde{Z})\delta_{V(X_i,Y_i)}.
\]

Part (b) follows directly by combining part (a) with Lemma 5.

For part (c), we use notation \( V_{lI} = V(X_l, Y_l) = f_1(Y_l|X_l)/f_2(Y_l|X_l), \ l = 1, \ldots, m, \) and \( V_2 = V(X_{m+1}, Y_{m+1}) = f_1(Y_{m+1}|X_{m+1})/f_2(Y_{m+1}|X_{m+1}). \) Recall that \( g(x) = f_{2,1}(x)/f_{1,1}(x). \)

For a given \( X_2 \), we consider

\[
2\mathbb{E}_\zeta U = \frac{\sum_{l=1}^{m} g(X_l)[\mathbb{I}(V_l < V_2) + \mathbb{I}(V_l \leq V_2)] + g(X_{m+1})}{\sum_{l=1}^{m} g(X_l) + g(X_{m+1})},
\]

where \( \mathbb{E}_\zeta \) denotes expectation taken only over \( \zeta \), keeping everything else as given. Because \( \mathbb{E}_g(X_l) = 1 \) for \( 1 \leq l \leq m \), it can be directly verified from the strong law of large numbers on the iid random variables \( X_1, \ldots, X_m \) that

\[
2\mathbb{E}_\zeta U \to \mathbb{E} \left[ g(X_1)\mathbb{I}(V_1 < V_2) X_{m+1}, Y_{m+1} \right] + \mathbb{E} \left[ g(X_1)\mathbb{I}(V_1 \leq V_2) X_{m+1}, Y_{m+1} \right]
\]
a.e. over \((X_1, \ldots, X_m, X_{m+1})\) as \( m \to \infty \). By construction we have \( U \in [0,2] \). By the dominated convergence theorem, we have

\[
2\mathbb{E}(U) \to \mathbb{E}g(X_1)\mathbb{I}(V_1 < V_2) + \mathbb{E}g(X_1)\mathbb{I}(V_1 \leq V_2). \tag{A.2}
\]

Lemma 6 implies that the right hand side of (A.2) is

\[
\mathbb{E}V_2'\mathbb{I}(V_2' < V_2) + \mathbb{E}V_2'\mathbb{I}(V_2' \leq V_2),
\]

41
where $V'_2$ is an iid copy of $V_2$.

By construction of $V$ we have $\mathbb{E}V_2 = \mathbb{E}V'_2 = 1$. Under $H_1$, $V(X_2, Y_2) = f_1(Y_2|X_2)/f_2(Y_2|X_2)$ cannot be degenerate. Therefore Lemma 7 implies that

$$\mathbb{E}\{V'_2 \mathbb{1}(V'_2 < V_2)\} + \mathbb{E}\{V'_2 \mathbb{1}(V'_2 \leq V_2)\} < 1,$$

which completes the proof.

**Proof of Theorem 3.** We first construct a coupling random variable $U'_k$ from $\hat{U}_k$. Recall that

$$\hat{U}_k = \sum_{l=1}^{m} \hat{p}_{k,l} \mathbb{1}(\hat{V}_{k,l} < \hat{V}_{k,m+1}) + \zeta_k \left\{ \hat{p}_{k,m+1} + \sum_{l=1}^{m} \hat{p}_{k,l} \mathbb{1}(\hat{V}_{k,l} = \hat{V}_{k,m+1}) \right\}$$

where $\zeta_k$ is a $U(0,1)$ random variable independent of everything else. The coupling random variable is

$$U'_k = \sum_{l=1}^{m} p_{k,l} \mathbb{1}(\hat{V}_{k,l} < \hat{V}_{k,m+1}) + \zeta_k \left\{ p_{k,m+1} + \sum_{l=1}^{m} p_{k,l} \mathbb{1}(\hat{V}_{k,l} = \hat{V}_{k,m+1}) \right\}. \tag{A.3}$$

By Lemma 2 we have $(U'_k | Z_k, D_{fit}) \sim U(0,1)$, where $Z_k = \{(X_{1i_{k,l}}, Y_{1i_{k,l}}) : 1 \leq l \leq m\} \cup \{(X_{2k}, Y_{2k})\}$ is the $k$th ranking batch whose ordering is ignored. Hence $(U'_k | D_{fit}) \overset{iid}{\sim} U(0,1)$.

It suffices to show that

$$\sqrt{K} \left| \frac{1}{K} \sum_{k=1}^{K} (\hat{U}_k - U'_k) \right| = o_P(1) \tag{A.4}$$

By construction, $|\hat{U}_k - U'_k| \leq \sum_{l=1}^{m+1} |\hat{p}_{k,l} - p_{k,l}|$. Then Lemma 8(a) implies that

$$\mathbb{E} \left\{ |\hat{U}_k - U'_k| | D_{fit} \right\} \leq C \max \{ \|\hat{g} - g\|_{2,p_1}, \|\hat{g} - g\|_{2,p_2} \}.$$ 

Then (A.4) follows from Assumption 2(b).

**Proof of Theorem 4.** Here we drop the subindex $k$ since we focus on only one ranking batch $(X_{1i_{k,l}}, Y_{1i_{k,l}})_{l=1}^{m}, (X_{2k}, Y_{2k})$, which is now written as $(X_{1l}, Y_{1l})_{l=1}^{m}, (X_{2}, Y_{2})$.

In the rest of the proof we use the notation $V_{1l} = V(X_{1l}, Y_{1l})$ for $1 \leq l \leq m$ and $V_2 = V(X_2, Y_2)$, and $\hat{V}_{1l}, \hat{g}_{1l}, \hat{\hat{g}}_{1l}, \hat{V}_{2}, \hat{g}_{2}, V_2$ are defined similarly. The construction of the conformal predictions statistic $\hat{U}$ can be written as

$$\hat{U} = \sum_{l=1}^{m} \hat{p}_{1l} \mathbb{1}(\hat{V}_{1l} < \hat{V}_2) + \zeta \left( \sum_{l=1}^{m} \hat{p}_{1l} \mathbb{1}(\hat{V}_{1l} = \hat{V}_2) + \hat{p}_2 \right),$$

where $\hat{p}_{1l} = \hat{g}_{1l}/(\sum_{l=1}^{m} \hat{g}_{1l} + \hat{g}_2)$, $l = 1, \ldots, m$, and $\hat{p}_2 = \hat{g}_2/(\sum_{l=1}^{m} \hat{g}_{1l} + \hat{g}_2)$. 

42
Again we consider the coupling random variable \( U' \) defined in (A.3). By Lemma 8(a) we have
\[
\mathbb{E}(\hat{U}|D_{\text{fit}}) \leq \mathbb{E}(U'|D_{\text{fit}}) + Cr_n, \tag{A.5}
\]
where \( r_n \) is defined as in Assumption 2(b). Because \( r_n\sqrt{K} = o_P(1) \) by Assumption 2(b), it suffices to show that \( \mathbb{E}(U'|D_{\text{fit}}) \leq 1/2 - \delta \) for some \( \delta > 0 \) uniformly as \( n \to \infty \) and \( m \geq m_0 \).

To this end, we consider the following ideal version
\[
U = \sum_{l=1}^{m} p_{1l}\mathbb{1}(V_{1l} < V_2) + \zeta \left( \sum_{l=1}^{m} p_{1l}\mathbb{1}(V_{1l} = V_2) + p_2 \right)
\]
where the \( p \)'s are defined analogously as \( \hat{p} \)'s with \( \hat{g} \) replaced by \( g \).

Lemma 2(c) guarantees that \( \mathbb{E}U \leq 1/2 - \delta \) for some \( \delta > 0 \) when \( m \geq m_0 \). Our proof reduces to controlling \( \mathbb{E}[(U' - U)|D_{\text{fit}}] \).

\[
2\mathbb{E}_\zeta(U' - U) = \frac{1}{m} \sum_{l=1}^{m} g_{1l} \left[ \mathbb{1}(\hat{V}_{1l} < \hat{V}_2) + \mathbb{1}(\hat{V}_{1l} = \hat{V}_2) - \{ \mathbb{1}(V_{1l} < V_2) + \mathbb{1}(V_{1l} \leq V_2) \} \right]
= \frac{A}{B}
\]

According to Lemma 8(b), conditioning on \( D_{\text{fit}} \), we have \( \mathbb{E}A \leq 4\|\hat{V} - V\|_{1,P_2} \). Moreover, \( \text{Var}(A) \leq m^{-1}\mathbb{E}g_{1l}^2 \leq C/m \). So by Chebyshev’s inequality
\[
P(|A| \geq 4\|\hat{V} - V\|_{1,P_2} + m^{-1/3}|D_{\text{fit}}) \leq Cm^{-1/3}.
\]

For the denominator \( B \) we have \( \mathbb{E}B \geq 1 \) and \( \text{Var}(B) \leq C/m \). So again by Chebyshev’s inequality
\[
P(B < 1/2) \leq \frac{C}{m}.
\]

Putting together using union bound,
\[
P \left( |U' - U| \geq 8\|\hat{V} - V\| + 2m^{-1/3}|D_{\text{fit}} \right) \leq Cm^{-1/3}.
\]

Because \( |U' - U| \leq 1 \) we then conclude that
\[
\mathbb{E}(|U' - U| |D_{\text{fit}}) \leq C(\|\hat{V} - V\|_{1,P_2} + m^{-1/3}). \tag{A.6}
\]

Combining (A.5) and (A.6) we have when \( m \) is large enough
\[
\mathbb{E}(\hat{U}|D_{\text{fit}}) \leq 1/2 - \delta/2 + C(\|\hat{V} - V\|_{1,P_2} + r_n),
\]

The claimed result follows since \( \|\hat{V} - V\|_{1,P_2} \) and \( r_n \) are both \( o_P(1) \) over the probability of \( D_{\text{fit}} \).