Conformal Prediction Under Covariate Shift

Rina Foygel Barber Emmanuel J. Candès Aaditya Ramdas Ryan J. Tibshirani

Abstract

We extend conformal prediction methodology beyond the case of exchangeable data. In particular, we show that a weighted version of conformal prediction can be used to compute distribution-free prediction intervals for problems in which the test and training covariate distributions differ, but the likelihood ratio between these two distributions is known—or, in practice, can be estimated accurately with access to a large set of unlabeled data (test covariate points). Our weighted extension of conformal prediction also applies more generally, to settings in which the data satisfies a certain weighted notion of exchangeability. We discuss other potential applications of our new conformal methodology, including latent variable and missing data problems.

1 Introduction

Let $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \dots, n$ denote training data that is assumed to be i.i.d. from an arbitrary distribution P. Given a desired coverage rate $1 - \alpha \in (0, 1)$, consider the problem of constructing a band $\widehat{C}_n : \mathbb{R}^d \to \{\text{subsets of } \mathbb{R}\}$, based on the training data such that, for a new i.i.d. point (X_{n+1}, Y_{n+1}),

$$\mathbb{P}\Big\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\Big\} \geq 1 - \alpha, \tag{1}$$

where this probability is taken over the $n + 1$ points (X_i, Y_i), $i = 1, \dots, n + 1$ (the n training points and the test point). Crucially, we will require (1) to hold with no assumptions whatsoever on the common distribution P.

Conformal prediction, a framework pioneered by Vladimir Vovk and colleagues in the 1990s, provides a means for achieving this goal, relying only on exchangeablility of the training and test data. The definitive reference is the book by Vovk et al. (2005); see also Shafer and Vovk (2008); Vovk et al. (2009); Vovk (2013); Burnaev and Vovk (2014), and `http://www.alrw.net` for an often-updated list of conformal prediction work by Vovk and colleagues. Moreover, we refer to Lei and Wasserman (2014); Lei et al. (2018) for recent developments in the areas of nonparametric and high-dimensional regression. In this work, we extend conformal prediction beyond the setting of exchangeable data, allowing for provably valid inference even when the training and test data are not drawn from the same distribution. We begin by reviewing the basics of conformal prediction, in this section. In Section 2, we describe an extension of conformal prediction to the setting of covariate shift, and give supporting empirical results. In Section 3, we cover the mathematical details behind our conformal extension. We conclude in Section 4, and discuss several other possible applications of our new conformal methodology.

1.1 Quantile lemma

Before explaining the basic ideas behind conformal inference (i.e., conformal prediction, we will use these two terms interchangeably), we introduce some notation. We denote by $\text{Quantile}(\beta; F)$ the level β quantile of a distribution F, i.e., for $Z \sim F$,

$$\text{Quantile}(\beta; F) = \inf\big\{z : \mathbb{P}\{Z \leq z\} \geq \beta\big\}.$$

In our use of quantiles, we will allow for distributions F on the augmented real line, $\mathbb{R} \cup \{\infty\}$. For values $v_1, \dots, v_n$, we write $v_{1:n} = \{v_1, \dots, v_n\}$ to denote their multiset. Note that this is unordered, and allows for multiple instances the same element; thus in the present case, if $v_i = v_j$ for $i \neq j$, then this value appears twice in $v_{1:n}$. To denote quantiles of the empirical distribution of the values $v_1, \dots, v_n$, we abbreviate

$$\text{Quantile}(\beta; v_{1:n}) = \text{Quantile}\bigg(\beta; \frac{1}{n}\sum_{i=1}^n \delta_{v_i}\bigg),$$

where δ_a denotes a point mass at a (i.e., the distribution that places all mass at the value a). The next result is a simple but key component underlying conformal prediction.

Lemma 1. *If $V_1, \dots, V_{n+1}$ are exchangeable random variables, then for any $\beta \in (0,1)$, we have*

$$\mathbb{P}\Big\{V_{n+1} \leq \mathrm{Quantile}\big(\beta; V_{1:n} \cup \{\infty\}\big)\Big\} \geq \beta.$$

Furthermore, if ties between $V_1, \dots, V_{n+1}$ occur with probability zero, then the above probability is upper bound by $\beta + 1/(n+1)$.

Proof. Consider the following useful fact about quantiles of a discrete distribution F, with support points $a_1, \dots, a_k \in \mathbb{R}$: denoting $q = \mathrm{Quantile}(\beta; F)$, if we reassign the points $a_i > q$ to arbitrary values strictly larger than q, yielding a new distribution $\widetilde{F}$, then the level β quantile remains unchanged, $\mathrm{Quantile}(\beta; F) = \mathrm{Quantile}(\beta; \widetilde{F})$. Using this fact,

$$V_{n+1} > \mathrm{Quantile}\big(\beta; V_{1:n} \cup \{\infty\}\big) \iff V_{n+1} > \mathrm{Quantile}\big(\beta; V_{1:(n+1)}\big),$$

or equivalently,

$$V_{n+1} \leq \mathrm{Quantile}\big(\beta; V_{1:n} \cup \{\infty\}\big) \iff V_{n+1} \leq \mathrm{Quantile}\big(\beta; V_{1:(n+1)}\big), \tag{2}$$

Moreover, it is straightforward to check that

$$V_{n+1} \leq \mathrm{Quantile}\big(\beta; V_{1:(n+1)}\big) \iff V_{n+1} \text{ is among the } \lceil \beta(n+1) \rceil \text{ smallest of } V_1, \dots, V_{n+1}.$$

By exchangeability, the latter event occurs with probability at least $\lceil \beta(n+1) \rceil/(n+1) \geq \beta$, which proves the lower bound; when there are almost surely no ties, it holds with probability exactly $\lceil \beta(n+1) \rceil/(n+1) \leq \beta + 1/(n+1)$, which proves the upper bound. □

1.2 Conformal prediction

We now return to the regression setting. Denote $Z_i = (X_i, Y_i)$, $i = 1, \dots, n$, and $Z_{1:n} = \{Z_1, \dots, Z_n\}$. We will use the abbreviation $Z_{-i} = Z_{1:n} \setminus \{Z_i\}$. In what follows, we describe the construction of a prediction band satisfying (1), using conformal inference, due to Vovk et al. (2005). We first choose a score function $\mathcal{S}$, whose arguments consist of a point (x, y), and a multiset Z.[1] Informally, a low value of $\mathcal{S}((x, y), Z)$ indicates that the point (x, y) "conforms" to Z, whereas a high value indicates that (x, y) is atypical relative to the points in Z. For example, we might choose to define $\mathcal{S}$ by

$$\mathcal{S}\big((x, y), Z\big) = |y - \widehat{\mu}(x)|, \tag{3}$$

where $\widehat{\mu} : \mathbb{R}^d \to \mathbb{R}$ is a regression function that was fitted by running an algorithm $\mathcal{A}$ on (x, y) and Z.

Next, at each $x \in \mathbb{R}^d$, we define the conformal prediction interval[2] $\widehat{C}_n(x)$ by repeating the following procedure for each $y \in \mathbb{R}$. We calculate the *nonconformity scores*

$$V_i^{(x,y)} = \mathcal{S}\big(Z_i, Z_{-i} \cup \{(x, y)\}\big),\ i = 1, \dots, n, \quad \text{and} \quad V_{n+1}^{(x,y)} = \mathcal{S}\big((x, y), Z_{1:n}\big), \tag{4}$$

and include y in our prediction interval $\widehat{C}_n(x)$ if

$$V_{n+1}^{(x,y)} \leq \mathrm{Quantile}\big(1 - \alpha; V_{1:n}^{(x,y)} \cup \{\infty\}\big),$$

where $V_{1:n}^{(x,y)} = \{V_1^{(x,y)}, \dots, V_n^{(x,y)}\}$. Importantly, the symmetry in the construction of the nonconformity scores (4) guarantees exact coverage in finite samples. The next theorem summarizes this coverage result. The lower bound is a standard result in conformal inference, due to Vovk et al. (2005); the upper bound, as far as we know, was first pointed out by Lei et al. (2018).

[1] We emphasize that by defining Z to be a multiset, we are treating its points as unordered. Hence, to be perfectly explicit, the score function $\mathcal{S}$ cannot accept the points in Z in any particular order, and it must take them in as unordered. The same is true of the base algorithm $\mathcal{A}$ used to define the fitted regression function $\widehat{\mu}$, in the choice of absolute residual score function (3).

[2] For convenience, throughout, we will refer to $\widehat{C}_n(x)$ as an "interval", even though this may actually be a union of multiple nonoverlapping intervals. Similarly, for simplicity, we will refer to $\widehat{C}_n$ as a "band".

Theorem 1 (Vovk et al. 2005; Lei et al. 2018)**.** *Assume that* $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$*,* $i = 1, \ldots, n+1$ *are exchangeable. For any score function* $\mathcal{S}$*, and any* $\alpha \in (0,1)$*, define the conformal band (based on the first* n *samples) at* $x \in \mathbb{R}^d$ *by*

$$\widehat{C}_n(x) = \Big\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq \mathrm{Quantile}\big(1-\alpha; V_{1:n}^{(x,y)} \cup \{\infty\}\big) \Big\}, \tag{5}$$

where $V_i^{(x,y)}$*,* $i = 1, \ldots, n+1$ *are as defined in* (4)*. Then* $\widehat{C}_n$ *satisfies*

$$\mathbb{P}\Big\{ Y_{n+1} \in \widehat{C}_n(X_{n+1}) \Big\} \geq 1 - \alpha.$$

Furthermore, if ties between $V_1^{(X_{n+1},Y_{n+1})}, \ldots, V_{n+1}^{(X_{n+1},Y_{n+1})}$ *occur with probability zero, then this probability is upper bounded by* $1 - \alpha + 1/(n+1)$*.*

Proof. To lighten notation, abbreviate $V_i = V_i^{(X_{n+1},Y_{n+1})}$, $i = 1, \ldots, n+1$. Observe

$$Y_{n+1} \in \widehat{C}_n(X_{n+1}) \iff V_{n+1} \leq \mathrm{Quantile}\big(1-\alpha; V_{1:n} \cup \{\infty\}\big).$$

By the symmetric construction of the nonconformity scores in (4),

$$(Z_1, \ldots, Z_{n+1}) \overset{d}{=} (Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) \iff (V_1, \ldots, V_{n+1}) \overset{d}{=} (V_{\sigma(1)}, \ldots, V_{\sigma(n+1)}),$$

for any permutation σ of the numbers $1, \ldots, n+1$. Therefore, as $Z_1, \ldots, Z_{n+1}$ are exchangeable, so are $V_1, \ldots, V_{n+1}$, and applying Lemma 1 gives the result. □

Remark 1. Theorem 1 is stated for exchangeable samples (X_i, Y_i), $i = 1, \ldots, n+1$, which is (significantly) weaker than assuming i.i.d. samples. As we will see in what follows, it is furthermore possible to relax the exchangeability assumption, under an appropriate modification to the conformal procedure.

Remark 2. If we use an appropriate random tie-breaking rule (to determine the rank of V_{n+1} among $V_1, \ldots, V_{n+1}$), then the upper bounds in Lemma 1 and Theorem 1 hold in general (without assuming there are no ties almost surely).

The result in Theorem 1, albeit very simple to prove, is quite remarkable. It gives us a recipe for distribution-free prediction intervals, with nearly exact coverage, starting from an arbitrary score function $\mathcal{S}$; e.g., absolute residuals with respect to a fitted regression function from any base algorithm $\mathcal{A}$, as in (3). For more discussion of conformal prediction, its properties, and its variants, we refer to Vovk et al. (2005); Lei et al. (2018) and references therein.

2 Covariate shift

In this paper, we are concerned with settings in which the data (X_i, Y_i), $i = 1, \ldots, n+1$ are no longer exchangeable. Our primary focus will be a setting in which we observe data according to

$$\begin{aligned} (X_i, Y_i) &\overset{\text{i.i.d.}}{\sim} P = P_X \times P_{Y|X}, \; i = 1, \ldots, n, \\ (X_{n+1}, Y_{n+1}) &\sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}, \; \text{independently}. \end{aligned} \tag{6}$$

Notice that the conditional distribution of $Y|X$ is assumed to be the same for both the training and test data. Such a setting is often called *covariate shift* (e.g., see Shimodaira 2000; Quinonero-Candela et al. 2009; see also Remark 4 below for more discussion of this literature). The key realization is the following: if we know the ratio of test to training covariate likelihoods, $\mathrm{d}\widetilde{P}_X/\mathrm{d}P_X$, then we can still perform a modified of version conformal inference, using a quantile of a suitably weighted empirical distribution of nonconformity scores. The next subsection gives the details; following this, we describe a more computationally efficient conformal procedure, and give an empirical demonstration.

2.1 Weighted conformal prediction

In conformal prediction, we compare the value of a nonconformity score at a test point to the empirical distribution of nonconformity scores at the training points. In the covariate shift case, where the covariate distributions $P_X, \widetilde{P}_X$ in our training and test sets differ, we will now weight each nonconformity score $V_i^{(x,y)}$ (which measures how well $Z_i =$

(X_i, Y_i) conforms to the other points) by a probability proportional to the likelihood ratio $w(X_i) = \mathrm{d}\widetilde{P}_X(X_i)/\mathrm{d}P_X(X_i)$. Therefore, we will no longer be interested in the empirical distribution

$$\frac{1}{n+1}\sum_{i=1}^{n}\delta_{V_i^{(x,y)}} + \frac{1}{n+1}\delta_\infty,$$

as in Theorem 1, but rather, a weighted version

$$\sum_{i=1}^{n} p_i^w(x)\delta_{V_i^{(x,y)}} + p_{n+1}^w(x)\delta_\infty,$$

where the weights are defined by

$$p_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)},\ i = 1,\dots,n, \quad \text{and} \quad p_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}. \tag{7}$$

Due this careful weighting, draws from the discrete distribution in the second to last display resemble nonconformity scores computed on the test population, and thus, they "look exchangeable" with the nonconformity score at our test point. Our main result below formalizes these claims.

Corollary 1. *Assume data from the model (6). Assume that $\widetilde{P}_X$ is absolutely continuous with respect to P_X, and denote $w = \mathrm{d}\widetilde{P}_X/\mathrm{d}P_X$. For any score function $\mathcal{S}$, and any $\alpha \in (0,1)$, define for $x \in \mathbb{R}^d$,*

$$\widehat{C}_n(x) = \left\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \le \mathrm{Quantile}\left(1-\alpha; \sum_{i=1}^n p_i^w(x)\delta_{V_i^{(x,y)}} + p_{n+1}^w(x)\delta_\infty\right)\right\}, \tag{8}$$

where $V_i^{(x,y)}$, $i = 1,\dots,n+1$ are as in (4), and $p_i^w(x)$, $i = 1,\dots,n+1$ are as in (7). Then $\widehat{C}_n$ satisfies

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

Corollary 1 is a special case of a more general result that we present later in Theorem 2, which extends conformal inference to a setting in which the data are what we call *weighted exchangeable*. The proof is given in Section 3.5.

Remark 3. The same result as in Corollary 1 holds if $w \propto \mathrm{d}\widetilde{P}_X/\mathrm{d}P_X$, i.e., with an unknown normalization constant, because this normalization constant cancels out in the calculation of probabilities in (7).

Remark 4. Though the basic premise of covariate shift—and certainly the techniques employed in addressing it—are related to much older ideas in statistics, the specific setup (6) has recently generated great interest in machine learning: e.g., see Sugiyama and Muller (2005); Sugiyama et al. (2007); Quinonero-Candela et al. (2009); Agarwal et al. (2011); Wen et al. (2014); Reddi et al. (2015); Chen et al. (2016) and references therein). The focus is usually on correcting estimators, model evaluation, or model selection approaches to account for covariate shift. Correcting distribution-free prediction intervals, as we examine in this work, is (as far as we know) a new contribution. As one might expect, the likelihood ratio $\mathrm{d}\widetilde{P}_X/\mathrm{d}P_X$, a key component of our conformal construction in Corollary 1, also plays a critical role in much of the literature on covariate shift.

2.2 Weighted split conformal

In general, constructing a conformal prediction band can be computationally intensive, though this of course depends on the choice of score function. Consider the use of absolute residuals as in (3). To compute the nonconformity scores in (4), we must first run our base algorithm $\mathcal{A}$ on the data set $Z_{1:n} \cup \{(x,y)\}$ to produce a fitted regression function $\widehat{\mu}$, and then calculate

$$V_i^{(x,y)} = |Y_i - \widehat{\mu}(X_i)|,\ i = 1,\dots,n, \quad \text{and} \quad V_{n+1}^{(x,y)} = |y - \widehat{\mu}(x)|.$$

As the formation of the conformal set in (5) (ordinary case) or (8) (covariate shift case) requires us to do this for each $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ (which requires refitting $\widehat{\mu}$ each time), this can clearly become computationally burdersome.

A fast alternative, known as *split conformal prediction* (Papadopoulos et al., 2002; Lei et al., 2015), resolves this issue by taking the score function $\mathcal{S}$ to be defined using absolute residuals with respect to a *fixed* regression function, typically, one that has been trained on an preliminary data set. Denote by $(X_1^0, Y_1^0), \dots, (X_{n_0}^0, Y_{n_0}^0)$ this preliminary data set, used for fitting the regression function μ_0, and consider the score function

$$\mathcal{S}\big((x,y), Z\big) = |y - \mu_0(x)|.$$

Given data $(X_1, Y_1), \dots, (X_n, Y_n)$, independent of $(X_1^0, Y_1^0), \dots, (X_{n_0}^0, Y_{n_0}^0)$, we calculate

$$V_i^{(x,y)} = |Y_i - \mu_0(X_i)|, \ i = 1, \dots, n, \quad \text{and} \quad V_{n+1}^{(x,y)} = |y - \mu_0(x)|.$$

The conformal prediction interval (5), defined at a point $x \in \mathbb{R}^d$, reduces to

$$\widehat{C}_n(x) = \mu_0(x) \pm \text{Quantile}\Big(1-\alpha; \big\{|Y_i - \mu_0(X_i)|\big\}_{i=1}^n \cup \{\infty\}\Big), \tag{9}$$

and by Theorem 1 it has coverage at least $1-\alpha$, conditional on $(X_1^0, Y_1^0), \dots, (X_{n_0}^0, Y_{n_0}^0)$. This coverage result holds because, when we treat μ_0 as fixed (meaning, condition on $(X_1^0, Y_1^0), \dots, (X_{n_0}^0, Y_{n_0}^0)$), the scores $V_1^{(x,y)}, \dots, V_{n+1}^{(x,y)}$ scores are exchangeable for $(x,y) = (X_{n+1}, Y_{n+1})$, because $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$ are.

As split conformal prediction can be seen as a special case of conformal prediction, in which the regression function μ_0 is treated as fixed, Corollary 1 also applies to the split scenario, and guarantees that the band defined for $x \in \mathbb{R}^d$ by

$$\widehat{C}_n(x) = \mu_0(x) \pm \text{Quantile}\bigg(1-\alpha; \sum_{i=1}^n p_i^w(x)\delta_{|Y_i - \mu_0(X_i)|} + p_{n+1}^w(x)\delta_\infty\bigg), \tag{10}$$

where the probabilities are as in (7), has coverage at least $1-\alpha$, conditional on $(X_1^0, Y_1^0), \dots, (X_{n_0}^0, Y_{n_0}^0)$.

2.3 Airfoil data example

We demonstrate the use of conformal prediction in the covariate shift setting in an empirical example. 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 $d = 5$ dimensions (log frequency, angle of attack, chord length, free-stream velocity, and suction side log displacement thickness). For efficiency, we use the split conformal prediction methods in (9) and (10) for the unweighted and weighted case, respectively.

Creating training data, test data, and covariate shift. We repeated an experiment for 5000 trials, where for each trial we randomly partitioned the data $\{(X_i, Y_i)\}_{i=1}^N$ into three sets $D_{\text{pre}}, D_{\text{train}}, D_{\text{test}}$, and also constructed a covariate shift test set D_{shift}, which have the following roles.

- D_{pre}, containing 25% of the data, is used to prefit a regression function μ_0 via linear regression, to be used in constructing split conformal intervals.
- D_{train}, containing 25% of the data, is our training set, i.e., (X_i, Y_i), $i = 1, \dots, n$, used to compute the residual quantiles in the construction of the split conformal prediction intervals (9) and (10).
- D_{test}, containing 50% of the data, is our test set (as these data points are exchangeable with those in D_{train}, there is no covariate shift in this test set).
- D_{shift} is a second test set, constructed to simulate covariate shift, by sampling 25% of the points from D_{test} with replacement, with probabilities proportional to

$$w(x) = \exp(x^T\beta), \quad \text{where} \quad \beta = (-1, 0, 0, 0, 1). \tag{11}$$

As the original data points $D_{\text{train}} \cup D_{\text{test}}$ can be seen as draws from the same underlying distribution, we can view $w(x)$ as the likelihood ratio of covariate distributions between the test set D_{shift} and training set D_{train}. Note that the test covariate distribution $\widetilde{P}_X$, which satisfies $\mathrm{d}\widetilde{P}_X \propto \exp(x^T\beta)\mathrm{d}P_X$ as we have defined it here, is called an *exponential tilting* of the training covariate distribution P_X. Figure 1 visualizes the effect of this exponential tilting, in the airfoil data set, with our choice $\beta = (-1, 0, 0, 0, 1)$. Only the 1st and 5th dimensions of the covariate distribution are tilted; the bottom row of Figure 1 plots the marginal densities of the 1st and 5th covariates (estimated via kernel smoothing) before and after the tilt. The top row plots the response versus the 1st and 5th covariates, simply to highlight the fact that there is heteroskedasticity, and thus we might expect the shift in the covariate distribution to have some effect on the validity of the ordinary conformal prediction intervals.

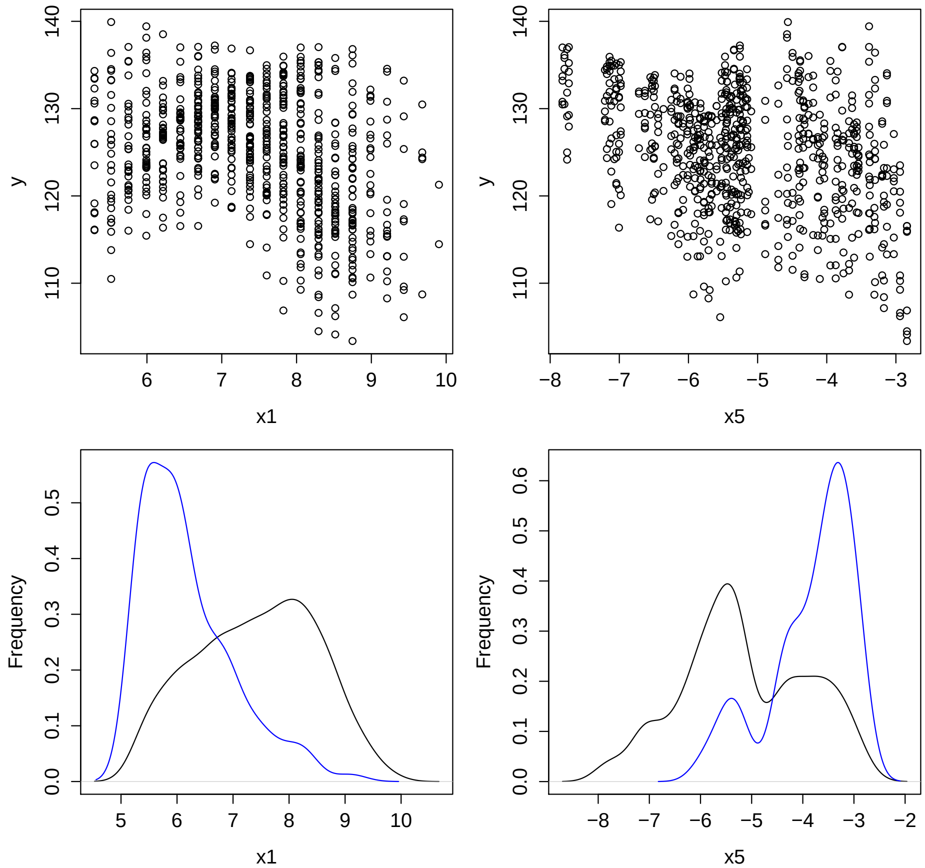

Figure 1: The top row plots the response in (a randomly chosen half of) the airfoil data set, versus the 1st and 5th covariates. The bottom row plots kernel density estimates for the 1st and 5th covariates, in black. Also displayed are kernel density estimates for the 1st and 5th covariates after exponential tilting (11), in blue.

Loss of coverage of ordinary conformal prediction under covariate shift. First, we examine the performance of ordinary split conformal prediction (9). The nominal coverage level was set to be 90% (meaning $\alpha = 0.1$), here and throughout. The results are shown in the top row of Figure 2. In each of the 5000 trials, we computed the empirical coverage from the split conformal intervals over points in the test sets, and the histograms show the distribution of these empirical coverages over the trials. We see that for the original test data D_{test} (no covariate shift, shown in red), split conformal works as expected, with the average of the empirical coverages (over the 5000 trials) being 90.2%; but for the nonuniformly subsampled test data D_{shift} (covariate shift, in blue), split conformal considerably undercovers, with its average coverage being 82.3%.

Coverage of weighted conformal prediction with oracle weights. Next, displayed in the middle row of Figure 2, we consider weighted split conformal prediction (10), to cover the points in D_{shift} (shown in orange). At the moment, we assume oracle knowledge of the true weight function w in (11) needed to calculate the probabilities in (7). We see that this brings the coverage back to the desired level, with the average coverage being 90.8%. However, the histogram is more dispersed than it is when there is no covariate shift (compare to the top row, in red). This is because, by using a quantile of the weighted empirical distribution of nonconformity scores, we are relying on a reduced "effective sample size". Given training points $X_1, \ldots, X_n$, and a likelihood ratio w of test to training covariate distributions, a popular heuristic formula from the covariate shift literature for the effective sample size of $X_1, \ldots, X_n$ is (Gretton et al., 2009; Reddi et al., 2015):

$$\widehat{n} = \frac{[\sum_{i=1}^n |w(X_i)|]^2}{\sum_{i=1}^n |w(X_i)|^2} = \frac{\|w(X_{1:n})\|_1^2}{\|w(X_{1:n})\|_2^2},$$

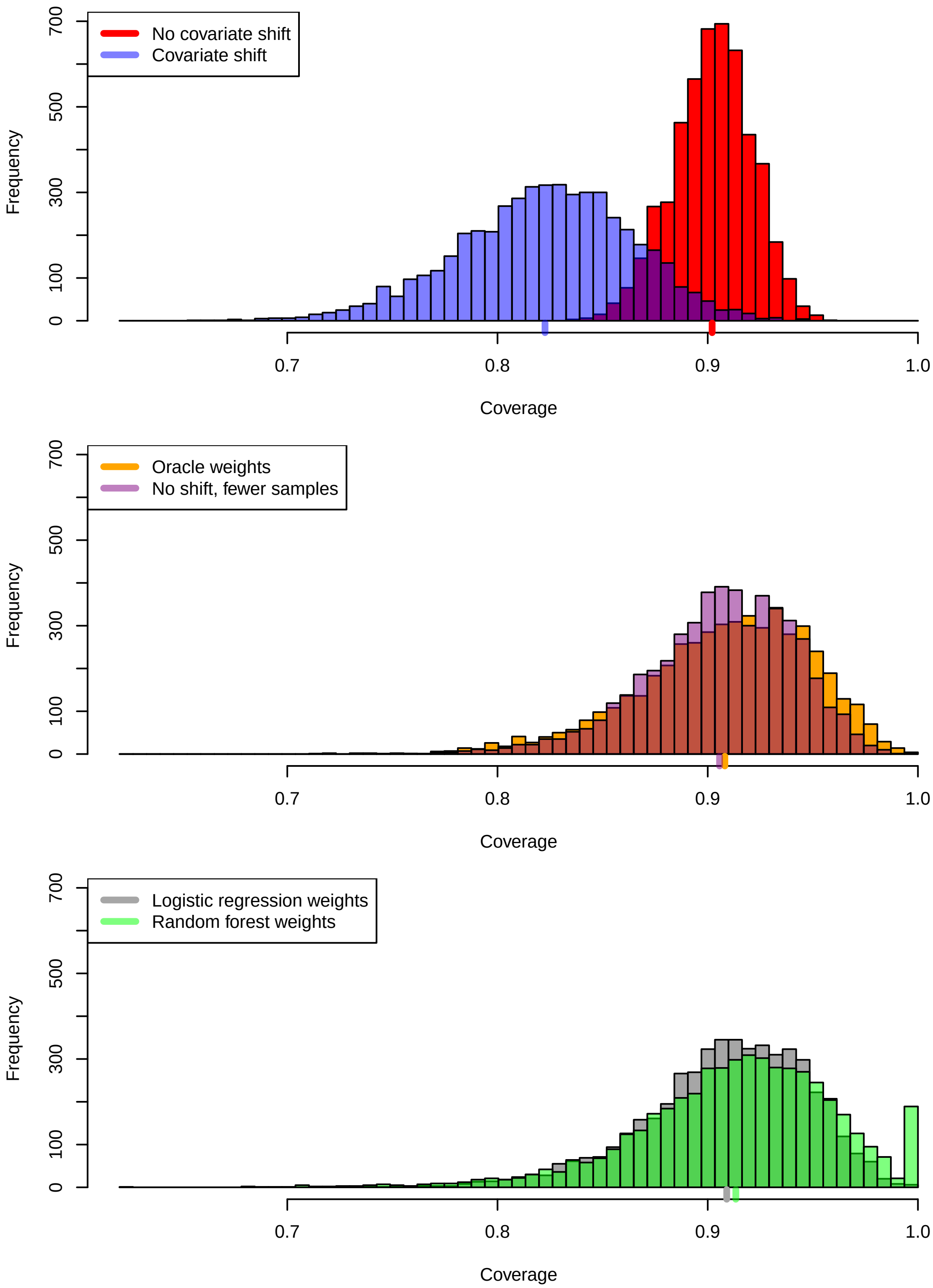

Figure 2: Empirical coverages of conformal prediction intervals, computed using 5000 different random splits of the airfoil data set. The averages of empirical coverages in each histogram are marked on the x-axis.

where we abbreviate $w(X_{1:n}) = (w(X_1), \dots, w(X_n)) \in \mathbb{R}^n$. To compare weighted conformal prediction against the unweighted method at the same effective sample size, in each trial, we ran unweighted split conformal on the original test set D_{test}, but we used only $\widehat{n}$ subsampled points from D_{train} to compute the quantile of nonconformity scores, when constructing the prediction interval (9). The results (the middle row of Figure 2, in purple) line up very closely with those from weighted conformal, which demonstrates that apparent overdispersion in the coverage histogram from the latter is fully explained by the reduced effective sample size.

Coverage of weighted conformal with estimated weights. Denote by $X_1, \dots, X_n$ the covariate points in D_{train} and by $X_{n+1}, \dots, X_{n+m}$ the covariate points in D_{shift}. Here we describe how to estimate $w = \mathsf{d}\widetilde{P}_X / \mathsf{d}P_X$, the likelihood ratio of interest, by applying logistic regression or random forests (more generally, any classifier that outputs estimated probabilities of class membership) to the feature-class pairs (X_i, C_i), $i = 1, \dots, n+m$, where $C_i = 0$ for $i = 1, \dots, n$ and $C_i = 1$ for $i = n+1, \dots, n+m$. Noting that

$$\frac{\mathbb{P}(C=1|X=x)}{\mathbb{P}(C=0|X=x)} = \frac{\mathbb{P}(C=1)}{\mathbb{P}(C=0)} \frac{\mathsf{d}\widetilde{P}_X}{\mathsf{d}P_X}(x),$$

we can take the conditional odds ratio $w(x) = \mathbb{P}(C=1|X=x)/\mathbb{P}(C=0|X=x)$ as an equivalent representation for the oracle weight function (since we only need to know the likelihood ratio up to a proportionality constant, recall Remark 3). Therefore, if $\widehat{p}(x)$ is an estimate of $\mathbb{P}(C=1|X=x)$ obtained by fitting a classifier to the data (X_i, C_i), $i = 1, \dots, n+m$, then we can use

$$\widehat{w}(x) = \frac{\widehat{p}(x)}{1 - \widehat{p}(x)} \tag{12}$$

as our estimated weight function for the calculation of probabilities (7), needed for conformal (8) or split conformal (10). There is in fact a sizeable literature on density ratio estimation, and the method just describe falls into a class called *probabilistic classification* approaches; two other classes are based on moment matching, and minimization of ϕ-divergences (e.g., Kullblack-Leibler divergence). For a comprehensive review of these approaches, and supporting theory, see Sugiyama et al. (2012).

The bottom row of Figure 2 shows the results from applying weighted split conformal prediction to cover the points in D_{shift}, where the weight function $\widehat{w}$ has been estimated as in (12), using logistic regression (in gray) and random forests (in green) to fit the class probability function $\widehat{p}$. Notice that logistic regression is well-specified in this example, because it assumes the log odds is a linear function of x, which is exactly as in (11). Random forests, of course, allows more flexibility in the fitted model. Both classification approaches deliver weights that translate into good coverage, with logistic regression yielding average coverage of 90.6% and random forests yielding average coverage of 90.9%. Further, their histograms are only a little more dispersed than that for the oracle weights (middle row, in orange). In the random forests histogram, there is a small bump at 1, which occurs because in a small number of cases split conformal using random forests weights returns very wide intervals (elaborated next).

Lengths of weighted conformal intervals. Figure 3 conveys the same setup as Figure 2, but displays histograms of the median lengths of prediction intervals rather than empirical coverages (meaning, in each of the 5000 trials, we ran unweighted or weighted split conformal prediction to cover test points, and report the median length of the prediction intervals over the test sets). We see no differences in the lengths of ordinary split conformal intervals (top row) when there is or is not covariate shift, as expected since these two settings differ only in the distributions of their test sets, but use the same procedure and have the same distribution of the training data. We see that the oracle-weighted split conformal intervals are longer than the ordinary split conformal intervals that use an equivalent effective sample size (middle row). This is also as expected, since in the former situation, the regression function μ_0 was fit on training data D_{train} of a different distribution than D_{shift}, and μ_0 itself should ideally be adjusted to account for covariate shift (plenty of methods for this are available from the covariate shift literature, but we left it unadjusted for simplicity). Lastly, we see that the random forests-weighted split conformal intervals are more variable, and in some cases, much longer, than the logistic regression-weighted split conformal intervals (bottom row, difficult to confirm visually because the bars in the histogram lie so close to the x-axis). In fact, with the random forests approach, about 3% of the intervals computed over all repetitions were infinite, which explains the bump in the coverage histogram at 1, at the bottom of Figure 2.

Weighted conformal when there is actually no covariate shift. Lastly, Figure 4 compares the empirical coverages and median lengths of split conformal intervals to cover points in D_{test} (no covariate shift), using the ordinary unweighted

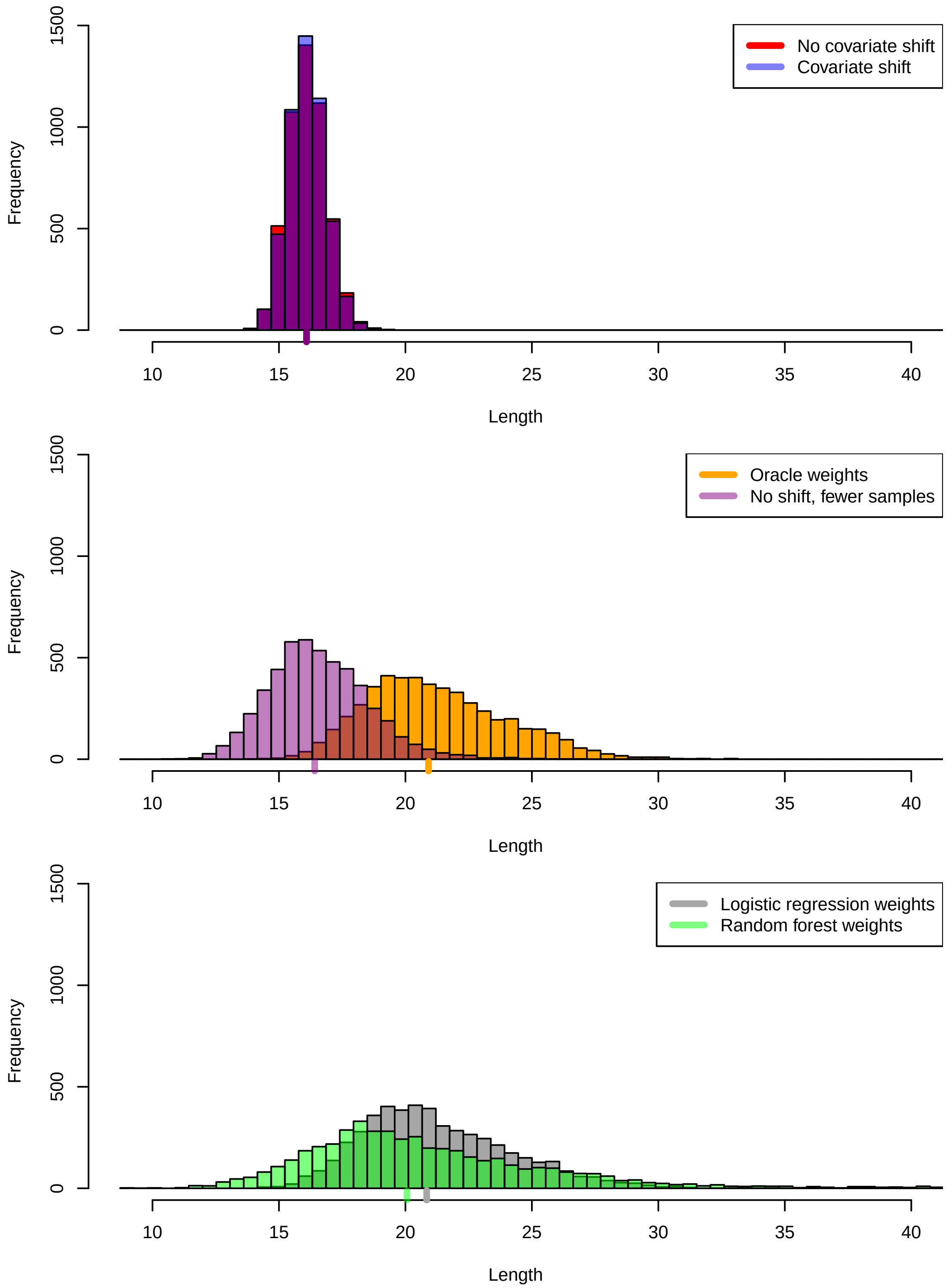

Figure 3: Median lengths of conformal prediction intervals, computed using 5000 different random splits of the airfoil data set. The averages of median lengths in each histogram are marked on the x-axis.

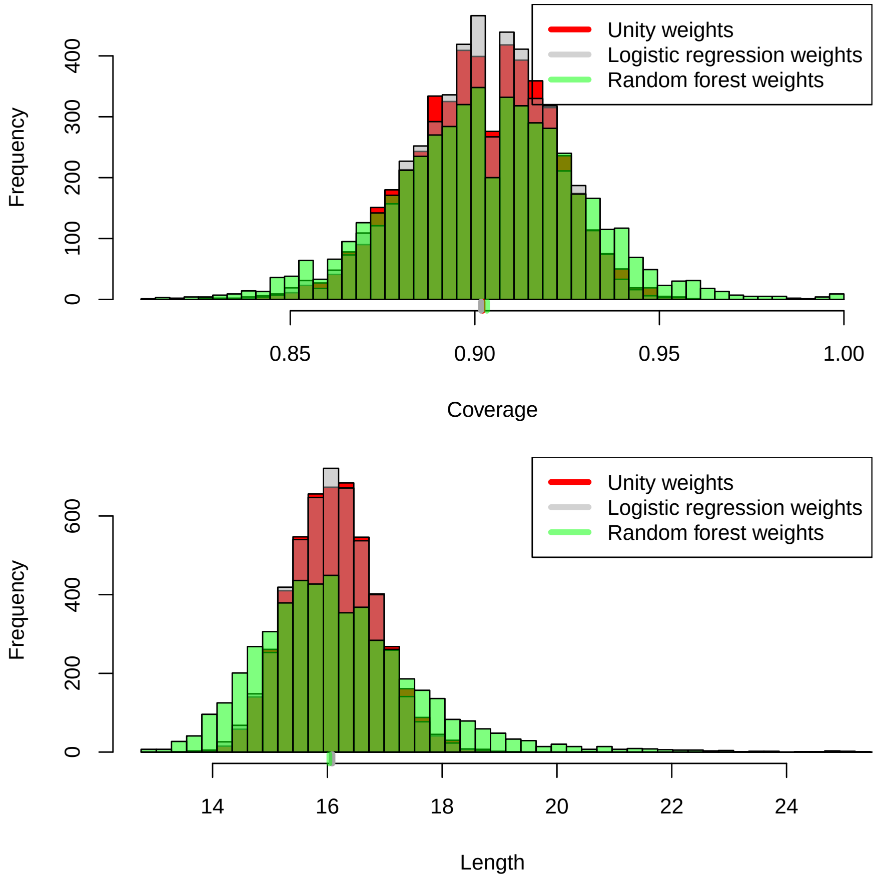

Figure 4: Empirical coverages and median lengths from conformal prediction, on the airfoil data set, with no covariate shift.

approach (in red), the logistic regression-weighted approach (in gray), and the random forests-weighted approach (in green). The unweighted and logistic regression approaches are very similar. The random forests approach yields slightly more dispersed coverages and lengths. This is because random forests are very flexible, and in the present case of no covariate shift, the estimated weights from random forests in each repetition are in general further from constant (compared to those from logistic regression). Still, random forests must not be overfitting dramatically here, since the coverages and lengths are still reasonable.

3 Weighted exchangeability

In this section, we develop a general result on conformal prediction for settings in which the data satisfy what we call *weighted exchangeability*. In the first subsection, we take a look back at the key quantile result in Lemma 1, and present an alternative proof from a somewhat different perspective. Then we precisely define weighted exchangeability, extend Lemma 1 to this new (and broader) setting, and extend conformal prediction as well. The last subsection provides a proof of Corollary 1, as a special case of our general conformal result.

3.1 Alternate proof of Lemma 1

The general strategy we pursue here is to condition on the unlabeled multiset of values obtained by our random variables $V_1, \dots, V_{n+1}$, and then inspect the probabilities that the last random variable V_{n+1} attains each one of these values. For simplicity, we assume that there are almost surely no ties among the scores $V_1, \dots, V_{n+1}$, so that we can work with sets rather than multisets (our arguments apply to the general case as well, but the notation is more cumbersome).

Denote by E_v the event that $\{V_1, \ldots, V_{n+1}\} = \{v_1, \ldots, v_{n+1}\}$, and consider

$$\mathbb{P}\{V_{n+1} = v_i \,|\, E_v\}, \; i = 1, \ldots, n+1. \tag{13}$$

Denote by f the probability density function[3] of the joint distribution $V_1, \ldots, V_{n+1}$. Exchangeability implies that

$$f(v_1, \ldots, v_{n+1}) = f(v_{\sigma(1)}, \ldots, v_{\sigma(n+1)})$$

for any permutation σ of the numbers $1, \ldots, n+1$. Thus, for each i, we have

$$\begin{aligned}\mathbb{P}\{V_{n+1} = v_i \,|\, E_v\} &= \frac{\sum_{\sigma:\sigma(n+1)=i} f(v_{\sigma(1)}, \ldots, v_{\sigma(n+1)})}{\sum_\sigma f(v_{\sigma(1)}, \ldots, v_{\sigma(n+1)})} \\ &= \frac{\sum_{\sigma:\sigma(n+1)=i} f(v_1, \ldots, v_{n+1})}{\sum_\sigma f(v_1, \ldots, v_{n+1})} \\ &= \frac{n!}{(n+1)!} = \frac{1}{n+1}. \end{aligned} \tag{14}$$

This shows that the distribution of $V_{n+1}|E_v$ is uniform on the set $\{v_1, \ldots, v_{n+1}\}$, i.e.,

$$V_{n+1}|E_v \sim \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{v_i},$$

and it follows immediately that

$$\mathbb{P}\bigg\{V_{n+1} \leq \text{Quantile}\bigg(\beta; \, \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{v_i}\bigg) \,\bigg|\, E_v\bigg\} \geq \beta.$$

On the event E_v, we have $\{V_1, \ldots, V_{n+1}\} = \{v_1, \ldots, v_{n+1}\}$, so

$$\mathbb{P}\bigg\{V_{n+1} \leq \text{Quantile}\bigg(\beta; \, \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{V_i}\bigg) \,\bigg|\, E_v\bigg\} \geq \beta.$$

Because this true for any v, we can marginalize to obtain

$$\mathbb{P}\bigg\{V_{n+1} \leq \text{Quantile}\bigg(\beta; \, \frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{V_i}\bigg)\bigg\} \geq \beta,$$

which, as argued in (2), is equivalent to the desired lower bound in the lemma. (The upper bound follows similarly.)

3.2 Weighted exchangeability

The alternate proof in the last subsection is certainly more complicated than our first proof of Lemma 1. What good did it do? Recall, we proceeded as if we had observed the unordered set of nonconformity scores $\{V_1, \ldots, V_{n+1}\} = \{v_1, \ldots, v_{n+1}\}$ but had forgotten the labels (as if we forgot which random variable V_i was associated with which value v_j). We then reduced the construction of a prediction interval for V_{n+1} to the computation of probabilities in (13), that the last random variable V_{n+1} attains each one of the observed values $v_1, \ldots, v_{n+1}$. The critical point was that this strategy isolated the role of exchangeability: it was used precisely to compute these probabilities, in (14). As we will show, the same probabilities in (13) can be calculated in a broader setting of interest, beyond the exchangeable one.

We first define a generalized notion of exchangeability.

[3] More generally, f may be the Radon-Nikodym derivative with respect to an arbitrary base measure; this measure may be discrete, continuous, or neither; henceforth, we will use the term "density" just for simplicity.

Definition 1. Random variables $V_1, \dots, V_n$ are said to be *weighted exchangeable*, with weight functions $w_1, \dots, w_n$, if the density[4] f of their joint distribution can be factorized as

$$f(v_1, \dots, v_n) = \prod_{i=1}^{n} w_i(v_i) \cdot g(v_1, \dots, v_n),$$

where g is any function that does not depend on the ordering of its inputs, i.e., $g(v_{\sigma(1)}, \dots, v_{\sigma(n)}) = g(v_1, \dots, v_n)$ for any permutation σ of $1, \dots, n$.

Clearly, weighted exchangeability with weight functions $w_i \equiv 1$ for $i = 1, \dots, n$ reduces to ordinary exchangeability. Furthermore, independent draws (where all marginal distributions are absolutely continuous with respect to, say, the first one) are always weighted exchangeable, with weight functions given by the appropriate Radon-Nikodym derivatives, i.e., likelihood ratios. This is stated next; the proof follows directly from Definition 1 and is omitted.

Lemma 2. *Let $Z_i \sim P_i$, $i = 1, \dots, n$ be independent draws, where each P_i is absolutely continuous with respect to P_1, for $i \geq 2$. Then $Z_1, \dots, Z_n$ are weighted exchangeable, with weight functions $w_1 \equiv 1$, and $w_i = \mathsf{d}P_i/\mathsf{d}P_1$, $i \geq 2$.*

Lemma 2 highlights an important special case (which we note, includes the covariate shift model in (6)). But it is worth being clear that our definition of weighted exchangeability encompasses more than independent sampling, and allows for a nontrivial dependency structure between the variables, just as exchangeability is broader than the i.i.d. case.

3.3 Weighted quantile lemma

Now we give a weighted generalization of Lemma 1.

Lemma 3. *Let Z_i, $i = 1, \dots, n+1$ be weighted exchangeable random variables, with weight functions $w_1, \dots, w_{n+1}$. Let $V_i = \mathcal{S}(Z_i, Z_{-i})$, where $Z_{-i} = Z_{1:(n+1)} \setminus \{Z_i\}$, for $i = 1, \dots, n+1$, and $\mathcal{S}$ is an arbitrary score function. Define*

$$p_i^w(z_1, \dots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}, \; i = 1, \dots, n+1, \tag{15}$$

where the summations are taken over permutations σ of the numbers $1, \dots, n+1$. Then for any $\beta \in (0,1)$,

$$\mathbb{P}\bigg\{ V_{n+1} \leq \text{Quantile}\bigg(\beta; \sum_{i=1}^{n} p_i^w(Z_1, \dots, Z_{n+1}) \delta_{V_i} + p_{n+1}^w(Z_1, \dots, Z_{n+1}) \delta_{\infty}\bigg)\bigg\} \geq \beta.$$

Proof. We follow the same general strategy in the alternate proof of Lemma 1 in Section 3.1. As before, we assume for simplicity that $V_1, \dots, V_{n+1}$ are distinct almost surely (but the result holds in the general case as well).

Let E_z denote the event that $\{Z_1, \dots, Z_{n+1}\} = \{z_1, \dots, z_{n+1}\}$, and let $v_i = \mathcal{S}(z_i, z_{-i})$, where $z_{-i} = z_{1:(n+1)} \setminus \{z_i\}$, for $i = 1, \dots, n+1$. Let f be the density function of the joint sample $Z_1, \dots, Z_{n+1}$. For each i,

$$\mathbb{P}\{V_{n+1} = v_i \,|\, E_z\} = \mathbb{P}\{Z_{n+1} = z_i \,|\, E_z\} = \frac{\sum_{\sigma:\sigma(n+1)=i} f(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})}{\sum_{\sigma} f(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})},$$

and as $Z_1, \dots, Z_{n+1}$ are weighted exchangeable,

$$\begin{aligned}
\frac{\sum_{\sigma:\sigma(n+1)=i} f(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})}{\sum_{\sigma} f(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})} &= \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) \cdot g(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) \cdot g(z_{\sigma(1)}, \dots, z_{\sigma(n+1)})} \\
&= \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) \cdot g(z_1, \dots, z_{n+1})}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) \cdot g(z_1, \dots, z_{n+1})} \\
&= p_i^w(z_1, \dots, z_{n+1}).
\end{aligned}$$

[4] As before, f may be the Radon-Nikodym derivative with respect to an arbitrary base measure.

In other words,

$$V_{n+1}|E_z \sim \sum_{i=1}^{n+1} p_i^w(z_1,\dots,z_{n+1})\delta_{v_i},$$

which implies that

$$\mathbb{P}\bigg\{V_{n+1} \le \text{Quantile}\bigg(\beta;\, \sum_{i=1}^{n+1} p_i^w(z_1,\dots,z_{n+1})\delta_{v_i}\bigg) \,\bigg|\, E_z\bigg\} \ge \beta.$$

This is equivalent to

$$\mathbb{P}\bigg\{V_{n+1} \le \text{Quantile}\bigg(\beta;\, \sum_{i=1}^{n+1} p_i^w(Z_1,\dots,Z_{n+1})\delta_{V_i}\bigg) \,\bigg|\, E_z\bigg\} \ge \beta,$$

and after marginalizing,

$$\mathbb{P}\bigg\{V_{n+1} \le \text{Quantile}\bigg(\beta;\, \sum_{i=1}^{n+1} p_i^w(Z_1,\dots,Z_{n+1})\delta_{V_i}\bigg)\bigg\} \ge \beta.$$

Finally, as in (2), this is equivalent to the claim in the lemma. □

Remark 5. When $V_1,\dots,V_{n+1}$ are exchangeable, we have $w_i \equiv 1$ for $i=1,\dots,n$, and thus $p_i^w \equiv 1$ for $i=1,\dots,n$ as well. In this special case, then, the lower bound in Lemma 3 reduces to the ordinary unweighted lower bound in Lemma 1. Meanwhile, a obtaining meaningful upper bound on the probability in question in Lemma 3, as was done in Lemma 1 (under the assumption of no ties, almost surely), does not seem possible without further conditions on the weight functions in consideration. This is because the largest "jump" in the discontinuous cumulative distribution function of $V_{n+1}|E_z$ is of size $\max_{i=1,\dots,n+1} p_i^w(z_1,\dots,z_{n+1})$, which can potentially be very large; by comparison, in the unweighted case, this jump is always of size $1/(n+1)$.

3.4 Weighted conformal prediction

A weighted version of conformal prediction follows immediately from Lemma 3.

Theorem 2. *Assume that $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i=1,\dots,n+1$ are weighted exchangeable with weight functions $w_1,\dots,w_{n+1}$. For any score function $\mathcal{S}$, and any $\alpha \in (0,1)$, define the weighted conformal band (based on the first n samples) at a point $x \in \mathbb{R}^d$ by*

$$\widehat{C}_n(x) = \bigg\{y \in \mathbb{R} : V_{n+1}^{(x,y)} \le \text{Quantile}\bigg(1-\alpha;\, \sum_{i=1}^{n} p_i^w\big(Z_1,\dots,Z_n,(x,y)\big)\delta_{V_i^{(x,y)}} + p_{n+1}^w\big(Z_1,\dots,Z_n,(x,y)\big)\delta_\infty\bigg)\bigg\}, \tag{16}$$

where $V_i^{(x,y)}$, $i=1,\dots,n+1$ are as in (4), and p_i^w, $i=1,\dots,n+1$ are as in (15). Then $\widehat{C}_n$ satisfies

$$\mathbb{P}\Big\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\Big\} \ge 1-\alpha.$$

Proof. Abbreviate $V_i = V_i^{(X_{n+1},Y_{n+1})}$, $i=1,\dots,n+1$. By construction,

$$Y_{n+1} \in \widehat{C}_n(X_{n+1}) \iff V_{n+1} \le \text{Quantile}\bigg(1-\alpha;\, \sum_{i=1}^{n} p_i^w(Z_1,\dots,Z_{n+1})\delta_{V_i} + p_{n+1}^w(Z_1,\dots,Z_{n+1})\delta_\infty\bigg),$$

and applying Lemma 3 gives the result. □

Remark 6. As explained in Section 2.2, the split conformal method is simply a special case of the conformal prediction framework, where we take the score function to be $\mathcal{S}((x,y),Z) = |y-\mu_0(x)|$, with μ_0 precomputed on a preliminary data set $(X_1^0,Y_1^0),\dots,(X_{n_0}^0,Y_{n_0}^0)$. Hence, the result in Theorem 2 carries over to the split conformal method as well, in which case the weighted conformal prediction interval in (16) simplifies to

$$\widehat{C}_n(x) = \mu_0(x) \pm \text{Quantile}\bigg(1-\alpha;\, \sum_{i=1}^{n} p_i^w\big(Z_1,\dots,Z_n,(x,y)\big)\delta_{|Y_i-\mu_0(X_i)|} + p_{n+1}^w\big(Z_1,\dots,Z_n,(x,y)\big)\delta_\infty\bigg).$$

By Theorem 2, this has coverage at least $1-\alpha$, conditional on $(X_1^0,Y_1^0),\dots,(X_{n_0}^0,Y_{n_0}^0)$.

3.5 Proof of Corollary 1

We return to the case of covariate shift, and show that Corollary 1 follows from the general weighted conformal result. By Lemma 2, we know that the independent draws $Z_i = (X_i, Y_i)$, $i = 1, \dots, n+1$ are weighted exchangeable with $w_i \equiv 1$ for $i = 1, \dots, n$, and $w_{n+1}((x, y)) = w(x)$. In this special case, the probabilities in (15) simplify to

$$p_i^w(z_1, \dots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} w(x_i)}{\sum_\sigma w(x_{\sigma(n+1)})} = \frac{w(x_i)}{\sum_{j=1}^{n+1} w(x_j)},\ i = 1, \dots, n+1,$$

in other words, $p_i^w(Z_1, \dots, Z_n, (x, y)) = p_i^w(x)$, $i = 1, \dots, n+1$, where the latter are as in (7). Applying Theorem 2 gives the result.

4 Discussion

We described an extension of conformal prediction to handle weighted exchangeable data. This covers exchangeable data, and independent (but not identically distributed) data, as special cases. In general, the new weighted methodology requires computing quantiles of a weighted discrete distribution of nonconformity scores, which is combinatorially hard. But the computations simplify dramatically for a case of significant practical interest, in which the test covariate distribution $\widetilde{P}_X$ differs from the training covariate distribution P_X by a known likelihood ratio $\mathsf{d}\widetilde{P}_X/\mathsf{d}P_X$ (and the conditional distribution $P_{Y|X}$ remains unchanged). In this case, known as covariate shift, the new weighted conformal prediction methodology is just as easy, computationally, as ordinary conformal prediction. When the likelihood ratio $\mathsf{d}\widetilde{P}_X/\mathsf{d}P_X$ is not known, it can be estimated given access to unlabeled data (test covariate points), which we showed empirically, on a low-dimensional example, can still yield correct coverage.

Beyond the setting of covariate shift that we have focused on (as the main application in this paper), our weighted conformal methodology can be applied to several other closely related settings, where ordinary conformal prediction will not directly yield correct coverage. We discuss three such settings below.

Graphical models with covariate shift. Assume that the training data $(Z, X, Y) \sim P$ has the Markovian structure $Z \to X \to Y$. As an example, to make matters concrete, suppose Z is a low-dimensional covariate (such as ancenstry information), X is a high-dimensional set of features for a person (such as genetic measurements), and Y is a real-valued outcome of interest (such as life expectancy). Suppose that on the test data $(Z, X, Y) \sim \widetilde{P}$, the distribution of Z has changed, causing a change in the distribution of X, and thus causing a change in the distribution of the unobserved Y (however the distribution of $X|Z$ is unchanged). One plausible solution to this problem would be to just ignore Z in both training and test sets, and run weighted conformal prediction on only (X, Y), treating this like a usual covariate shift problem. But, as X is high-dimensional, this would require estimating a ratio of two high-dimensional densities, which would be difficult. As Z is low-dimensional, we can instead estimate the weights by estimating the likelihood ratio of Z between test and training sets, which follows because for the joint covariate (Z, X),

$$\frac{\widetilde{P}_{Z,X}(z, x)}{P_{Z,X}(z, x)} = \frac{\widetilde{P}_Z(z) P_{X|Z=z}(x)}{P_Z(z) P_{X|Z=z}(x)} = \frac{\widetilde{P}_Z(z)}{P_Z(z)}.$$

This may be a more tractable quantity to estimate for the purpose of weighted conformal inference. These ideas may be generalized to more complex graphical settings, which we leave to future work.

Missing covariates with known summaries. As another concrete example, suppose that hospital A has collected a private training data set $(Z, X, Y) \sim P^A$ where $Z \in \{0, 1\}$ is a sensitive patient covariate, $X \in \mathbb{R}^d$ represents other covariates, and $Y \in \mathbb{R}$ is a response that is expensive to measure. Hospital B also has its own data set, but in order to save money and not measure the responses for their patients, it asks hospital A for help to produce prediction intervals for these responses. Instead of sharing the collected data $(Z, X) \sim P^B$ for each patient with hospital A, due to privacy concerns, hospital B only provides hospital A with the X covariate for each patient, along with a summary statistic for Z, representing the fraction of Z values that equal one (more accurately, the probability of drawing a patient with $Z = 1$ from their underlying patient population). Assume that $P^A_{X|Z} = P^B_{X|Z}$ (e.g., if Z is the sex of the patient, then this assumes there is one joint distribution on X for males and one for females, which does not depend on the hospital). The likelihood ratio of covariate distributions thus again reduces to calculating the likelihood ratio of Z between P^B and P^A, which we can easily do, and use weighted conformal prediction.

Towards local conditional coverage? We finish by describing how the weighted conformal prediction methodology can be used to construct prediction bands that satisfy a certain approximate (locally-smoothed) notion of conditional coverage. Given i.i.d. points (X_i, Y_i), $i = 1, \dots, n+1$, consider instead of our original goal (1),

$$\mathbb{P}\Big\{Y_{n+1} \in \widehat{C}_n(x_0) \,\Big|\, X_{n+1} = x_0\Big\} \geq 1 - \alpha. \tag{17}$$

This is (exact) conditional coverage at $x_0 \in \mathbb{R}^d$. As it turns out, maintaining that (17) hold for almost all[5] $x_0 \in \mathbb{R}^d$ and all distributions P is far too strong: Vovk (2012); Lei and Wasserman (2014) prove that any method with such a property must yield an interval $\widehat{C}_n(x_0)$ with infinite expected length at any non-atom point[6] x_0, for any underlying distribution P.

Thus we must relax (17) and seek some notion of approximate conditional coverage, if we hope to achieve it with a nontrivial prediction band. Some relaxations were recently considered in Barber et al. (2019), most of which were also impossible to achieve in a nontrivial way. A different, but natural relaxation of (17) is

$$\frac{\int \mathbb{P}\big(Y_{n+1} \in \widehat{C}_n(x_0) \,|\, X_{n+1} = x\big) K\big(\frac{x - x_0}{h}\big) \, \mathsf{d}P_X(x)}{\int K\big(\frac{x - x_0}{h}\big) \, \mathsf{d}P_X(x)} \geq 1 - \alpha, \tag{18}$$

where K is kernel function and $h > 0$ is bandwidth parameter. Here we are asking for a prediction band whose average conditional coverage, in some locally-weighted sense around x, is at least $1 - \alpha$. We can equivalently write (18) as

$$\mathbb{P}\Big\{Y_{n+1} \in \widehat{C}_n(x_0) \,\Big|\, X_{n+1} = x_0 + h\omega\Big\} \geq 1 - \alpha, \tag{19}$$

where the probability is taken over the $n+1$ data points and an independent draw ω from a distribution whose density is proportional to K. (For example, when we take $K(x) = \exp(-\|x\|_2^2/2)$, the Gaussian kernel, we have $\omega \sim N(0, I)$.) As we can see from (18) (or (19)), this kind of locally-weighted guarantee should be close to a guarantee on conditional coverage, when the bandwidth h is small.

In order to achieve (18) in a distribution-free manner, we can invoke the weighted conformal inference methodology. In particular, note that we can once more rewrite (19) as

$$\mathbb{P}_{x_0}\Big\{Y_{n+1} \in \widehat{C}_n(\widetilde{X}_{n+1})\Big\} \geq 1 - \alpha, \tag{20}$$

where the probability is taken over training points (X_i, Y_i), $i = 1, \dots, n$, that are i.i.d. from $P = P_X \times P_{Y|X}$ and an independent test point $(\widetilde{X}_{n+1}, Y_{n+1})$, from $\widetilde{P} = \widetilde{P}_X \times P_{Y|X}$, where $\mathsf{d}\widetilde{P}_X / \mathsf{d}P_X \propto K((\cdot - x_0)/h)$. The subscript x_0 on the probability operator in (20) emphasizes the dependence of the test covariate distribution on x_0. Note that this precisely fits into the covariate shift setting (6). To be explicit, for any score function $\mathcal{S}$, and any $\alpha \in (0,1)$, given a center point $x_0 \in \mathbb{R}^d$ of interest, define

$$\widehat{C}_n(x) = \left\{ y \in \mathbb{R} : V_{n+1}^{(x,y)} \leq \mathrm{Quantile}\left(1 - \alpha;\, \frac{\sum_{i=1}^n K\big(\frac{X_i - x_0}{h}\big)\delta_{V_i^{(x,y)}} + K\big(\frac{x - x_0}{h}\big)\delta_\infty}{\sum_{i=1}^n K\big(\frac{X_i - x_0}{h}\big) + K\big(\frac{x - x_0}{h}\big)}\right)\right\},$$

where $V_i^{(x,y)}$, $i = 1, \dots, n+1$, are as in (4). Then by Corollary 1,

$$\mathbb{P}_{x_0}\Big\{Y_{n+1} \in \widehat{C}_n(X_{n+1}; x_0)\Big\} \geq 1 - \alpha. \tag{21}$$

This is "almost" of the desired form (20) (equivalently (18), or (19)), except for one important caveat. The prediction band $\widehat{C}_n(\,\cdot\,; x_0)$ in (21) was constructed based on the center point x_0 (as suggested by our notation) to have the property in (21). If we were to ask for local conditional coverage at a new point x_0, then the entire band $\widehat{C}_n(\cdot; x_0)$ must change (must be recomputed) in order to accommodate the new guarantee.

In other words, we have not provided a recipe for the construction of a single band $\widehat{C}_n$ based on the training data (X_i, Y_i), $i = 1, \dots, n$ that has the property (20) at all $x_0 \in \mathbb{R}^d$ (or almost all $x_0 \in \mathbb{R}^d$). We have only described a way to satisfy such a property if the center point x_0 were specified in advance. Combined with our somewhat pessimistic results in Barber et al. (2019), many important practical and philosophical problems in assumption-lean conditional predictive inference remain open.

[5]This is meant to be interpreted with respect to P_X, i.e., for all $x_0 \in \mathbb{R}^d$ except on a set whose probability under P_X is zero.

[6]This is a point x_0 in the support of P_X such that $P_X\{B_r(x_0)\} \to 0$ as $r \to 0$, where $B_r(x_0)$ is the ball of radius r centered at x_0.

Acknowledgements

The authors thank the American Institute of Mathematics for supporting and hosting our collaboration. R.F.B. was partially supported by the National Science Foundation under grant DMS-1654076 and by an Alfred P. Sloan fellowship. E.J.C. was partially supported by the Office of Naval Research under grant N00014-16-1-2712, by the National Science Foundation under grant DMS-1712800, and by a generous gift from TwoSigma. R.J.T. was partially supported by the National Science Foundation under grant DMS-1554123.

References

Deepak Agarwal, Lihong Li, and Alex Smola. Linear-time estimators for propensity scores. *International Conference on Artificial Intelligence and Statistics*, 2011.

Rina Foygel Barber, Emmanuel J. Candes, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*, 2019.

Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. *Annual Conference on Learning Theory*, 2014.

Xiangli Chen, , Mathew Monfort, Anqi Liu, and Brian Da Ziebart. Robust covariate shift regression. *International Conference on Artificial Intelligence and Statistics*, 2016.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019. URL `http://archive.ics.uci.edu/ml`.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Scholkopf. Covariate shift by kernel mean matching. In *Dataset Shift in Machine Learning*, chapter 8, pages 131–160. MIT Press, 2009.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B*, 76(1):71–96, 2014.

Jing Lei, Alessandro Rinaldo, and Larry Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1–2):29–43, 2015.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. *European Conference on Machine Learning*, 2002.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

Sashank J. Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. *AAAI Conference on Artificial Intelligence*, 2015.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9: 371–421, 2008.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Masashi Sugiyama and Klaus-Robert Muller. Input-dependent estimation of generalization error under covariate shift. *Statistics & Decisions*, 23(4):249–279, 2005.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Muller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(985–1005), 2007.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Asian Conference on Machine Learning*, 2012.

Vladimir Vovk. Transductive conformal predictors. *Symposium on Conformal and Probabilistic Prediction with Applications*, 2013.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Vladimir Vovk, Ilia Nouretdinov, and Alex Gammerman. On-line predictive linear regression. *Annals of Statistics*, 37 (3):1566–1590, 2009.

Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. *International Conference on Machine Learning*, 2014.