The Annals of Statistics
2023, Vol. 51, No. 2, 816–845
https://doi.org/10.1214/23-AOS2276

CONFORMAL PREDICTION BEYOND EXCHANGEABILITY

By Rina Foygel Barber[1,a], Emmanuel J. Candès[2,b], Aaditya Ramdas[3,c] and Ryan J. Tibshirani[3,d]

[1]*Department of Statistics, University of Chicago,* [a]*rina@uchicago.edu*
[2]*Departments of Statistics and Mathematics, Stanford University,* [b]*candes@stanford.edu*
[3]*Departments of Statistics and Machine Learning, Carnegie Mellon University,* [c]*aramdas@cmu.edu,* [d]*ryantibs@cmu.edu*

Conformal prediction is a popular, modern technique for providing valid predictive inference for arbitrary machine learning models. Its validity relies on the assumptions of exchangeability of the data, and symmetry of the given model fitting algorithm as a function of the data. However, exchangeability is often violated when predictive models are deployed in practice. For example, if the data distribution drifts over time, then the data points are no longer exchangeable; moreover, in such settings, we might want to use a nonsymmetric algorithm that treats recent observations as more relevant. This paper generalizes conformal prediction to deal with both aspects: we employ weighted quantiles to introduce robustness against distribution drift, and design a new randomization technique to allow for algorithms that do not treat data points symmetrically. Our new methods are provably robust, with substantially less loss of coverage when exchangeability is violated due to distribution drift or other challenging features of real data, while also achieving the same coverage guarantees as existing conformal prediction methods if the data points are in fact exchangeable. We demonstrate the practical utility of these new tools with simulations and real-data experiments on electricity and election forecasting.

1. Introduction. The field of conformal prediction addresses a challenging modern problem: given a "black box" algorithm that fits a predictive model to available training data, how can we calibrate prediction intervals around the output of the model so that these intervals are guaranteed to achieve some desired coverage level?

As an example, consider a holdout set approach. Suppose we have a pre-fitted model $\widehat{\mu}$ mapping features X to a prediction of a real-valued variable Y (e.g., $\widehat{\mu}$ is the output of some machine learning algorithm trained on a prior data set), and a fresh holdout set of data $(X_1, Y_1), \dots, (X_n, Y_n)$ not used for training. We can then use the empirical quantiles of the errors $|Y_i - \widehat{\mu}(X_i)|$ on the holdout set to compute a prediction interval around our prediction $\widehat{\mu}(X_{n+1})$ that aims to cover the unseen response Y_{n+1}. Split conformal prediction (Vovk, Gammerman and Shafer (2005)) formalizes this method, and gives guaranteed predictive coverage when the data points (X_i, Y_i) are drawn i.i.d. from *any* distribution (see Section 2). However, the validity of this method hinges on the assumption that the data points are drawn independently from the *same* distribution, or more generally, that $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$ are exchangeable.

In many applied domains, this assumption is often substantially violated, due to distribution drift, correlations between data points, or other phenomena. As an example, Figure 1 shows results from an experiment on a real data set monitoring electricity usage in Australia (the ELEC2 data set (Harries (1999)), which we return to in Section 5.2). We see that over

Received March 2022; revised February 2023.
MSC2020 subject classifications. Primary 62G35; secondary 62F40, 62G86.
Key words and phrases. Conformal prediction, distribution-free inference, robust statistics, jackknife, exchangeability.

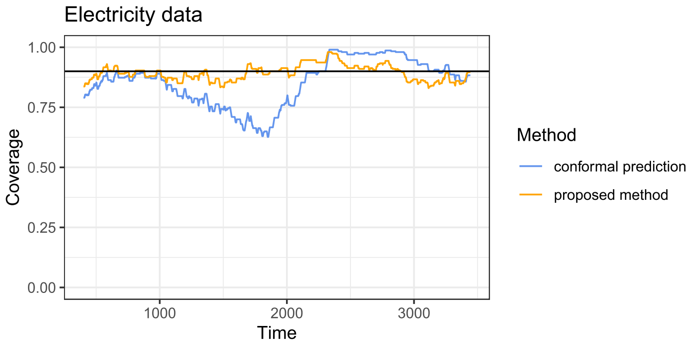

FIG. 1. *Empirical results from a real data set (details will be given in Section 5.2).*

a substantial stretch of time, conformal prediction loses coverage, its intervals decreasing far below the target 90% coverage level, while our proposed method, *nonexchangeable conformal prediction*, is able to maintain approximately the desired coverage level. In this paper, we will see how to quantify the loss of coverage due to violations of exchangeability, and how we can modify the conformal prediction methodology to regain predictive coverage even in the presence of distribution drift or other violations of exchangeability.

1.1. *Beyond exchangeability.* In this paper, we will consider three important classes of methods for distribution-free prediction: split conformal, full conformal, and the jackknife+. (We give background on split and full conformal in Section 2, and on jackknife+ in Appendix B.) These methods rely on exchangeability in two different ways:

- The data $Z_i = (X_i, Y_i)$ are assumed to be exchangeable (for example, i.i.d.).
- The algorithm $\mathcal{A}$, which maps data to a fitted model $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$, is assumed to treat the data points symmetrically, to ensure that exchangeability of the data points Z_i still holds even after we observe the fitted model(s).

In this work, we aim to provide distribution-free prediction guarantees when we drop both of these assumptions:

- We may have data points Z_i that are not exchangeable—for instance, they may be independent but nonidentically distributed (e.g., due to distribution drift), or there may be dependence among them that creates nonexchangeability (e.g., correlation over space or time).
- We may wish to use an algorithm $\mathcal{A}$ that does not treat the input data points symmetrically—for example, if Z_i denotes data collected at time i, we may prefer to fit a model $\widehat{\mu}$ that places higher weight on more recent data points.

1.2. *Our contributions.* We generalize the split conformal, full conformal, and jackknife+ methods (detailed later) to allow for both of these sources of nonexchangeability. Our procedures can recover the original variants if a symmetric algorithm is employed. We will provide coverage guarantees that are identical to existing guarantees if the data points are in fact exchangeable, and only slightly lower if the deviation from exchangeability is mild.

To elaborate, let us define the *coverage gap* as the loss in coverage compared to what is achieved under exchangeability. For example, in split conformal prediction run with a desired coverage level $1 - \alpha$, we have

$$\text{Coverage gap} = (1 - \alpha) - \mathbb{P}\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\},$$

since, under exchangeability, the method guarantees coverage with probability $1-\alpha$. To give an informal preview of our results, we write $Z_i = (X_i, Y_i)$ to denote the ith data point and

$$Z = (Z_1, \dots, Z_{n+1}) \tag{1}$$

to denote the full (training and test) data sequence, and let Z^i denote this sequence after swapping the test point (X_{n+1}, Y_{n+1}) with the ith training point (X_i, Y_i):

$$Z^i = (Z_1, \dots, Z_{i-1}, Z_{n+1}, Z_{i+1}, \dots, Z_n, Z_i). \tag{2}$$

To enable robustness, our methods will allow for weights: let $w_i \in [0, 1]$ denote a prespecified weight placed on data point i. We will see that the coverage gap can be bounded as

$$\text{Coverage gap} \leq \frac{\sum_{i=1}^n w_i \cdot \mathsf{d}_{\mathsf{TV}}(Z, Z^i)}{1 + \sum_{i=1}^n w_i}, \tag{3}$$

where d_{TV} denotes the total variation distance between distributions. Notably, we do not make any assumption on the joint distribution of the $n+1$ points. Of course, the result will only be meaningful if we are able to select fixed (non-data-dependent) weights w_i such that this upper bound is likely to be small. In particular, if we use these methods, we are implicitly assuming that this weighted sum of total variation terms is small. In contrast, past work on conformal prediction in a model-free setting has relied on the assumption that the data are *exactly* exchangeable, since no prior results offered an analysis of the coverage gap under violations of exchangeability. See Section 4.3 for further discussion.

Note that the upper bound in (3) is a far stronger result than simply asking whether the data is "nearly exchangeable." For instance, in a time series, it might be the case that $\mathsf{d}_{\mathsf{TV}}(Z, Z^i)$ is quite small but $\mathsf{d}_{\mathsf{TV}}(Z, Z_\pi) \approx 1$ for most permutations π. In words, if the observations are noisy, then permuting only two data points might not be detectable—but if there is nonstationarity or dependence over time then Z is likely to be far from exchangeable.

Several further remarks are in order. First, for $w_i \equiv 1$ and a symmetric algorithm, the proposed weighted methods will reduce to the usual conformal (or jackknife+) methods. Thus, the result (3) also quantifies the degradation in coverage of standard methods in nonexchangeable settings. Second, this result has new implications in exchangeable settings: if the data points are in fact exchangeable (with i.i.d. as a special case), then $Z \overset{\mathrm{d}}{=} Z^i$ and the coverage gap bound in (3) is equal to zero (here we use $\overset{\mathrm{d}}{=}$ for equality in distribution). Therefore, our use of a weighted conformal procedure (rather than choosing $w_i \equiv 1$, which is the original unweighted procedure) does not hurt coverage if the data are exchangeable. Finally, the result provides insights on why one might prefer to use our new weighted procedures in (possibly) nonexchangeable settings: it can provide robustness in the case of distribution shift. To elaborate, consider a setting where the data points Z_i are independent, but are not identically distributed due to distribution drift. The following result relates $\mathsf{d}_{\mathsf{TV}}(Z, Z^i)$ to the distributions of the individual data points.

LEMMA 1. *If $Z_1, \dots, Z_{n+1}$ are independent, then*

$$\mathsf{d}_{\mathsf{TV}}(Z, Z^i) \leq 2\mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1}) - \mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1})^2 \leq 2\mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1}).$$

Combining this lemma with (3), we can see that if we are able to place small weights w_i on data points Z_i with large total variation distance $\mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1})$, then the coverage gap will be low. For example, under distribution drift, we might have $\mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1})$ decreasing with i; we can achieve a low coverage gap by using, say, weights $w_i = \rho^{n+1-i}$ for some $\rho < 1$. We will return to this example in Section 4.4.

We will also see that the result in (3) actually stems from a stronger result:

$$\text{Coverage gap} \leq \frac{\sum_{i=1}^n w_i \cdot \mathsf{d}_{\mathsf{TV}}(R(Z), R(Z^i))}{1 + \sum_{i=1}^n w_i}. \tag{4}$$

Here $R(Z)$ denotes a vector of residuals: for split conformal prediction, this is the vector with entries $R(Z)_i = |Y_i - \widehat{\mu}(X_i)|$, where $\widehat{\mu}$ is a pre-fitted model, while for full conformal the entries are again given by $R(Z)_i = |Y_i - \widehat{\mu}(X_i)|$ but now $\widehat{\mu}$ is the model obtained by running $\mathcal{A}$ on the entire data sequence Z. Now $R(Z^i)$ is simply the same function applied to the swapped data Z^i instead of Z—that is, the residuals are computed after swapping data points i and $n+1$ in the data set. (We also later generalize to any outcome space $\mathcal{Y}$ and to other definitions of residuals.)

Clearly, the bound in (4) is strictly stronger than (3), because the total variation distance between any function applied to each of Z and Z^i, cannot be larger than $\mathsf{d}_{\mathsf{TV}}(Z, Z^i)$ itself—and in many cases, the bound in (4) may be substantially tighter. For example, if the data is high dimensional, with $Z_i = (X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ for large p, then the distance $\mathsf{d}_{\mathsf{TV}}(Z, Z^i)$ may be extremely large since Z and Z^i each contain $p+1$ dimensions of information about each data point. On the other hand, if we only observe the residuals (e.g., $R(Z)_i = |Y_i - \widehat{\mu}(X_i)|$ for each i), then this reveals only a one-dimensional summary of each data point; this typically reduces the distance between the two distributions, and by a considerable amount if the distribution drift occurs in features that happen to be irrelevant for prediction and are thus ignored by $\widehat{\mu}$. In Section 4.4, we will see a specific example demonstrating the potentially large gap between these two upper bounds.

2. Background and related work.

We briefly review several distribution-free prediction methods that offer guarantees under an exchangeability assumption on the data and symmetry of the underlying algorithm. We also set up notation that will be useful later in the paper.

Split conformal prediction. Split conformal prediction (Vovk, Gammerman and Shafer (2005)) (also called inductive conformal prediction) is a holdout method for constructing prediction intervals around a pre-trained model. Specifically, given a model $\widehat{\mu} : \mathcal{X} \to \mathbb{R}$ that was fitted on an initial training data set, and given n additional data points $(X_1, Y_1), \ldots, (X_n, Y_n)$ (the holdout set), we define residuals

$$R_i = \bigl|Y_i - \widehat{\mu}(X_i)\bigr|, \quad i = 1, \ldots, n,$$

and then compute the prediction interval at the new feature vector X_{n+1} as

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \bigl(\text{the } \lceil (1-\alpha)(n+1) \rceil\text{-th smallest of } R_1, \ldots, R_n\bigr).$$

Equivalently, we can write

$$\widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^n \frac{1}{n+1} \cdot \delta_{R_i} + \frac{1}{n+1} \cdot \delta_{+\infty}\right), \tag{5}$$

where $\mathsf{Q}_\tau(\cdot)$ denotes the τ-quantile of its argument,[1] and δ_a denotes the point mass at a. This method is well-known to guarantee distribution-free predictive coverage at the target level $1-\alpha$.

A drawback of the split conformal method is the loss of accuracy due to sample splitting, since the pre-trained model $\widehat{\mu}$ needs to be independent from the holdout set—in practice, if only n labeled data points are available in total, we might use $n/2$ data points for training $\widehat{\mu}$,

[1] If the quantile is not unique, then $\mathsf{Q}_\tau(\cdot)$ denotes the smallest possible τ-quantile, throughout this paper.

and then the procedure defined in (5) above would actually be run with a holdout set of size $n/2$ in place of n. In this paper, however, we will continue to write n to denote the holdout set size for the split conformal method, in order to allow for universal notation across different methods.

Full conformal prediction. To avoid the cost of data splitting, an alternative is the full conformal method (Vovk, Gammerman and Shafer (2005)), also referred to as transductive conformal prediction. Fix any regression algorithm

$$\mathcal{A} : \bigcup_{n\geq 0} (\mathcal{X} \times \mathbb{R})^n \to \{\text{measurable functions } \widehat{\mu} : \mathcal{X} \to \mathbb{R}\},$$

which maps a data set containing any number of pairs (X_i, Y_i), to a fitted regression function $\widehat{\mu}$. The algorithm $\mathcal{A}$ is required to treat data points symmetrically, that is,[2]

$$\mathcal{A}\big((x_{\pi(1)}, y_{\pi(1)}), \dots, (x_{\pi(n)}, y_{\pi(n)})\big) = \mathcal{A}\big((x_1, y_1), \dots, (x_n, y_n)\big) \tag{6}$$

for all $n \geq 1$, all permutations π on $[n] := \{1, \dots, n\}$, and all $\{(x_i, y_i)\}_{i=1,\dots,n}$. Next, for each $y \in \mathbb{R}$, let

$$\widehat{\mu}^y = \mathcal{A}\big((X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, y)\big)$$

denote the trained model, fitted to the training data together with a hypothesized test point (X_{n+1}, y), and let

$$R_i^y = \begin{cases} |Y_i - \widehat{\mu}^y(X_i)|, & i = 1, \dots, n, \\ |y - \widehat{\mu}^y(X_{n+1})|, & i = n+1. \end{cases} \tag{7}$$

The prediction set (which might or might not be an interval) for feature vector X_{n+1} is then defined as

$$\widehat{C}_n(X_{n+1}) = \left\{ y \in \mathbb{R} : R_{n+1}^y \leq \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_i^y} \right) \right\}. \tag{8}$$

The full conformal method is known to guarantee distribution-free predictive coverage at the target level $1 - \alpha$:

THEOREM 1 (Full conformal prediction (Vovk, Gammerman and Shafer (2005))). *If the data points $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ are i.i.d.* (*or more generally, exchangeable*), *and the algorithm $\mathcal{A}$ treats the input data points symmetrically as in* (6), *then the full conformal prediction set defined in* (8) *satisfies*

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

The same result holds true for split conformal.

Indeed, since split conformal can be viewed as a special case of full conformal (by considering a trivial algorithm $\mathcal{A}$ that returns the same fixed *pre-fitted* function $\widehat{\mu}$ regardless of the input data), this theorem also implies that the same coverage result holds for the split conformal method (5). For completeness, and to set up our proof strategy, we will give a succinct proof of this theorem in Section 6.1.

By avoiding data splitting, full conformal often (but not always) yields more precise prediction intervals than split conformal. This potential benefit comes at a steep computational cost, since in order to compute the prediction set (8) we need to rerun the model training algorithm $\mathcal{A}$ for each $y \in \mathbb{R}$ (or in practice, for each y in a fine grid). Luckily, in certain special

[2] If $\mathcal{A}$ is a randomized algorithm, then this equality is only required to hold in a distributional sense.

cases such as ordinary least squares, kernel ridge regression (Burnaev and Vovk (2014)), or the Lasso (Lei (2019)), the prediction set (8) can be computed more efficiently using specialized techniques.

As a compromise between the greater computational efficiency of split conformal and the greater statistical efficiency of full conformal, the jackknife+ and CV+ methods (Barber et al. (2021)) (closely related to cross-conformal prediction (Vovk (2015))) use a cross-validation-style approach for distribution-free predictive inference. For example, for jackknife+, the procedure requires fitting n leave-one-out models $\widehat{\mu}_{-i}$. Later, in Appendix B, we will give more detailed background on these methods, and provide a nonexchangeable version of the jackknife+.

General nonconformity scores. In the exchangeable setting, conformal prediction (both split and full) was initially proposed in terms of "nonconformity scores" $\widehat{S}(X_i, Y_i)$, where $\widehat{S}$ is a fitted function that measure the extent to which a data point (X_i, Y_i) is unusual relative to a training data set (Vovk, Gammerman and Shafer (2005)) (whose dependence we make implicit in the notation). For simplicity, so far we have only presented the most commonly used nonconformity score, which is the residual from the fitted model

$$\widehat{S}(X_i, Y_i) := |Y_i - \widehat{\mu}(X_i)| \tag{9}$$

(where $\widehat{\mu}$ is pre-trained for split conformal, and $\widehat{\mu} = \mathcal{A}((X_j, Y_j) : j \in [n+1])$ for full conformal). We will also present our new methods with this particular form of score. In many settings, other nonconformity scores can be more effective—for example, Kivaranovic, Johnson and Leeb (2020), Romano, Patterson and Candès (2019) propose scores based on quantile regression that often lead to tighter prediction intervals in practice. Our proposed nonexchangeable conformal prediction procedures can also be extended to allow for general nonconformity scores—we will return to this generalization in Appendix A.

Further related work. Conformal prediction was pioneered by Vladimir Vovk and various collaborators in the early 2000s; the book by Vovk, Gammerman and Shafer (2005) details their advances and remains a critical resource. The recent spurt of interest in these ideas in the field of statistics was catalyzed by Jing Lei, Larry Wasserman, and colleagues (see, e.g., Lei et al. (2018), Lei, Robins and Wasserman (2013), Lei and Wasserman (2014)). For gentle introduction and more history, we refer to the tutorials by Shafer and Vovk (2008) and Angelopoulos and Bates (2023).

Tibshirani et al. (2019) extended conformal prediction to handle nonexchangeable data under an assumption called *covariate shift*, where training and test data can have a different X distribution, but are assumed to have an identical distribution of Y given X. The data is reweighted using the likelihood ratio to compare the test and training covariate distributions (with this likelihood ratio assumed to be known or accurately approximable), coverage can be guaranteed via an argument based on a concept that they called *weighted exchangeability*.

Our current work differs from Tibshirani et al. (2019) in several fundamental ways, such that neither work subsumes the other in terms of methodology or theory. In their work, the covariate shift assumption must hold, and the aforementioned high-dimensional likelihood ratio must be known exactly or well approximated for correct coverage. Furthermore, the weights on the data points are then calculated as a function of the data point (X_i, Y_i) to compensate for the known distribution shift. In the present work, on the other hand, the weights are required to be *fixed* rather than data-dependent, and can compensate for *unknown* violations of the exchangeability assumption, as long as the violations are small (to ensure a low coverage gap). Moreover, our theory can handle nonsymmetric algorithms that treat different data points differently, and in particular, can depend on their order. Finally, and importantly, if there was actually no distribution shift, and the data happened to be exchangeable, their weighted algorithm does not have any coverage guarantee, while ours retains exact coverage.

Since its publication, the ideas and methods from Tibshirani et al. (2019) have been applied and extended in several ways. For example, Podkopaev and Ramdas (2021) demonstrate that reweighting can also deal with *label shift* (the marginal distribution of Y changes from training to test, but the conditional distribution of X given Y is assumed unchanged). Lei and Candès (2021) show how reweighting can be extended to causal inference setups for predictive inference on individual treatment effects, and Candès, Lei and Ren (2023) show how to apply these ideas in the context of censored outcomes in survival analysis. Fannjiang et al. (2022) use reweighting in a setup where the test covariate distribution is under the statistician's control. A different weighted approach is taken in Guan (2023), called "localized" conformal prediction, where the weight on data point i is determined as a function of the distance $\|X_i - X_{n+1}\|_2$, to enable predictive coverage that holds locally (in neighborhoods of X space, that is, an approximation of prediction that holds conditional on the value of X_{n+1}). Each of these works also contributes new ideas to problem-specific challenges (and differs substantially from the work proposed here, both in terms of methods and the nature of the resulting guarantees), but we omit the details for brevity.

Conformal methods have also be used for sequential tests for exchangeability of the underlying data (Vovk (2021)), and these sequential tests can form the basis of sequential algorithms for changepoint detection (Volkhonskiy et al. (2017)) or outlier detection (Bates et al. (2023)). This line of work is differs from ours in that they employ conformal prediction for detecting nonexchangeability, but do not provide algorithms or guarantees for the use of conformal methods for predictive inference on nonexchangeable data. Several other recent works propose conformal inference type methods for time series (Chernozhukov, Wüthrich and Yinchu (2018), Stankeviciute, Alaa and van der Schaar (2021), Xu and Xie (2021)), but these results require exchangeability assumptions or other distributional conditions (e.g., assuming a strongly mixing time series), while in our present work we aim to avoid these conditions.

The recent work of Gibbs and Candès (2021) takes a different approach towards handling distribution drift in an online manner. Informally, they compare the current attained coverage to the target $1 - \alpha$ level, and if the former is bigger (or smaller) than the latter, then they iteratively increase (or decrease) the nominal level α_t to employ for the next prediction. Zaffran et al. (2022) build further on this approach, allowing for adaptivity to the amount of dependence in the time series. An alternative approach is that of Cauchois, Gupta and Ali (2020), where robustness is introduced under the assumption that the test distribution is bounded in f-divergence from the distribution of the training data points.

For data that is instead drawn from a *spatial* domain, the recent work of Mao, Martin and Reich (2023) uses weighted conformal prediction with higher weights assigned to data points drawn at spatial locations near that of the test point (or, as a special case, giving a weight of 1 to the nearest neighbors of the test point, and weight 0 to all other points), but their theoretical guarantees require distributional assumptions.

Finally, we return full circle to the book of Vovk, Gammerman and Shafer (2005), which has chapters that discuss moving beyond exchangeability, for example using Mondrian conformal prediction (and its generalization, online compression models). Mondrian methods informally divide the observations into groups, and assume that the observations within each group are still exchangeable (e.g., class-conditional conformal classification). We also note the work of Dunn, Wasserman and Ramdas (2022) that studies the case of two-layer hierarchical models (like random effect models) that shares strength across groups. These works involve very different ideas from those presented in the current paper.

3. Nonexchangeable conformal prediction.

We now present our new nonexchangeable conformal prediction method, in both its split and full versions, in this section. For clarity of the exposition, we will use $|y - \widehat{\mu}(x)|$ as the score used to measure the nonconformity of a point (x, y) in the data set, as in (9), but our methods and accompanying theoretical guarantees can be extended in a straightforward way to arbitrary nonconformity scores—we give details for this extension in Appendix A.

3.1. *Robust inference through weighted quantiles.* As described above, our new methodology moves beyond the exchangeable setting by allowing both for nonexchangeable data, and for nonsymmetric algorithms. For simplicity, we will first consider only the first extension—the data points $Z_i = (X_i, Y_i)$ are no longer required to be exchangeable, but the model fitting algorithm $\mathcal{A}$ will still be assumed to be symmetric for now. The next subsection generalizes the method to allow nonsymmetric algorithms as well.

For our nonexchangeable conformal methods, we choose weights $w_1, \ldots, w_n \in [0, 1]$, with the intuition that a higher weight w_i should be assigned to a data point Z_i that is "trusted" more, that is, that we believe comes from (nearly) the same distribution as the test point Z_{n+1}. We assume the weights w_i are fixed (see Section 4.5 for further discussion on this point). For instance, if data point Z_i occurs at time i, and we are concerned about distribution drift, we might choose weights $w_1 \le \cdots \le w_n$ so that our prediction interval relies mostly on recent data points and places little weight on data from the distant past. Alternatively, in a spatial setting, if data point i is collected at a (prespecified) location L_i, then the weight w_i might be chosen as a function of the distance $\mathrm{dist}(L_i, L_{n+1})$, with the intuition that data points collected nearby in the spatial domain are more likely to have similar distributions.

We now modify the split and full conformal predictive inference methods to use weighted quantiles, rather than the original definitions where all data points are implicitly given equal weight. To simplify notation, in what follows, given $w_i \in [0, 1]$, $i = 1, \ldots, n$, we will define normalized weights

$$(10) \qquad \tilde{w}_i = \frac{w_i}{w_1 + \cdots + w_n + 1}, \quad i = 1, \ldots, n \quad \text{and} \quad \tilde{w}_{n+1} = \frac{1}{w_1 + \cdots + w_n + 1}.$$

Nonexchangeable split conformal with a symmetric algorithm. The prediction interval is given by

$$(11) \qquad \widehat{C}_n(X_{n+1}) = \widehat{\mu}(X_{n+1}) \pm \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{w}_i \cdot \delta_{R_i} + \tilde{w}_{n+1} \cdot \delta_{+\infty}\right),$$

where $R_i = |Y_i - \widehat{\mu}(X_i)|$ for the pre-trained model $\widehat{\mu}$, as before.

Nonexchangeable full conformal with a symmetric algorithm. The prediction set is given by

$$(12) \qquad \widehat{C}_n(X_{n+1}) = \left\{ y : R^y_{n+1} \le \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{R^y_i}\right)\right\},$$

where as before, we define $\widehat{\mu}^y = \mathcal{A}((X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, y))$ by running the algorithm $\mathcal{A}$ on the training data together with the hypothesized test point (X_{n+1}, y), and define R^y_i as in (7) from before.

Notice that for both methods, their original (unweighted) versions are recovered by choosing weights $w_1 = \cdots = w_n = 1$.

The theoretical results for this section, which we previewed in (3) and (4), will follow as a corollary of more general results that also accommodate nonsymmetric algorithms (introduced next); we avoid restating the results here for brevity. In addition, the interested reader may already jump forward to Appendix C to examine a different style of result on the robustness of weighted (and unweighted) conformal methods—using symmetric algorithms—under a Huber-style adversarial contamination model (which relies on stronger assumptions to allow for a tighter guarantee).

3.2. *Enhanced predictions with nonsymmetric algorithms.* Now, we will allow the algorithm $\mathcal{A}$ to be an arbitrary function of the data points, removing the requirement of a symmetric algorithm. This generalization will require only a small modification to the previous conformal method to ensure validity, and can result in more accurate predictors and boost efficiency of the resulting prediction sets, as we will demonstrate in the experiments (Section 5).

To begin, let us give some examples of algorithms that do not treat data points symmetrically, to see what types of settings we want to handle:

- *Weighted regression.* The algorithm $\mathcal{A}$ might fit a model $\widehat{\mu}(x) = x^\top \widehat{\beta}$ where the parameter vector $\widehat{\beta}$ is fitted via a weighted regression. Specifically, for nonnegative weights t_i, consider solving

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_i t_i \cdot \ell\left(X_i^\top \beta, Y_i\right) + h(\beta) \right\}, \tag{13}$$

  for some loss function ℓ and penalty function h. For example, weighted least squares would be obtained by taking the loss function $\ell(u, y) = (u - y)^2$.
- *Adapting to changepoints.* In a streaming data setting, if sudden changes may occur in the data distribution, then the quality of our predictions will suffer if our models are always trained on the full set of available training data without accounting for possible changepoints. We might therefore aim to improve the model by building in a changepoint detection step. Assume data points arrive in an ordered fashion so that $i = 1$ is the first arrival, $i = 2$ the second, and so on. Then, we might have

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i > \widehat{T}} \ell\left(X_i^\top \beta, Y_i\right) + h(\beta) \right\}, \tag{14}$$

  for some loss function ℓ and penalty function h, where $\widehat{T}$ is the time of the most recent detected changepoint (or $\widehat{T} = 0$ if no changepoint is detected). To be clear, here the algorithm $\mathcal{A}$ incorporates estimation of both $\widehat{T}$ and of $\widehat{\beta}$.
- *Autoregressive models.* Suppose that the response Y_{n+1} is best predicted by combining information from the features X_{n+1} together with response Y_n from the previous time point—for example, we might solve for

$$(\widehat{\beta}, \widehat{\gamma}) = \arg\min_{(\beta, \gamma) \in \mathbb{R}^p \times \mathbb{R}} \left\{ \sum_i \left(Y_i - \left(X_i^\top \beta + \gamma Y_{i-1}\right)\right)^2 \right\}, \tag{15}$$

  to return a fitted function of the form $\widehat{\mu}(x, y_{\mathrm{prev}}) := x^\top \widehat{\beta} + \widehat{\gamma} \cdot y_{\mathrm{prev}}$.

To accommodate these and many other settings, we will now define $\mathcal{A}$ as

$$\mathcal{A} : \bigcup_{n \geq 0} (\mathcal{X} \times \mathbb{R} \times \mathcal{T})^n \to \{\text{measurable functions } \widehat{\mu} : \mathcal{X} \to \mathbb{R}\}, \tag{16}$$

mapping a data sequence containing any number of "tagged" data points $(X_i, Y_i, t_i) \in \mathcal{X} \times \mathbb{R} \times \mathcal{T}$, to a fitted regression function $\widehat{\mu}$. The tag t_i associated with data point (X_i, Y_i) can play a variety of different roles, depending on the application:

- t_i can provide the weight for data point i in a weighted regression;
- t_i can indicate the time or spatial location at which data point i is sampled;
- t_i can simply indicate the order of the data points (i.e., setting $t_i = i$ for each i), so that $\mathcal{A}$ is "aware" that data point (X_i, Y_i) is the ith data point, and is thus able to use the ordering of the data points when fitting the model.

In particular, the algorithm $\mathcal{A}$ is no longer required to treat the input data points (X_i, Y_i) symmetrically, because if we swap (X_i, Y_i) with (X_j, Y_j) (and the algorithm receives tagged data points (X_j, Y_j, t_i) and (X_i, Y_i, t_j)), the fitted model may indeed change.[3] As for the weights w_i, we require the tags $t_1, \dots, t_{n+1}$ to be fixed.

With the added flexibility of a nonsymmetric regression algorithm, we will need a key modification to the methods defined earlier in Section 3.1 to maintain predictive coverage. Our modification requires that, before applying the model fitting algorithm $\mathcal{A}$, we first randomly swap the tags of two of the data points in the ordering. First, draw a random index $K \in [n+1]$ from the multinomial distribution that takes the value i with probability $\tilde{w}_i$ (defined in (10)):

$$K \sim \sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_i. \tag{17}$$

Note that K is drawn independently from the data. We will apply our algorithm to the data Z^K (defined in (2)) in place of Z. In particular, the tagged data points are now (X_{n+1}, Y_{n+1}, t_K) and (X_K, Y_K, t_{n+1}), that is, these two data points have swapped tags. This modification is carried out as follows.

Nonexchangeable split conformal with a nonsymmetric algorithm. For split conformal, the model $\widehat{\mu}$ is pre-fitted on separate data, and does not depend on the data points (X_i, Y_i) of the holdout set—in other words, $\widehat{\mu}$ is trivially a symmetric function of the (X_i, Y_i) points. Thus, no modification is needed, and our prediction interval (11) is unaltered.

Nonexchangeable full conformal with a nonsymmetric algorithm. First, for any $y \in \mathbb{R}$ and any $k \in [n+1]$, define

$$\widehat{\mu}^{y,k} = \mathcal{A}\big((X_{\pi_k(i)}, Y^y_{\pi_k(i)}, t_i) : i \in [n+1]\big),$$

where π_k is the permutation on $[n+1]$ swapping indices k and $n+1$ (and π_{n+1} is the identity permutation), and where we define

$$Y_i^y = \begin{cases} Y_i, & i = 1, \dots, n, \\ y, & i = n+1. \end{cases}$$

In other words, $\widehat{\mu}^{y,k}$ is fitted by applying the algorithm $\mathcal{A}$ to the training data $(X_1, Y_1), \dots, (X_n, Y_n)$ together with the hypothesized test point (X_{n+1}, y), but with the kth and $(n+1)$st data points swapped (note that the tags t_k and t_{n+1} are now assigned to data points (X_{n+1}, y) and (X_k, Y_k), respectively, after this swap).

Define the residuals from this model,

$$R_i^{y,k} = \begin{cases} |Y_i - \widehat{\mu}^{y,k}(X_i)|, & i = 1, \dots, n, \\ |y - \widehat{\mu}^{y,k}(X_{n+1})|, & i = n+1. \end{cases}$$

Then, after drawing a random index K as in (17), the prediction set is given by

$$\widehat{C}_n(X_{n+1}) = \left\{ y : R_{n+1}^{y,K} \le \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{R_i^{y,K}} \right) \right\}. \tag{18}$$

Symmetric algorithms as a special case. The symmetric setting, discussed in Section 3.1, is actually a special case of the broader setting defined here. Specifically, for any symmetric algorithm $\mathcal{A}$ that acts on (untagged) data points (x_i, y_i), we can trivially regard it as an

[3] For many common examples, the algorithm $\mathcal{A}$ will instead be symmetric as a function of the *tagged* data points (X_i, Y_i, t_i), but we do not require this assumption in this work.

algorithm $\mathcal{A}'$ that acts on tagged data points (x_i, y_i, t_i) by simply ignoring the tags. For this reason, we will only give theoretical results for the general forms of the methods given in this section, but our theorems apply also to the symmetric setting considered in Section 3.1.

The swap step. In the case of a nonsymmetric algorithm, our swap step requires that $\mathcal{A}$ is run on the swapped data set—that is, with data points (X_{n+1}, Y_{n+1}, t_K) and (X_K, Y_K, t_{n+1}), rather than the original tagged data points (X_K, Y_K, t_K) and $(X_{n+1}, Y_{n+1}, t_{n+1})$. This swap step is necessary for our theoretical guarantees to hold (in fact, it plays a key role in the theory even for *symmetric* algorithms, with fixed weights w_i as in (12), even though the fitted models are unchanged in that case).

Of course, for nonsymmetric algorithm $\mathcal{A}$, the swap step will alter the fitted model $\widehat{\mu}$ produced by $\mathcal{A}$ and thus may affect the precision of the resulting prediction interval. The extent to which this swap will perturb the output of the algorithm $\mathcal{A}$, will undoubtedly depend on the nature of the algorithm itself. In many practical situations, we would not expect the random swap to have a large impact on the output of the method, since many algorithms $\mathcal{A}$ applied to a large number of data points are often not very sensitive to perturbing the training set in this fashion, although interestingly, our theoretical results do not rely on any stability conditions or any assumptions of this type. (For comparison, if we were to instead permute the data at random before applying the algorithm $\mathcal{A}$—that is, use a permutation π chosen uniformly at random, rather than the single swap permutation π_K, so that the algorithm is now trained on tagged data points $(X_{\pi(i)}, Y_{\pi(i)}, t_i)$—then this would restore the symmetric algorithm assumption, but could potentially result in a highly inaccurate model since the information carried by the tags is now meaningless.)

However, in certain settings it may be the case that the fitted model $\widehat{\mu}$ returns predictions that are far less accurate due to the swap. For instance, this may be the case in an autoregressive setting where the tag t_{n+1} indicates the most recent data point and thus plays a disproportionately large role in the resulting predictions. We leave the important question of practical implementation for such settings, and the question of how to choose algorithms $\mathcal{A}$ that will not be too sensitive to the swap, to future work.

4. Theory.

In this section, we establish theory on the coverage of our proposed method. Since split conformal is a special case of full conformal (even in this nonexchangeable setting), we only present theory for the nonexchangeable full conformal method.

We first need to define the map from a data sequence $z = (z_1, \dots, z_{n+1}) \in (\mathcal{X} \times \mathbb{R})^{n+1}$, with entries $z_i = (x_i, y_i)$, to a vector of residuals $R(z)$. Given z, we first define the model

$$\widehat{\mu} = \mathcal{A}\big((x_i, y_i, t_i) : i \in [n+1]\big).$$

Then define the residual vector $R(z) \in \mathbb{R}^{n+1}$ with entries

$$\big(R(z)\big)_i = \big|y_i - \widehat{\mu}(x_i)\big|, \quad i = 1, \dots, n+1.$$

4.1. *Lower bounds on coverage.*

Recall the notation Z_i, Z, Z^i defined in (1) and (2). We now present our coverage guarantee for nonexchangeable full conformal (and consequently, the same bound holds for nonexchangeable split conformal as a special case). This theorem can be viewed as a generalization of Theorem 1.

THEOREM 2 (Nonexchangeable full conformal prediction). *Let $\mathcal{A}$ be an algorithm mapping a sequence of triplets (X_i, Y_i, t_i) to a fitted function as in* (16). *Then the nonexchangeable full conformal method defined in* (18) *satisfies*

$$\mathbb{P}\big\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\big\} \geq 1 - \alpha - \sum_{i=1}^{n} \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}\big(R(Z), R(Z^i)\big).$$

The same result holds true for nonexchangeable split conformal.

To summarize, we see that the coverage gap is bounded by $\sum_{i=1}^n \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}(R(Z), R(Z^i))$. Since it holds that

$$\mathsf{d}_{\mathsf{TV}}(R(Z), R(Z^i)) \le \mathsf{d}_{\mathsf{TV}}(Z, Z^i)$$

for each i, we therefore also see that

$$\text{Coverage gap} \le \sum_i \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}(Z, Z^i).$$

This last bound is arguably more interpretable, but could also be significantly more loose, and we consider it an important point that the coverage gap depends on the total variation between swapped residual vectors, and not the swapped raw data vectors. Finally, recalling Lemma 1, we see that in the case of independent data points, we have

$$\text{Coverage gap} \le 2 \sum_i \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}\big((X_i, Y_i), (X_{n+1}, Y_{n+1})\big).$$

4.2. *Upper bounds on coverage.* To complement the results in the last subsection, it is also possible to verify, for the nonexchangeable conformal method, that the procedure does not substantially overcover—that is, under mild deviations from exchangeability, our method is not overly conservative.

For the exchangeable setting, Lei et al. ((2018), Theorem 2.1) show that, in a setting where the residuals R_i (for split conformal) or R_i^y (for full conformal) are distinct with probability 1, conformal prediction satisfies

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

Here we give the analogous results for our nonexchangeable methods.

THEOREM 3. *For any algorithm $\mathcal{A}$ as in* (16), *if $R_1^{Y_{n+1},K}, \dots, R_n^{Y_{n+1},K}, R_{n+1}^{Y_{n+1},K}$ are distinct with probability* 1, *then the nonexchangeable full conformal method* (18) *satisfies*

$$\mathbb{P}\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\} < 1 - \alpha + \tilde{w}_{n+1} + \sum_{i=1}^n \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}(R(Z), R(Z^i)).$$

The same result holds true for nonexchangeable split conformal.

(In the split conformal context, since $R_i^{Y_{n+1},K} = |Y_i - \widehat{\mu}(X_i)|$ for the pre-fitted function $\widehat{\mu}$, the statement becomes simpler—we simply require that the residuals $|Y_i - \widehat{\mu}(X_i)|$ are distinct with probability 1.)

From this result, we see that if $\tilde{w}_{n+1} = \frac{1}{w_1 + \cdots + w_n + 1}$ is small (which corresponds to the effective sample size of our weighted method being large), then mild violations of exchangeability can only lead to mild undercoverage (as in Theorem 2) or to mild overcoverage.

Of course, when we use these methods in practice, it would be useful to know whether overcoverage or undercoverage is to be expected; however, without further assumptions, this cannot be determined in advance. As a simple example, if the data exhibits mild violations of exchangeability due to the conditional variance of $Y|X$ changing over time, then we might see undercoverage if $\mathrm{Var}(Y|X)$ increases over time (and thus the residual of the test point (X_{n+1}, Y_{n+1}) is larger than typical training residuals), or overcoverage if $\mathrm{Var}(Y|X)$ is instead decreasing over time.

4.3. *Remarks on the theorems.* A few comments are in order to help us further understand the implications of these theoretical results.

New results in the exchangeable setting. We point out that when the data happen to be exchangeable, that is, $\mathsf{d}_{\mathsf{TV}}(Z, Z^i) = 0$ for all i, then the above results are new and cannot be inferred from the existing conformal literature. In particular, existing conformal methods are not able to handle nonsymmetric algorithms, which limits their applicability in many practical settings (e.g., streaming data, as described above). In addition, our results show that, under exchangeability, there is no coverage lost by introducing fixed weights w_i into the quantile calculations used for constructing the prediction interval; this means that we are free to use these weights to help ensure robustness against nonexchangeability without sacrificing any guarantees if indeed exchangeability happens to hold.

Robustness results for the original algorithms. Another interesting implication of these new bounds is that they yield robustness results for the original algorithms. In more detail, the original split conformal (5) and full conformal (8) algorithms presented in Section 2 can be viewed as special cases of our proposed nonexchangeable methods (11) and (18), respectively, by taking weights $w_1 = \cdots = w_n = 1$ and using a symmetric $\mathcal{A}$, that is, without tags. (As we will see in Appendix B below, the same is true for viewing jackknife+ as a special case of the nonexchangeable jackknife+.) In this setting, our theorems establish a new robustness result,

$$\text{Coverage gap} \le \frac{\sum_{i=1}^n \mathsf{d}_{\mathsf{TV}}(R(Z), R(Z^i))}{n+1} \le \frac{\sum_{i=1}^n \mathsf{d}_{\mathsf{TV}}(Z, Z^i)}{n+1}.$$

For example, in the case of independent data points, applying Lemma 1 we obtain

$$\text{Coverage gap} \le \frac{2\sum_{i=1}^n \mathsf{d}_{\mathsf{TV}}((X_i, Y_i), (X_{n+1}, Y_{n+1})))}{n+1}.$$

These new bounds ensure robustness of existing methods against mild violations of the exchangeability (or i.i.d.) assumption, and thus help explain the success of these methods on real data, where the exchangeability assumption may not hold.

Choosing the weights. Our theoretical results above confirm the intuition that we should give higher weights w_i to data points (X_i, Y_i) that we believe are drawn from a similar distribution as (X_{n+1}, Y_{n+1}), and lower weights to those that are less reliable. As is always the case with inference methods, we are faced with a tradeoff: if many weights w_i are chosen to be quite low, then this reduces the effective sample size of the method (e.g., for split conformal prediction, we are reducing the effective sample size for estimating the empirical quantile of the residual distribution). Thus, overly low weights will often lead to wider prediction intervals—at the extreme, if we choose $w_1 = \cdots = w_n = 0$, this yields a coverage gap of zero but results in $\widehat{C}_n(X_{n+1}) \equiv \mathbb{R}$, a completely uninformative prediction interval. How to choose weights optimally (and, even how to quantify optimality) is an interesting and important question that we leave for future work.

Is the guarantee useful? While the upper bound on the coverage gap holds with no assumptions on the distribution of the data, the result is meaningless if this upper bound is extremely large. Thus, we would ideally use these methods in settings where we have some *a priori* knowledge about the properties of the data distribution, so that the weights w_i can be chosen in advance in such a way that we believe the resulting coverage gap is likely to be small. We emphasize in practice we likely only need qualitative (not quantitative) knowledge of the likely deviations from exchangeability—for example, under gradual distribution drift, a geometric decay as in $w_i = \rho^{n+1-i}$ will likely lead to a low coverage gap, without requiring knowledge of the exact rate or nature of the distribution drift. On the other hand, if the test point comes from a new distribution that bears no resemblance to the training data, neither our bound nor any other method would be able to guarantee valid coverage without further assumptions. An important open question is whether it may be possible to determine, in an adaptive way, whether coverage will likely hold for a particular data set, or whether that data set exhibits high deviations from exchangeability such that the coverage gap may be large.

4.4. *Examples.* Before turning to our empirical results, we pause to give several examples of settings where the coverage gap bound is favorable.

Bounded distribution drift. First, consider a setting where the data points (X_i, Y_i) are independent, but experience distribution drift over time. In this type of setting, we would want to choose weights w_i that decay as we move into the distant past, for example, $w_i = \rho^{n+1-i}$ for some decay parameter $\rho \in (0, 1)$. If we assume that the distribution drift is bounded with a Lipschitz-type condition,

$$\mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1}) \leq \epsilon \cdot (n+1-i), \quad i = 1, \dots, n+1,$$

for some $\epsilon > 0$, then the coverage gap for our proposed methods is bounded as

$$\begin{aligned}\text{Coverage gap} &\leq \sum_i \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}\big(Z, Z^i\big) \leq \sum_i \tilde{w}_i \cdot 2\mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1}) \\ &\leq \sum_{i=1}^{n} \frac{\rho^{n+1-i}}{1 + \sum_{j=1}^{n} \rho^{n+1-j}} \cdot 2\epsilon \cdot (n+1-i) \leq \frac{2\epsilon}{1-\rho},\end{aligned}$$

which is small as long as the distribution drift parameter ϵ is sufficiently small.

Changepoints. In other settings with independent data points (X_i, Y_i), we might have periodic large changes in the distribution rather than the gradual drift studied above—that is, we may be faced with a changepoint. Suppose that the most recent changepoint occurred k time steps ago, so that $\mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1}) = 0$ for $i > n-k$ (but, before that time, the distribution might be arbitrarily different from the test point, so we might even have $\mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1}) = 1$ for $i \leq n-k$). In this setting, again taking weights $w_i = \rho^{n+1-i}$ that decay as we move into the past, we have

$$\text{Coverage gap} \leq \sum_{i=1}^{n} \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}\big(Z, Z^i\big) \leq \sum_{i=1}^{n-k} \tilde{w}_i = \frac{\sum_{i=1}^{n-k} \rho^{n+1-i}}{1 + \sum_{i=1}^{n} \rho^{n+1-i}} \leq \rho^k.$$

This yields a small coverage gap as long as k is large, that is, as long as we have plenty of data observed after the most recent changepoint.

Covariate time series. Next, to highlight the distinction between $\mathsf{d}_{\mathsf{TV}}(Z, Z^i)$ and $\mathsf{d}_{\mathsf{TV}}(R(Z), R(Z^i))$, we will consider a setting where the data points (X_i, Y_i) are no longer independent. Suppose that $Y_i = X_i^\top \beta + \epsilon_i$ where $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ but where the covariates X_i are not i.i.d. For example, the covariates may be dependent due to a time series structure, or may be independent but not identically distributed. Writing $X \in \mathbb{R}^{(n+1)\times p}$ to denote the covariate matrix (with ith row X_i), we will assume that $\text{vec}(X) \sim \mathcal{N}(0, \Sigma)$ for some $\Sigma \in \mathbb{R}^{(n+1)p \times (n+1)p}$, allowing for both nonindependent and/or nonidentically distributed rows X_i. Now consider running full conformal with least squares regression as the base algorithm, so that we have residuals

$$R(Z) = Y - X\big(X^\top X\big)^{-1} X^\top Y = \mathcal{P}_X^\perp(Y) = \mathcal{P}_X^\perp(\epsilon),$$

where $Y = (Y_1, \dots, Y_{n+1})$, $\epsilon = (\epsilon_1, \dots, \epsilon_{n+1})$, and $\mathcal{P}_X^\perp$ denotes projection to the orthogonal complement of the column span of X. In the Supplementary Material (Barber et al. (2023)), we prove that

$$\mathsf{d}_{\mathsf{TV}}\big(R(Z), R\big(Z^i\big)\big) \leq \sqrt{8}\kappa_\Sigma \cdot \frac{p}{\sqrt{n+1-p}}, \tag{19}$$

where κ_Σ is the condition number of Σ; if $n \gg p^2$ while κ_Σ is bounded, then this total variation distance is very small.

On the other hand, it is likely that $\mathsf{d}_{\mathsf{TV}}(Z, Z^i)$ is very large (it may even be close to the largest possible value of 1), unless the covariates are essentially exchangeable. For example, in dimension $p = 1$, we can consider the autoregressive model $X_i = \gamma X_{i-1} + \mathcal{N}(0, 1-\gamma^2)$, with $X_1 \sim \mathcal{N}(0, 1)$, so $X_1, \dots, X_{n+1}$ are identically distributed. Then, for $2 \le i \le n$ we have

$$\begin{aligned}\mathsf{d}_{\mathsf{TV}}(Z, Z^i) &\ge \mathsf{d}_{\mathsf{TV}}(X_i - \gamma X_{i-1}, X_{n+1} - \gamma X_{i-1}) \\ &= \mathsf{d}_{\mathsf{TV}}\big(\mathcal{N}(0, 1-\gamma^2), \mathcal{N}(0, 1+\gamma^2 - 2\gamma^{n+3-i})\big),\end{aligned}$$

which is proportional to γ^2. This shows that $\mathsf{d}_{\mathsf{TV}}(R(Z), R(Z^i))$ can be vanishingly small even when $\mathsf{d}_{\mathsf{TV}}(Z, Z^i)$ is bounded away from zero.

4.5. *Extensions and explorations.* We now briefly describe several extensions of our general framework.

Additive versus multiplicative bounds. In our theoretical results above, the reduction in coverage is additive—that is, the probability $\mathbb{P}\{Y_{n+1} \notin \widehat{C}_n(X_{n+1})\}$ has the form $\alpha + \Delta$, where the term Δ reflects the extent to which the exchangeability assumption is violated (as measured by total variation distance). If the target noncoverage level α is extremely low, then this additive bound may represent a substantial increase in the probability of error. In Appendix C, we give an alternative bound under a Huber contamination model, which is multiplicative rather than additive, but holds only for the symmetric algorithm case.

Fixed versus data-dependent weights. Throughout this paper, we have worked under the assumption that the weights w_i on the conformal residuals, as well as the tags t_i used in model fitting in the nonsymmetric case, are fixed a priori. In contrast, when weighted versions of conformal prediction are used for addressing problems such as covariate shift (Tibshirani et al. (2019)), data censoring (Candès, Lei and Ren (2023)), or local coverage (Guan (2023)), the weights are data-dependent, that is, $w_i = w(X_i)$ or $w_i = w(X_i, X_{n+1})$, in each of these settings. We pause here to comment on this distinction.

In practical applications of our proposed methods, it may be the case that we would like to use weights and/or tags that are somehow random—for example, if each data point (X_i, Y_i) is gathered at a random time T_i, the weight w_i and tag t_i might then need to depend on T_i. In the setting where the weights w_i and/or tags t_i may be random or data-dependent, our results will still apply if the terms $\mathsf{d}_{\mathsf{TV}}(Z, Z^i)$ appearing in our bounds on the coverage gap are replaced with suitable conditional versions,

$$\text{Coverage gap} \le \mathbb{E}\left[\sum_{i=1}^n \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}(Z, Z_i | w_1, \dots, w_n, t_1, \dots, t_{n+1})\right],$$

where now the ith term on the right-hand side is the total variation distance between the *conditional* distributions of Z and Z^i, conditioning on the weights and tags. We might therefore consider a possible extension that unifies the proposed framework with the weighted conformal prediction (Tibshirani et al. (2019)) and/or localized conformal prediction (LCP) (Guan (2023)) methods, where the weight w_i (and, potentially, the tag t_i) placed on data point i might now additionally incorporate data-dependent information—for example, the weight w_i might depend on both the index i, as in our framework, and on $\|X_i - X_{n+1}\|_2$, as in the LCP framework. We leave a more detailed investigation of data dependent weights for future work.

Are these results assuming the data is approximately exchangeable? Finally, we point out that these coverage gap bounds are very different in flavor than simply assuming that Z is "nearly exchangeable." In particular, in a setting where $\mathsf{d}_{\mathsf{TV}}(Z, \tilde{Z})$ is small for some exchangeable $\tilde{Z}$, it follows immediately that the coverage gap is bounded by $\mathsf{d}_{\mathsf{TV}}(Z, \tilde{Z})$ for (unweighted) split or full conformal, since these methods are guaranteed to have coverage $1-\alpha$ with exchangeable data $\tilde{Z}$. By comparison, our coverage gap bound $\sum_i \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}(Z, Z^i)$ is substantially stronger.

To see this through an example, consider a distribution where the covariates X_i are i.i.d., and where $Y_i \sim \mathrm{Bernoulli}(0.5 + (-1)^i \cdot \epsilon)$, for some small constant $\epsilon > 0$. Suppose that we run conformal prediction without weights, $w_i \equiv 1$. Then we have $\mathsf{d}_{\mathsf{TV}}(Z_i, Z_{n+1}) \leq 2\epsilon$ for all i, and so our coverage gap bound ensures that conformal prediction has coverage at least $1-\alpha-4\epsilon$. On the other hand, we have

$$\mathsf{d}_{\mathsf{TV}}(Z, \tilde{Z}) \approx 1 \quad \text{for any exchangeable } \tilde{Z}.$$

To verify the above claim, note that under the distribution of Z, we have

$$\sum_{i=1}^{\lfloor \frac{n+1}{2} \rfloor} \mathbb{1}\{Y_{2i-1} < Y_{2i}\} + B_i \mathbb{1}\{Y_{2i-1} = Y_{2i}\} \sim \mathrm{Binomial}\left(\left\lfloor \frac{n+1}{2} \right\rfloor, 0.5 + \epsilon\right),$$

where $B_i \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(0.5)$, while under any exchangeable distribution, the left-hand side above is distributed as $\mathrm{Binomial}(\lfloor \frac{n+1}{2} \rfloor, 0.5)$, and these two binomial distributions have total variation distance ≈ 1, for large n. Thus, in this example, we see that our coverage gap is low even though it is not the case that Z is "nearly exchangeable."

5. Experiments. In this section, we examine the empirical performance of nonexchangeable full conformal prediction, with residual weights and allowing for a nonsymmetric algorithm, against the original full conformal method. (Additional experiments that implement split conformal and jackknife+ can be found in the Supplementary Material.) We will see that adding weights enables robustness against changes in the data distribution (i.e., better coverage), while moving to a nonsymmetric algorithm enables shorter prediction intervals.[4]

5.1. *Simulations.* We consider three simulated data distributions:

- *Setting 1: i.i.d. data.* We generate $N = 2000$ i.i.d. data points (X_i, Y_i), with $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_4)$ and $Y_i \sim X_i^\top \beta + \mathcal{N}(0,1)$ for a coefficient vector $\beta = (2, 1, 0, 0)$.
- *Setting 2: changepoints.* We generate $N = 2000$ data points (X_i, Y_i), with $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_4)$ and $Y_i \sim X_i^\top \beta^{(i)} + \mathcal{N}(0,1)$. Here $\beta^{(i)}$ is the coefficient vector at time i, and changes two times over the duration of data collection:

$$\beta^{(1)} = \cdots = \beta^{(500)} = (2, 1, 0, 0),$$
$$\beta^{(501)} = \cdots = \beta^{(1500)} = (0, -2, -1, 0),$$
$$\beta^{(1501)} = \cdots = \beta^{(2000)} = (0, 0, 2, 1).$$

- *Setting 3: distribution drift.* We generate $N = 2000$ data points (X_i, Y_i), with $X_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_4)$ and $Y_i \sim X_i^\top \beta^{(i)} + \mathcal{N}(0,1)$. As before, $\beta^{(i)}$ is the coefficient vector at time i; but now we set $\beta^{(1)} = (2, 1, 0, 0)$, $\beta^{(N)} = (0, 0, 2, 1)$, and then compute each intermediate $\beta^{(i)}$ by linear interpolation.

[4] Code for reproducing the experiments in Sections 5.1 and 5.2 is available at https://rinafb.github.io/code/nonexchangeable_conformal.zip.

For each task, we implement the following three methods, with target coverage level $1 - \alpha = 0.9$.

- *CP+LS: full conformal prediction with least squares.* We consider the original definition of full conformal prediction (8), with $\widehat{\mu}$ the least squares fit, that is, $\mathcal{A}$ is the least squares regression algorithm.[5]
- *NexCP+LS: nonexchangeable full conformal with least squares.* We also run nonexchangeable full conformal prediction (12) using weights $w_i = 0.99^{n+1-i}$, and with the same algorithm $\mathcal{A}$ (least squares regression).
- *NexCP+WLS: nonexchangeable full conformal with weighted least squares.* Lastly, we use nonexchangeable full conformal prediction (12) but now with a nonsymmetric algorithm, weighted least squares regression. Specifically, to fit $\widehat{\mu}$ given tagged data points (x_i, y_i, t_i), the algorithm $\mathcal{A}$ will run weighted least squares regression placing weight t_i on data point (x_i, y_i). We implement the algorithm with $t_i = 0.99^{n+1-i}$, and again use weights $w_i = 0.99^{n+1-i}$.

After a burn-in period of the first 100 time points, at each time $n = 100, \dots, N - 1$ we run the methods with training data $i = 1, \dots, n$ and test point $n + 1$. The results shown are averaged over 200 independent replications of the simulation.

Our results are shown in Figure 2, and are summarized in Table 1. In terms of coverage, we see that all three methods have coverage $\approx 90\%$ across the time range of the experiment for the i.i.d. data setting (Setting 1), while for the changepoint (Setting 2) and distribution drift (Setting 3) experiments, the two proposed methods achieve approximately the desired coverage level, but the original full conformal method CP+LS undercovers. In particular, as expected, CP+LS shows steep drops in coverage in Setting 2 after changepoints, while in Setting 3 the coverage for CP+LS declines gradually over time as the distribution drift grows. The NexCP+LS and NexCP+WLS methods are better able to maintain coverage in these settings. (In fact, in Setting 2, we see that NexCP+WLS overcovers for a period of time after each changepoint—this is because, a short period of time after the changepoint, the fitted weighted least squares model is already quite accurate for the new data distribution, but the weights $\tilde{w}_i$ are still placing some weight on residuals from data points from before the changepoint, leading briefly to an overestimate of our model error.)

Turning to the prediction interval width, for the i.i.d. data setting (Setting 1), the three methods show similar mean widths, although the widths for NexCP+LS and NexCP+WLS are very slightly higher than for CP+LS; in addition, variability is higher for NexCP+LS and NexCP+WLS than for CP+LS, which is to be expected since using decaying weights w_i for computing the prediction intervals leads to a lower effective sample size. For the changepoint (Setting 2) and distribution drift (Setting 3) experiments, we see that NexCP+LS leads to wider prediction intervals than the original method CP+LS, which is to be expected since NexCP+LS is using the same model fitting algorithm but avoiding the undercoverage issue of CP+LS. More importantly, NexCP+WLS is able to construct narrower prediction intervals than CP+LS, while avoiding undercoverage. This is due to the fact that weighted least squares leads to more accurate fitted models. This highlights the utility of nonsymmetric algorithms for settings where data are not exchangeable.

[5] In principle, full conformal run with $\mathcal{A}$ given by least squares may return a prediction set $\widehat{C}(X_{n+1})$ that is a disjoint union of intervals, but this is rare in most typical settings. The same is true for NexCP and for weighted least squares. For interpretability, we implement each method to always return an interval (i.e., if $\widehat{C}(X_{n+1})$ happens to be a disjoint union of intervals, then we return its convex hull).

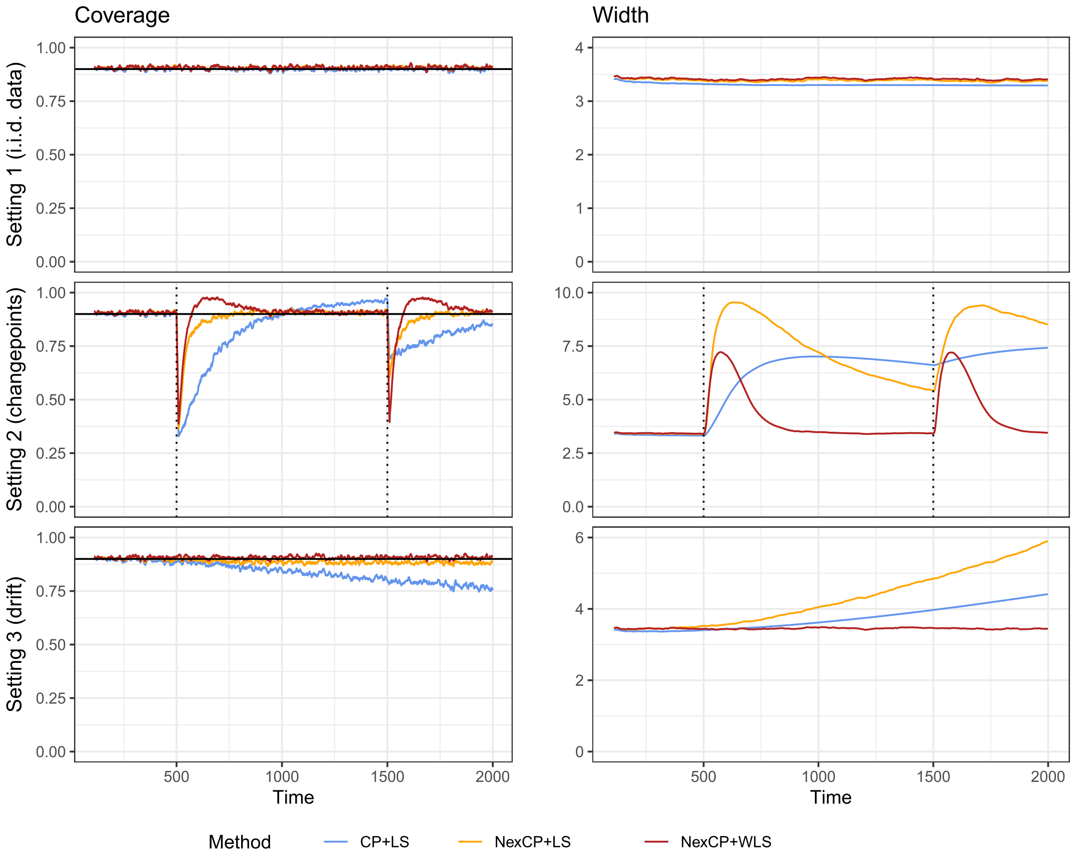

FIG. 2. *Simulation results showing mean prediction interval coverage and width, averaged over* 200 *independent trials. The displayed curves are smoothed by taking a rolling average with a window of* 10 *time points.*

5.2. *Electricity data set.* We now compare the three methods on a real data set. The ELEC2 data set[6] (Harries (1999)) tracks electricity usage and pricing in the states of New South Wales and Victoria in Australia, every 30 minutes over a 2.5 year period in 1996–1999. (This data set was previously analyzed by Vovk, Petej and Gammerman (2021) in the context of conformal prediction, finding distribution drift that violated exchangeability.)

For our experiment, we use four covariates: `nswprice` and `vicprice`, the price of electricity in each of the two states, and `nswdemand` and `vicdemand`, the usage demand in each of the two states. Our response variable is `transfer`, the quantity of electricity transferred between the two states. We work with a subset of the data, keeping only those

TABLE 1
Simulation results showing mean prediction interval coverage and width, averaged over all time points and over 200 *trials*

	Setting 1 (i.i.d. data)		Setting 2 (changepoints)		Setting 3 (drift)	
	Coverage	Width	Coverage	Width	Coverage	Width
CP+LS	0.900	3.31	0.835	5.99	0.838	3.73
NexCP+LS	0.907	3.39	0.884	6.83	0.888	4.29
NexCP+WLS	0.907	3.42	0.906	4.13	0.907	3.45

[6] Data was obtained from https://www.kaggle.com/yashsharan/the-elec2-dataset.

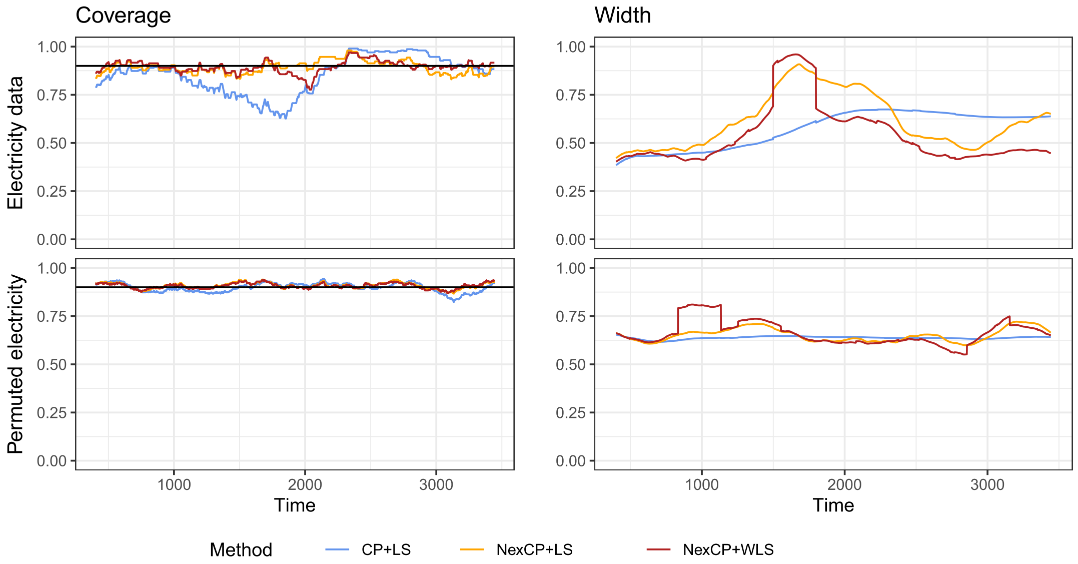

FIG. 3. *Electricity data results showing coverage and prediction interval width on the original data and the permuted data. The displayed curves are smoothed by taking a rolling average with a window of* 300 *time points.*

observations in the time range 9:00am–12:00pm (aiming to remove daily fluctuation effects), and discarding an initial stretch of time during which the value `transfer` is constant. After these steps, we have $N = 3444$ time points. We then implement the same three methods as in the simulations (CP+LS, NexCP+LS, and NexCP+WLS), using the exact same definitions and settings as before.

Our goal is to examine how distribution drift over the duration of this 2.5 year period will affect each of the three methods. As a sort of "control group," we also perform the experiment with a permuted version of this same data set—we draw a permutation π on $[N]$ uniformly at random, and then repeat the same experiment on the permuted data set $(X_{\pi(1)}, Y_{\pi(1)}), \dots, (X_{\pi(N)}, Y_{\pi(N)})$. The random permutation ensures that the distribution of this data set now satisfies exchangeability.

Our results are shown in Figure 3, and summarized in Table 2. On the original data set, we see that the unweighted method CP+LS shows some undercoverage, while both NexCP+LS and NexCP+WLS achieve nearly the desired 90% coverage level. In particular, CP+LS shows undercoverage during a long range of time around the middle of the duration of the experiment, and then recovers, showing the effects of distribution drift in this data set—this occurs as the response variable `transfer` is more noisy during the middle of the time range, as compared to the beginning and end of the time range. On the permuted data set, on the other

TABLE 2
Electricity data results showing coverage and prediction interval width on the original data and the permuted data, averaged over all time points

	Electricity data		Permuted electricity data	
	Coverage	Width	Coverage	Width
CP+LS	0.852	0.565	0.899	0.639
NexCP+LS	0.890	0.606	0.908	0.652
NexCP+WLS	0.893	0.527	0.908	0.663

hand, all three methods show coverage that is close to 90% throughout the time range, which is expected since the permuted data set is exchangeable.

Turning now to prediction interval width, on the original data set we see that the interval width of NexCP+LS is generally larger than that of NexCP+WLS, again demonstrating the advantage of a nonsymmetric algorithm. For the permuted data set, on the other hand, the interval widths are similar, although NexCP+LS and NexCP+WLS show higher variability; this is explained by the lower effective sample size that is introduced by weighting the data points, combined with the heavy-tailed nature of the data.

5.3. *Election data set.* Finally, we apply our weighted methods to predict how Americans voted in the 2020 U.S. presidential election. Our experiments in this subsection are inspired by the work of Cherian and Bronner (2020) for The Washington Post.

The left map in Figure 4 shows, county by county, the relative change in the number of votes for the Democratic Candidate between 2016 and 2020, defined as

$$Y = \frac{\mathrm{Dem}_{2020} - \mathrm{Dem}_{2016}}{\mathrm{Dem}_{2016}},$$

where Dem_{2020} is the number of Democratic votes in a given county in 2020 (and similarly for 2016). In our experiments, the covariate vector X includes information on the makeup of the county population by ethnicity, age, sex, median income and education (see the Supplementary Material for details and for information about the data sources), given the data that was available in 2020.

During real-time election forecasting, after observing the response Y for a subset of the counties (those counties that have reported), the problem is to predict the vote change Y in each of the counties where vote counts are not yet available. If the order in which counties report their vote totals were drawn uniformly at random, then the exchangeability of the resulting training and test sets would mean that conformal prediction can be applied in a straightforward manner to obtain valid predictive intervals for the unobserved counties. In practice, however, the time at which a county reports its votes may depend on various factor such as the time zone of the county, the size of the county, and so on. Therefore, if at any point in time we were to train on counties whose votes have already been reported, then this can create a division of training and test sets that violates exchangeability, and can thus lead to a failure of the predictive coverage guarantee.

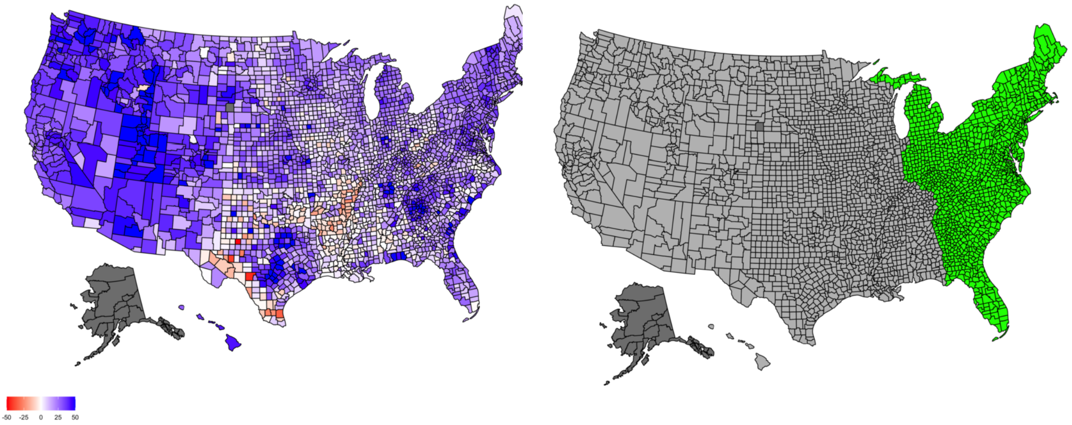

FIG. 4. *Left map: relative change in the number of votes for the Democratic presidential candidate from* 2016 *to* 2020. *Blue colors indicate an increase in Democratic votes and red colors indicate a decrease, with the darkest shade of blue (respectively, red) corresponding to a* 50% *increase (respectively, decrease). Right map: the counties that form the training set (in green), and all remaining counties are in the test set.*

TABLE 3
Election data results showing coverage, averaged over all test counties

	Coverage		Coverage
CP+LS	0.743	CP+QR	0.782
NexCP+LS	0.820	NexCP+QR	0.836
NexCP+WLS	0.840	NexCP+WQR	0.835

To mimic this type of biased split, for the current experiment, we use counties that fall under the Eastern time zone as our training set, and the remaining counties as the test set, as highlighted in the right-hand map in Figure 4. This results in 1119 training points and 1957 test points.

To run the experiment, we implement the same three full conformal methods as before (CP+LS, NexCP+LS, and NexCP+WLS). To define weights w_i for NexCP, we will use some available side information—namely, $X^{\text{prev}} \in \mathbb{R}^p$, which gives the 2016 measurements for the same set of demographic and socioeconomic variables as contained in X for 2020. The weights are then defined as $w_i = e^{-\gamma \|X_i^{\text{prev}} - X_{n+1}^{\text{prev}}\|_2}$, where we choose γ to satisfy

$$\frac{(\sum_{i=1}^n w_i + 1)^2}{\sum_{i=1}^n w_i^2 + 1} = 100,$$

essentially corresponding to an effective training sample size of 100 once we use the weighted training sample. We note that, since these weights depend only on data from 2016, we can treat these weights as fixed (i.e., these weights were determined "earlier" than gathering the data set $\{(X_i, Y_i)\}_{i=1}^{3076}$ in 2020). By using these weights within nonexchangeable conformal, we are implicitly invoking a hypothesis that counties which had similar demographics in 2016 will generate approximately exchangeable data in 2020. Finally, for NexCP+WLS, we use the same choice for the tags used for running the weighted least squares regression, that is, setting $t_i = w_i$.

In addition, we also repeat the entire experiment with quantile regression in place of linear regression, and use a corresponding choice of the nonconformity score function—specifically, after fitting a lower 5% percentile function $\hat{q}_{0.05}(\cdot)$, and an upper 95% percentile function $\hat{q}_{0.95}(\cdot)$ to the data, the nonconformity score is given by $\widehat{S}(X_i, Y_i) = \max\{\hat{q}_{0.05}(X_i) - Y_i, Y_i - \hat{q}_{0.95}(X_i)\}$, as in Romano, Patterson and Candès (2019). This yields three additional methods: conformal prediction with quantile regression (CP+QR), nonexchangeable conformal with quantile regression (NexCP+QR), and nonexchangeable conformal with weighted quantile regression (NexCP+WQR), where the weights w_i and the tags t_i are defined the same way as in linear regression.

Table 3 shows the resulting predictive coverage, averaged over the test set, for each of the three methods, when they are run with target coverage level $1 - \alpha = 0.9$. We can see that CP undercovers substantially, particularly when combined with least squares, due to the construction of nonexchangeable training and test counties. In contrast, NexCP (with or without the nonsymmetric algorithm) is able to achieve a coverage level that is much closer to the target level 90%.

6. Proofs.

In this section, we give proofs of all theorems presented so far.

6.1. *Background*: *Proof of Theorem* 1.

To help build intuition for the proof techniques we will use later on, we reformulate Vovk, Gammerman and Shafer (2005)'s proofs of these results, and we then explain some of the challenges in extending these existing results to our new setting.

Let $R_i = R_i^{Y_{n+1}}$ denote the ith residual, at the hypothesized value $y = Y_{n+1}$. By our assumptions, the data points $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$ are i.i.d. (or exchangeable), and the fitted model $\widehat{\mu} = \widehat{\mu}^{Y_{n+1}} = \mathcal{A}((X_1, Y_1), \dots, (X_{n+1}, Y_{n+1}))$ is constructed via an algorithm $\mathcal{A}$ that treats these $n+1$ data points symmetrically. The residuals $R_i = |Y_i - \widehat{\mu}(X_i)|$ are thus exchangeable.

Now define the set of "strange" points

$$\mathcal{S}(R) = \left\{ i \in [n+1] : R_i > \mathsf{Q}_{1-\alpha}\left(\sum_{j=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_j} \right) \right\}.$$

That is, an index i corresponds to a "strange" point if its residual R_i is one of the $\lfloor \alpha(n+1) \rfloor$ largest elements of the list $R_1, \dots, R_{n+1}$. By definition, this can include at most $\alpha(n+1)$ entries of the list, that is,

$$\left| \mathcal{S}(R) \right| \leq \alpha(n+1).$$

Next, by definition of the full conformal prediction set, we see that $Y_{n+1} \notin \widehat{C}_n(X_{n+1})$ (i.e., coverage fails) if and only if $R_{n+1} > \mathsf{Q}_{1-\alpha}(\sum_{i=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_i})$, or equivalently, if and only if the test point $n+1$ is "strange," that is, $n+1 \in \mathcal{S}(R)$. Therefore, we have

$$\begin{aligned} \mathbb{P}\{Y_{n+1} \notin \widehat{C}_n(X_{n+1})\} = \mathbb{P}\{n+1 \in \mathcal{S}(R)\} &= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{P}\{i \in \mathcal{S}(R)\} \\ &= \frac{1}{n+1} \mathbb{E}\left[\sum_{i=1}^{n+1} \mathbb{1}\{i \in \mathcal{S}(R)\} \right] = \frac{1}{n+1} \mathbb{E}\left[\left| \mathcal{S}(R) \right| \right] \\ &\leq \frac{1}{n+1} \cdot \alpha(n+1) = \alpha, \end{aligned}$$

where the second equality holds due to the exchangeability of $R_1, \dots, R_{n+1}$.

Challenges for the new algorithms. We will now see why the above proof does not obviously extend to our nonexchangeable conformal method, even if we were to assume that the data points (X_i, Y_i) are exchangeable. First, suppose that $\mathcal{A}$ is symmetric (i.e., we do not use tags t_i). For the original full conformal prediction method, in the proof of Theorem 1, exchangeability of the data points is used to verify that $\mathbb{P}\{n+1 \in \mathcal{S}(R)\} = \mathbb{P}\{i \in \mathcal{S}(R)\}$ for each $i \in [n]$, or equivalently,

$$\mathbb{P}\left\{ R_{n+1} > \mathsf{Q}_{1-\alpha}\left(\sum_{j=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_j} \right) \right\} = \mathbb{P}\left\{ R_i > \mathsf{Q}_{1-\alpha}\left(\sum_{j=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_j} \right) \right\}.$$

This equality holds since the residuals R_i are exchangeable (by assumption on the data) and since $\mathsf{Q}_{1-\alpha}(\sum_{j=1}^{n+1} \frac{1}{n+1} \cdot \delta_{R_j})$ is a symmetric function of $R_1, \dots, R_{n+1}$. For the nonexchangeable full conformal algorithm proposed in (12), on the other hand, we would need to check whether

$$\mathbb{P}\left\{ R_{n+1} > \mathsf{Q}_{1-\alpha}\left(\sum_{j=1}^{n+1} \tilde{w}_j \cdot \delta_{R_j} \right) \right\} \overset{?}{=} \mathbb{P}\left\{ R_i > \mathsf{Q}_{1-\alpha}\left(\sum_{j=1}^{n+1} \tilde{w}_j \cdot \delta_{R_j} \right) \right\}.$$

Even when the residuals R_i are exchangeable (i.e., when the data points are exchangeable and the algorithm is symmetric), the weighted quantile $\mathsf{Q}_{1-\alpha}(\sum_{j=1}^{n+1} \tilde{w}_j \cdot \delta_{R_j})$ is no longer a symmetric function of $R_1, \dots, R_{n+1}$ if the weights $\tilde{w}_j$ are not all equal, and therefore, the equality will no longer be true in general.

Next, if we use nonsymmetric algorithms that take tagged data points (X_i, Y_i, t_i) as input, the situation becomes even more complex—even if the data points (X_i, Y_i) are exchangeable, the residuals $R_1, \dots, R_{n+1}$ may no longer be exchangeable as they depend on a fitted model $\widehat{\mu}$ that treats the training data points nonsymmetrically.

Finally, in this paper we are of course primarily interested in the setting where the data points are no longer exchangeable, and in bounding the resulting coverage gap. This leads to additional challenges, all of which we address in the proofs below.

6.2. *Proof of Theorem* 2. Since nonexchangeable split conformal is simply a special case of nonexchangeable full conformal, we only need to prove the result for full conformal.

For each $k \in [n+1]$, denote

$$\widehat{\mu}^k = \widehat{\mu}^{Y_{n+1},k} = \mathcal{A}\big((X_{\pi_k(1)}, Y_{\pi_k(1)}, t_1), \dots, (X_{\pi_k(n+1)}, Y_{\pi_k(n+1)}, t_{n+1})\big),$$

where for any $k \in [n]$, as before π_k denotes the permutation on $[n+1]$ that swaps indices k and $n+1$, while π_{n+1} is the identity permutation. Then, for any $k \in [n+1]$, we can calculate

$$\big(R(Z^k)\big)_i = \big|Y_{\pi_k(i)} - \widehat{\mu}^k(X_{\pi_k(i)})\big|,$$

and therefore,

$$\big(R(Z^K)\big)_i = \begin{cases} R_i^{Y_{n+1},K}, & \text{if } i \neq K \text{ and } i \neq n+1, \\ R_{n+1}^{Y_{n+1},K}, & \text{if } i = K, \\ R_K^{Y_{n+1},K}, & \text{if } i = n+1. \end{cases} \tag{20}$$

The definition of the nonexchangeable full conformal prediction set (18) reveals

$$Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \quad \Longleftrightarrow \quad R_{n+1}^{Y_{n+1},K} > \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{R_i^{Y_{n+1},K}}\right),$$

and we can equivalently write this as

$$Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \quad \Longleftrightarrow \quad R_{n+1}^{Y_{n+1},K} > \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{w}_i \cdot \delta_{R_i^{Y_{n+1},K}} + \tilde{w}_{n+1} \cdot \delta_{+\infty}\right). \tag{21}$$

Next, we verify that deterministically (20) implies

$$\mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{w}_i \cdot \delta_{R_i^{Y_{n+1},K}} + \tilde{w}_{n+1} \cdot \delta_{+\infty}\right) \geq \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{(R(Z^K))_i}\right). \tag{22}$$

Indeed, if $K = n+1$, then $R(Z^K) = R^{Y_{n+1},K}$ by (20), and so the bound holds trivially. If instead $K \leq n$, then the distribution on the left-hand side of (22) equals

$$\begin{aligned} &\sum_{i=1}^{n} \tilde{w}_i \cdot \delta_{R_i^{Y_{n+1},K}} + \tilde{w}_{n+1} \cdot \delta_{+\infty} \\ &\quad = \sum_{i=1,\dots,n; i \neq K} \tilde{w}_i \cdot \delta_{R_i^{Y_{n+1},K}} + \tilde{w}_K(\delta_{R_K^{Y_{n+1},K}} + \delta_{+\infty}) + (\tilde{w}_{n+1} - \tilde{w}_K)\delta_{+\infty}, \end{aligned}$$

while the distribution on the right-hand side of (22) can be rewritten as

$$\begin{aligned}\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{(R(Z^K))_i} &= \sum_{i=1,\dots,n;i\neq K} \tilde{w}_i \cdot \delta_{R_i^{Y_{n+1},K}} + \tilde{w}_K \delta_{R_{n+1}^{Y_{n+1},K}} + \tilde{w}_{n+1} \delta_{R_K^{Y_{n+1},K}} \\ &= \sum_{i=1,\dots,n;i\neq K} \tilde{w}_i \cdot \delta_{R_i^{Y_{n+1},K}} + \tilde{w}_K (\delta_{R_K^{Y_{n+1},K}} + \delta_{R_{n+1}^{Y_{n+1},K}}) + (\tilde{w}_{n+1} - \tilde{w}_K)\delta_{R_K^{Y_{n+1},K}},\end{aligned}$$

by applying (20). Since $w_K \in [0, 1]$ by assumption, we have $\tilde{w}_{n+1} \geq \tilde{w}_K$, which from the last two displays verifies that (22) must hold. Combining (21) and (22), we have

$$Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \quad \Longrightarrow \quad R_{n+1}^{Y_{n+1},K} > \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{(R(Z^K))_i}\right),$$

or equivalently by (20),

$$Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \quad \Longrightarrow \quad \bigl(R(Z^K)\bigr)_K > \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{(R(Z^K))_i}\right). \tag{23}$$

Next define a function $\mathcal{S}$ from $\mathbb{R}^{n+1}$ to subsets of $[n + 1]$, as follows: for any $r \in \mathbb{R}^{n+1}$,

$$\mathcal{S}(r) = \left\{i \in [n+1] : r_i > \mathsf{Q}_{1-\alpha}\left(\sum_{j=1}^{n+1} \tilde{w}_j \cdot \delta_{r_j}\right)\right\}. \tag{24}$$

These are the “strange” points—indices i for which r_i is unusually large, relative to the (weighted) empirical distribution of $r_1, \dots, r_{n+1}$. A direct argument (see, e.g., the deterministic inequality in Harrison ((2012), Lemma A.1)) shows that

$$\sum_{i\in\mathcal{S}(r)} \tilde{w}_i \leq \alpha \quad \text{for all } r \in \mathbb{R}^{n+1}, \tag{25}$$

that is, the (weighted) fraction of “strange” points cannot exceed α. From (23), we have that noncoverage of Y_{n+1} implies strangeness of point K:

$$Y_{n+1} \notin \widehat{C}_n(X_{n+1}) \quad \Longrightarrow \quad K \in \mathcal{S}\bigl(R(Z^K)\bigr). \tag{26}$$

Finally,

$$\begin{aligned}\mathbb{P}\bigl\{K \in \mathcal{S}\bigl(R(Z^K)\bigr)\bigr\} &= \sum_{i=1}^{n+1} \mathbb{P}\bigl\{K = i \text{ and } i \in \mathcal{S}\bigl(R(Z^i)\bigr)\bigr\} \\ &= \sum_{i=1}^{n+1} \tilde{w}_i \cdot \mathbb{P}\bigl\{i \in \mathcal{S}\bigl(R(Z^i)\bigr)\bigr\} \\ &\leq \sum_{i=1}^{n+1} \tilde{w}_i \cdot \bigl(\mathbb{P}\bigl\{i \in \mathcal{S}\bigl(R(Z)\bigr)\bigr\} + \mathsf{d}_{\mathsf{TV}}\bigl(R(Z), R(Z^i)\bigr)\bigr) \qquad (27) \\ &= \mathbb{E}\biggl[\sum_{i\in\mathcal{S}(R(Z))} \tilde{w}_i\biggr] + \sum_{i=1}^{n} \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}\bigl(R(Z), R(Z^i)\bigr) \\ &\leq \alpha + \sum_{i=1}^{n} \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}\bigl(R(Z), R(Z^i)\bigr),\end{aligned}$$

where the last step holds by (25), whereas step (27) holds because $K \perp\!\!\!\perp Z$ and $Z^i = \pi_i(Z)$ is a function of the data Z, and therefore, $K \perp\!\!\!\perp Z^i$.

7. Discussion.

Our main contribution in this paper was to demonstrate how conformal prediction, which has crucially relied on exchangeability, can be modified to handle nonsymmetric regression algorithms, and utilize weighted residual distributions in order to provide robustness against deviations from exchangeability in the data. With no assumptions whatsoever on the underlying joint distribution of the data, it is possible to give a coverage guarantee for both existing conformal methods, and our new proposed nonexchangeable conformal procedures. The coverage gap, expressing the extent to which the guaranteed coverage level is lower than what would be guaranteed under exchangeability, is bounded by a weighted sum of total variation distances between the residual vectors obtained by swapping the ith point with the $(n+1)$st point.

Our work opens the door to applying conformal prediction in applications where the data is globally likely far from exchangeable but locally deviates mildly from exchangeability. Tags and weights can be prudently used to downweight "far away" points during training and calibration, and recover reasonable coverage in practice. We hope our work will lead to more targeted methods that focus on custom design of nonsymmetric algorithms and weighting schemes to improve efficiency and robustness in specific applications, through the lens of nonexchangeable conformal prediction.

APPENDIX A: EXTENSION TO GENERAL NONCONFORMITY SCORES

In this section, we extend our new nonexchangeable inference methods for split and full conformal to the setting of general nonconformity scores. The response is no longer required to be real-valued, so we will consider the general setting with data points $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$.

For split conformal, as usual, we assume that the nonconformity score function $\widehat{S} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is pre-fitted. The nonexchangeable split conformal set is given by

$$\widehat{C}_n(X_{n+1}) = \left\{ y \in \mathcal{Y} : \widehat{S}(X_{n+1}, y) \le \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{w}_i \cdot \delta_{\widehat{S}(X_i, Y_i)} + \tilde{w}_{n+1} \cdot \delta_{+\infty} \right) \right\}. \tag{28}$$

For the special case $\widehat{S}(x, y) = |y - \widehat{\mu}(x)|$ (where $\widehat{\mu}$ is a pre-fitted function), note that this reduces to the previous definition (11) from before.

For full conformal, we now consider algorithms $\mathcal{A}$ of the form

$$\mathcal{A} : \bigcup_{n \ge 0} (\mathcal{X} \times \mathcal{Y} \times \mathcal{T})^n \to \{\text{measurable functions } \widehat{S} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}. \tag{29}$$

(As before, the symmetric algorithm setting, with no tags t_i, is simply a special case of this general formulation.) First, for any $y \in \mathbb{R}$ and any $k \in [n+1]$, define

$$\widehat{S}^{y,k} = \mathcal{A}\big((X_{\pi_k(i)}, Y^y_{\pi_k(i)}, t_i) : i \in [n+1] \big),$$

where the permutation π_k is defined as before (that swaps indices k and $n+1$), and where

$$Y_i^y = Y_i, \quad i = 1, \dots, n, \qquad Y_{n+1}^y = y,$$

as before. Define the scores from this model,

$$S_i^{y,k} = \widehat{S}^{y,k}(X_i, Y_i), \quad i = 1, \dots, n, \qquad S_{n+1}^{y,k} = \widehat{S}^{y,k}(X_{n+1}, y).$$

Then, after drawing a random index K as in (17), the prediction set is given by

$$\widehat{C}_n(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^{y,K} \le \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^{n+1} \tilde{w}_i \cdot \delta_{S_i^{y,K}} \right) \right\}. \tag{30}$$

For the special case $\widehat{S}(x, y) = |y - \widehat{\mu}(x)|$ (where $\widehat{\mu}$ is fitted on the same data), this again reduces to the previous definition (18) from before.

Importantly, the same theoretical result (i.e., Theorem 2) holds for these more general methods as well. The proof does not fundamentally rely on residual scores, and the modifications required for the general case are straightforward, so we omit the details here.

APPENDIX B: NONEXCHANGEABLE JACKKNIFE+

B.1. Background. The jackknife+ (Barber et al. (2021)) (closely related to "cross-conformal prediction" (Vovk (2015))) is a method that offers a compromise between the computational and statistical costs of the split and full conformal methods. For each $i = 1, \dots, n$, define the ith leave-one-out model as

$$\widehat{\mu}_{-i} = \mathcal{A}\big((X_1, Y_1), \dots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \dots, (X_n, Y_n)\big), \tag{31}$$

fitted to the training data with ith point removed. Define also the ith leave-one-out residual $R_i^{\mathrm{LOO}} = |Y_i - \widehat{\mu}_{-i}(X_i)|$, which avoids overfitting since data point (X_i, Y_i) is not used for training $\widehat{\mu}_{-i}$. The jackknife+ prediction interval is then given by[7]

$$\begin{aligned}\Bigg[&\mathsf{Q}_\alpha\Bigg(\sum_{i=1}^n \frac{1}{n+1}\cdot\delta_{\widehat{\mu}_{-i}(X_{n+1})-R_i^{\mathrm{LOO}}} + \frac{1}{n+1}\cdot\delta_{-\infty}\Bigg),\\ &\mathsf{Q}_{1-\alpha}\Bigg(\sum_{i=1}^n \frac{1}{n+1}\cdot\delta_{\widehat{\mu}_{-i}(X_{n+1})+R_i^{\mathrm{LOO}}} + \frac{1}{n+1}\cdot\delta_{+\infty}\Bigg)\Bigg].\end{aligned} \tag{32}$$

While in practice the jackknife+ generally provides coverage close to the target level $1-\alpha$ (and provably so under a stability assumption on $\mathcal{A}$), its theoretical guarantee only ensures $1-2\alpha$ probability of coverage in the worst case:

THEOREM 4 (Jackknife+ (Barber et al. (2021))). *If* $(X_1, Y_1), \dots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ *are i.i.d.* (*or more generally*, *exchangeable*), *and the algorithm* $\mathcal{A}$ *treats the input data points symmetrically as in* (6), *then the jackknife+ prediction interval defined in* (32) *satisfies*

$$\mathbb{P}\big\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\big\} \geq 1 - 2\alpha.$$

This method can be viewed as a form of n-fold cross-validation; more generally, the CV+ method (Barber et al. (2021)) uses K-fold cross-validation for any desired K, and obtains a similar distribution-free guarantee.

B.2. Methods. We next present the nonexchangeable jackknife+ method.

Nonexchangeable jackknife+ with a symmetric algorithm. We first consider the setting where the algorithm $\mathcal{A}$ is symmetric. To begin, we choose weights $w_i \in [0, 1]$, $i = 1, \dots, n$, which are fixed ahead of time, and as before, this gives rise to normalized weights as in (10). The prediction interval is then given by

$$\begin{aligned}\Bigg[&\mathsf{Q}_\alpha\Bigg(\sum_{i=1}^n \tilde{w}_i\cdot\delta_{\widehat{\mu}_{-i}(X_{n+1})-R_i^{\mathrm{LOO}}} + \tilde{w}_{n+1}\cdot\delta_{-\infty}\Bigg),\\ &\mathsf{Q}_{1-\alpha}\Bigg(\sum_{i=1}^n \tilde{w}_i\cdot\delta_{\widehat{\mu}_{-i}(X_{n+1})+R_i^{\mathrm{LOO}}} + \tilde{w}_{n+1}\cdot\delta_{+\infty}\Bigg)\Bigg],\end{aligned} \tag{33}$$

where $\widehat{\mu}_{-i}$ is defined as in (31), and $R_i^{\mathrm{LOO}} = |Y_i - \widehat{\mu}_{-i}(X_i)|$ as before.

[7] Abusing notation, here $\mathsf{Q}_\alpha(\cdot)$ is used to denote the largest possible α-quantile if it is not unique, while as before, $\mathsf{Q}_{1-\alpha}(\cdot)$ denotes the smallest possible $(1-\alpha)$-quantile if not unique.

Analogous to split and full conformal, here the original (unweighted) version of jackknife+ is recovered by choosing weights $w_1 = \cdots = w_n = 1$ in the new algorithm.

Nonexchangeable jackknife+ with a nonsymmetric algorithm. We now extend the nonexchangeable jackknife+ to allow for a nonsymmetric algorithm $\mathcal{A}$. For any $k \in [n+1]$ and any $i \in [n]$, define the model $\widehat{\mu}^k_{-i}$ as

$$\widehat{\mu}^k_{-i} = \mathcal{A}\big((X_{\pi_k(j)}, Y_{\pi_k(j)}, t_j) : j \in [n+1], \pi_k(j) \notin \{i, n+1\}\big).$$

As before, π_k is the permutation on $[n+1]$ that swaps indices k and $n+1$ (or, the identity permutation in the case $k = n+1$). Equivalently,

$$\widehat{\mu}^k_{-i} = \begin{cases} \mathcal{A}\big((X_j, Y_j, t_j) : j \in [n]\backslash\{i, k\}, (X_k, Y_k, t_{n+1})\big), & \text{if } k \in [n] \text{ and } k \neq i, \\ \mathcal{A}\big((X_j, Y_j, t_j) : j \in [n]\backslash\{i\}\big), & \text{if } k = n+1 \text{ or } k = i. \end{cases}$$

In other words, this model is fitted on the training data $(X_1, Y_1), \dots, (X_n, Y_n)$ but with the ith point removed, and furthermore the data point (X_k, Y_k) is given the tag t_{n+1} rather than t_k. (We note that computing the fitted model $\widehat{\mu}^k_{-i}$ does not require knowledge of the test point (X_{n+1}, Y_{n+1}), because $\pi_k(j) = n+1$ is excluded from the data set when running $\mathcal{A}$.) For the model $\widehat{\mu}^k_{-i}$, we define its corresponding leave-one-out residuals as

$$R_i^{k,\mathrm{LOO}} = \big|Y_i - \widehat{\mu}^k_{-i}(X_i)\big|.$$

To run the method, we first draw a random index K as in (17), and then compute the nonexchangeable jackknife+ prediction interval as

$$\left[\mathsf{Q}_\alpha\left(\sum_{i=1}^n \tilde{w}_i \cdot \delta_{\widehat{\mu}^K_{-i}(X_{n+1}) - R_i^{K,\mathrm{LOO}}} + \tilde{w}_{n+1} \cdot \delta_{-\infty}\right), \right.$$
$$\left. \mathsf{Q}_{1-\alpha}\left(\sum_{i=1}^n \tilde{w}_i \cdot \delta_{\widehat{\mu}^K_{-i}(X_{n+1}) + R_i^{K,\mathrm{LOO}}} + \tilde{w}_{n+1} \cdot \delta_{+\infty}\right)\right]. \tag{34}$$

Again, as was the case for nonexchangeable conformal, this method is a generalization of the symmetric case for nonexchangeable jackknife+, which was presented above.

B.3. Theory. As for the proof for nonexchangeable full conformal, we first need to define how we map a data sequence $z = (z_1, \dots, z_{n+1}) \in (\mathcal{X} \times \mathbb{R})^{n+1}$, with entries $z_i = (x_i, y_i)$, to the residuals $R_{\text{jack+}}(z)$. In the setting of jackknife+, however, the residuals are now a matrix rather than a vector. Given z, we define $\binom{n+1}{2}$ leave-two-out models: for each $i, j \in [n+1]$ with $i \neq j$, let

$$\widehat{\mu}_{-ij} = \widehat{\mu}_{-ji} = \mathcal{A}\big((x_k, y_k, t_k) : k \in [n+1]\backslash\{i, j\}\big).$$

Then define the matrix of residuals $R_{\text{jack+}}(z) \in \mathbb{R}^{(n+1)\times(n+1)}$ with entries

$$\big(R_{\text{jack+}}(z)\big)_{ij} = \big|y_i - \widehat{\mu}_{-ij}(x_i)\big|,$$

for all $i \neq j$, and zeros on the diagonal.

THEOREM 5 (Nonexchangeable jackknife+). *Let $\mathcal{A}$ be an algorithm mapping a sequence of triplets (X_i, Y_i, t_i) to a fitted function as in (16). Then the nonexchangeable jackknife+ defined in (34) satisfies*

$$\mathbb{P}\big\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\big\} \geq 1 - 2\alpha - \sum_{i=1}^n \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}\big(R_{\text{jack+}}(Z), R_{\text{jack+}}(Z^i)\big).$$

The proof of this result is given in the Supplementary Material. To summarize, we see that the coverage gap is bounded by

$$\sum_{i=1}^{n} \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}\big(R_{\text{jack+}}(Z), R_{\text{jack+}}(Z^i)\big).$$

As for the full conformal guarantee, this therefore implies the coverage gap is bounded by $\sum_i \tilde{w}_i \cdot \mathsf{d}_{\mathsf{TV}}(Z, Z^i)$, as well. Again, while this last bound can be viewed as more interpretable, in many settings it is substantially more loose.

While jackknife+ is defined specifically for the residual-based nonconformity score (i.e., the score $|y - \widehat{\mu}(x)|$ to measure the extent to which a data point (x, y) does not conform to observed trends in the data), in other settings we may wish to use alternative nonconformity scores. Jackknife+ is closely related to earlier work on the cross-conformal method (Vovk (2015), Vovk et al. (2018)). Unlike jackknife+, the cross-conformal method can be applied to arbitrary nonconformity scores. In the Supplementary Material, we will present a nonexchangeable version of the cross-conformal algorithm.

APPENDIX C: HUBER-ROBUSTNESS OF CONFORMAL PREDICTION

In this section, we consider an alternative form of robustness, which requires stricter assumptions on the distribution drift but will yield a stronger predictive coverage guarantee. First, consider a version of the classic Huber contamination model from robust statistics, where most of the data is i.i.d. from the target distribution $\mathcal{D}_{\text{target}}$ but some fraction ϵ of the data is arbitrarily corrupted. For simplicity, to start we consider observing training data point $Z_i = (X_i, Y_i)$ from the mixture model

$$\mathcal{D}_i = (1 - \epsilon)\mathcal{D}_{\text{target}} + \epsilon \mathcal{D}'_i. \tag{35}$$

Here $\mathcal{D}'_i$ denotes an arbitrary adversarial distribution, that could potentially corrupt the ith training data point. However, we want to ensure coverage with respect to the target distribution $\mathcal{D}_{\text{target}}$—that is, the test point $Z_{n+1} = (X_{n+1}, Y_{n+1})$ will be drawn from $\mathcal{D}_{\text{target}}$. Standard conformal prediction assumes $\epsilon = 0$. But, one may ask: how badly can such adversarial corruptions hurt coverage? Here, we will answer that question, but do so in a slightly more general manner. First, define a new measure of distance between distributions,

$$\mathsf{d}_{\text{mix}}(\mathcal{D}, \mathcal{D}') = \inf\{t \geq 0 : \mathcal{D} = (1 - t) \cdot \mathcal{D}' + t \cdot \mathcal{D}'' \text{ for some distribution } \mathcal{D}''\}. \tag{36}$$

Abusing notation, we will write $\mathsf{d}_{\text{mix}}(Z, Z') = \mathsf{d}_{\text{mix}}(\mathcal{D}, \mathcal{D}')$ if $Z \sim \mathcal{D}$ and $Z' \sim \mathcal{D}'$.

This “distance” can be thought of as measuring the contamination of $\mathcal{D}'$, in the Huber sense. Indeed, if the data did indeed come from the mixture model in (35), then we would have $\mathsf{d}_{\text{mix}}(Z_i, Z_{n+1}) \leq \epsilon$. (We note that d_{mix} is not a metric, and in particular, is not symmetric in its two arguments.)

We now state our theory for our weighted version of split conformal, full conformal, and jackknife+, in a more restricted setting where the data points are independent and the algorithm is symmetric. From this point on, we assume $w_1 + \cdots + w_n > 0$ to avoid a trivial setting. Define

$$\bar{w}_i = \frac{w_i}{w_1 + \cdots + w_n}, \quad i = 1, \dots, n.$$

THEOREM 6 (Multiplicative bounds). *Suppose that $Z_1, \dots, Z_{n+1}$ are independent. For any symmetric algorithm $\mathcal{A}$, the nonexchangeable full conformal method* (12) *satisfies*

$$\mathbb{P}\{Y_{n+1} \notin \widehat{C}_n(X_{n+1})\} \leq \frac{\alpha}{1 - \sum_{i=1}^{n} \bar{w}_i \cdot \mathsf{d}_{\text{mix}}(Z_i, Z_{n+1})}$$

(*which includes the nonexchangeable split conformal method* (11) *as a special case*), *and the nonexchangeable jackknife+ method* (33) *satisfies*

$$\mathbb{P}\{Y_{n+1} \notin \widehat{C}_n(X_{n+1})\} \leq \frac{2\alpha}{1 - \sum_{i=1}^n \bar{w}_i \cdot \mathsf{d}_{\text{mix}}(Z_i, Z_{n+1})}.$$

In particular, if each Z_i follows an ϵ-Huber contamination model relative to Z_{n+1} as in (35), then the bound on the noncoverage rate for both unweighted and weighted conformal methods inflates by a factor of at most $1/(1-\epsilon)$, that is, for split or full conformal prediction we get a noncoverage guarantee of $\alpha/(1-\epsilon)$ instead of the nominal level α. We note that the coverage gap here is multiplicative—that is, $\alpha/(1-\epsilon) \approx \alpha + \alpha\epsilon$, and so the coverage gap is proportional to α. If the target error level α is small, then this multiplicative bound can offer much tighter error control, as compared to the earlier additive bounds in Theorem 2, if the terms $\mathsf{d}_{\text{mix}}(Z_i, Z_{n+1})$ are small.

On the other hand, notice that in general, we have $\mathsf{d}_{\text{mix}}(Z_i, Z_{n+1}) \geq \mathsf{d}_{\text{TV}}(Z_i, Z_{n+1})$, and furthermore, it is possible to have $\mathsf{d}_{\text{mix}}(Z_i, Z_{n+1}) = 1$ even when $\mathsf{d}_{\text{TV}}(Z_i, Z_{n+1})$ is arbitrarily small. In a such setting the original additive bounds may give tighter results. Of course, an additional restriction is that the multiplicative bounds require independent data and symmetric algorithms, whereas the earlier theorems make no such assumptions.

Acknowledgments. The authors are grateful to the American Institute of Mathematics for supporting and hosting our collaboration. The authors are grateful to Vladimir Vovk for helpful feedback on an earlier draft of this paper. E.J.C. would like to thank John Cherian and Isaac Gibbs for their help with the presidential election data.

Funding. R.F.B. was supported by the National Science Foundation via grants DMS-1654076 and DMS-2023109, and by the Office of Naval Research via grant N00014-20-1-2337.

E.J.C. was supported by the Office of Naval Research grant N00014-20-1-2157, the National Science Foundation grant DMS-2032014, the Simons Foundation under award 814641, and the ARO grant 2003514594.

R.J.T. was supported by ONR grant N00014-20-1-2787.

SUPPLEMENTARY MATERIAL

Supplementary material for "Conformal prediction beyond exchangeability" (DOI: 10.1214/23-AOS2276SUPP; .pdf). This supplement contains details on the nonexchangeable cross-conformal method, additional proofs and calculations for the theoretical results in the main paper, additional simulations, and details for the election data set.

REFERENCES

ANGELOPOULOS, A. and BATES, S. (2023). Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **16** 495–591.

BARBER, R. F., CANDÈS, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2021). Predictive inference with the jackknife+. *Ann. Statist.* **49** 486–507. MR4206687 https://doi.org/10.1214/20-AOS1965

BARBER, R. F., CANDÈS, E. J., RAMDAS, A. and TIBSHIRANI, R. J. (2023). Supplement to "Conformal prediction beyond exchangeability." https://doi.org/10.1214/23-AOS2276SUPP

BATES, S., CANDÈS, E., LEI, L., ROMANO, Y. and SESIA, M. (2023). Testing for outliers with conformal p-values. *Ann. Statist.* **51** 149–178. MR4564852 https://doi.org/10.1214/22-aos2244

BURNAEV, E. and VOVK, V. (2014). Efficiency of conformalized ridge regression. In *Conference on Learning Theory* 605–622.

CANDÈS, E. J., LEI, L. and REN, Z. (2023). Conformalized survival analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 24–45. https://doi.org/10.1093/jrsssb/qkac004

CAUCHOIS, M., GUPTA, S. and ALI, A. (2020). Robust validation: Confident predictions even when distributions shift. arXiv preprint, arXiv:2008.04267.

CHERIAN, J. and BRONNER, L. (2020). How the Washington Post estimates outstanding votes for the 2020 presidential election. Available at https://s3.us-east-1.amazonaws.com/elex-models-prod/2020-general/write-up/election_model_writeup.pdf.

CHERNOZHUKOV, V., WÜTHRICH, K. and YINCHU, Z. (2018). Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference on Learning Theory* 732–749. PMLR.

DUNN, R., WASSERMAN, L. and RAMDAS, A. (2022). Distribution-free prediction sets for two-layer hierarchical models. *J. Amer. Statist. Assoc*. To appear.

FANNJIANG, C., BATES, S., ANGELOPOULOS, A. N., LISTGARTEN, J. and JORDAN, M. I. (2022). Conformal prediction for the design problem. arXiv preprint arXiv:2202.03613.

GIBBS, I. and CANDÈS, E. J. (2021). Adaptive conformal inference under distribution shift. *Adv. Neural Inf. Process. Syst*. **34**.

GUAN, L. (2023). Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika* **110** 33–50. MR4565442 https://doi.org/10.1093/biomet/asac040

HARRIES, M. (1999). Splice-2 comparative evaluation: Electricity pricing. Tech. rept., Univ.New South Wales.

HARRISON, M. T. (2012). Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika* **99** 57–69. MR2899663 https://doi.org/10.1093/biomet/asr079

KIVARANOVIC, D., JOHNSON, K. D. and LEEB, H. (2020). Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

LEI, J. (2019). Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika* **106** 749–764. MR4031197 https://doi.org/10.1093/biomet/asz046

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc*. **113** 1094–1111. MR3862342 https://doi.org/10.1080/01621459.2017.1307116

LEI, J., ROBINS, J. and WASSERMAN, L. (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc*. **108** 278–287. MR3174619 https://doi.org/10.1080/01621459.2012.751873

LEI, J. and WASSERMAN, L. (2014). Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **76** 71–96. MR3153934 https://doi.org/10.1111/rssb.12021

LEI, L. and CANDÈS, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **83** 911–938. MR4349122

MAO, H., MARTIN, R. and REICH, B. (2023). Valid model-free spatial prediction. *J. Amer. Statist. Assoc.* To appear.

PODKOPAEV, A. and RAMDAS, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*. PMLR.

ROMANO, Y., PATTERSON, E. and CANDÈS, E. J. (2019). Conformalized quantile regression. *Adv. Neural Inf. Process. Syst*. **32**.

SHAFER, G. and VOVK, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res*. **9** 371–421. MR2417240

STANKEVICIUTE, K., ALAA, A. M. and VAN DER SCHAAR, M. (2021). Conformal time-series forecasting. *Adv. Neural Inf. Process. Syst*. **34**.

TIBSHIRANI, R. J., BARBER, R. F., CANDÈS, E. J. and RAMDAS, A. (2019). Conformal prediction under covariate shift. *Adv. Neural Inf. Process. Syst*. **32**.

VOLKHONSKIY, D., BURNAEV, E., NOURETDINOV, I., GAMMERMAN, A. and VOVK, V. (2017). Inductive conformal martingales for change-point detection. In *Conformal and Probabilistic Prediction and Applications* 132–153. PMLR.

VOVK, V. (2015). Cross-conformal predictors. *Ann. Math. Artif. Intell*. **74** 9–28. MR3353894 https://doi.org/10.1007/s10472-013-9368-4

VOVK, V. (2021). Testing randomness online. *Statist. Sci*. **36** 595–611. MR4323055 https://doi.org/10.1214/20-sts817

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

VOVK, V., NOURETDINOV, I., MANOKHIN, V. and GAMMERMAN, A. (2018). Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications* 37–51. PMLR.

VOVK, V., PETEJ, I. and GAMMERMAN, A. (2021). Protected probabilistic classification. In *Conformal and Probabilistic Prediction and Applications* 297–299. PMLR.

XU, C. and XIE, Y. (2021). Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*. PMLR.

ZAFFRAN, M., FÉRON, O., GOUDE, Y. and JOSSE, J. (2022). Adaptive conformal predictions for time series. In *International Conference on Machine Learning*. PMLR.