Unifying Different Theories of Conformal Prediction

Rina Foygel Barber[*]   Ryan J. Tibshirani[†]

Abstract

This paper establishes a unified framework for understanding the methodology and theory behind several different methods in the conformal prediction literature, which includes standard conformal prediction (CP), weighted conformal prediction (WCP), nonexchangeable conformal prediction (NexCP), and randomly-localized conformal prediction (RLCP), among others. At the crux of our framework is the idea that conformal methods are based on revealing *partial information* about the data at hand, and positing a conditional distribution for the data given the partial information. Different methods arise from different choices of partial information, and of the corresponding (approximate) conditional distribution. In addition to recovering and unifying existing results, our framework leads to both new theoretical guarantees for existing methods, and new extensions of the conformal methodology.

1 Introduction

As machine learning algorithms become increasingly embedded in prediction systems, it has become increasingly important to address a question of reliability: how can we quantify our uncertainty in the predictions that are produced by black-box models? Conformal prediction (Vovk et al., 2005) is a framework for providing predictive inference around the output of a black-box algorithm, offering a guarantee of predictive coverage that relies only on assuming the data points are exchangeable, without placing further assumptions on the distribution of the data or any assumptions on the algorithm used for modeling.

To be more precise, conformal prediction works in a setting in which we observe training data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$, where each X_i is a feature (e.g., a vector of covariates) and each Y_i is a response. We also have a test point (X_{n+1}, Y_{n+1}), where the feature X_{n+1} is observed, and our task is to predict the unobserved response Y_{n+1}. Many different methods from statistics and machine learning can be applied to produce a fitted model $\hat{f} : \mathcal{X} \to \mathcal{Y}$ to (often accurately) predict Y from X. How can we build a prediction interval around $\hat{f}(X_{n+1})$ to quantify our uncertainty about Y_{n+1}? To answer this question, conformal prediction relies on the assumption that the data is exchangeable, i.e., the assumption that the distribution of $((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}))$ is invariant to permuting these $n + 1$ data points. Section 2 gives more background on this method.

---

[*]Department of Statistics, University of Chicago
[†]Department of Statistics, University of California, Berkeley

In recent years, the conformal prediction literature has seen a flurry of development. Some of this work features new choices for the score function—a key component of the conformal framework, or specialized (efficient) implementations for certain prediction algorithms. While these developments can lead to substantially improved performance in practice, they do not represent fundamental modifications of the conformal framework. Other recent developments have extended the framework itself, to handle different relaxations of exchangeability. These extensions are the focus of our work, and include the following:

- Weighted conformal prediction (Tibshirani et al., 2019), which applies likelihood-based weights to accommodate settings where the training and test data come from different distributions, with a known shift. Tibshirani et al. (2019) focus on the case of covariate shift, while Podkopaev and Ramdas (2021) study the problem of label shift.

- Nonexchangeable conformal prediction (Barber et al., 2023), which applies a different sort of weighting to improve the robustness of the prediction sets to mild but unknown violations of the exchangeability assumption.

- Localized conformal prediction (Guan, 2023) and randomly-localized conformal prediction (Hore and Barber, 2023), which modify the method to place higher weight on data points closer to the test point X_{n+1}, in order to aim for better locally adaptivity, i.e., coverage which approximately holds once we condition on X_{n+1}.

- Generalized weighted conformal prediction (Prinster et al., 2024), which offers a generalized view through the lens of understanding distributions of all possible permutations of the data, to handle arbitrary forms of nonexchangeability; earlier work by Fannjiang et al. (2022) treats an important special case arising in experimental design.

Our contributions. In this work, we construct a unified theory that brings together these methods and their supporting theory. At a high level, this unified view follows a simple but general recipe: given a dataset $Z = ((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}))$, we suppose that we observe partial information U about Z. The construction of a conformal prediction set then relies on calculating the (exact or approximate) distribution of the test point (X_{n+1}, Y_{n+1}) conditional on the partial information U. For example, in standard conformal prediction, we define U to be the unordered collection of the data points in Z, and the test point (X_{n+1}, Y_{n+1}) is equally likely to be any one of the $n + 1$ data points in this collection, due to exchangeability.

We will see that this unified view recreates the existing extensions of standard conformal prediction highlighted above, offering a shared theory that explains these seemingly disparate generalizations which have been discovered in the recent literature. We will also see that the unified view allows us to fluidly derive new results, ultimately leading to broader applicability of the conformal prediction framework.

Additional related work. While the focus of our work is on unifying various extensions of conformal prediction under departures from exchangeability, many other new developments have been recently made in the literature on conformal prediction.

While conformal prediction can be paired with any modeling algorithm (to make predictions), and any score function (to measure the predictive model's accuracy), the utility of the

method relies on choosing this pairing appropriately for the problem at hand. There is by now a rich literature investigating different choices, and we highlight just a few contributions. For the setting of a real-valued response Y, popular options include the residual score and scaled residual score methods studied by Lei et al. (2018), the quantile score studied by Romano et al. (2019), a method for identifying regions of high density by Izbicki et al. (2022), and a nearest-neighbor method studied by Györfi and Walk (2019). For the setting of a categorical response, various novel score functions have also been proposed and studied by Sadinle et al. (2019); Romano et al. (2020); Angelopoulos et al. (2021), among others.

Other lines of work focus on developing methodology and/or theory in different contexts. We highlight two such lines. First, if the data in fact follows a specified model (which can be known exactly or approximately), then it is possible to derive theory which ensures that the conformal prediction method is competitive with model-based methods; see, e.g., Chapter 5 of Angelopoulos et al. (2024b) for an overview of results of this type. Second, in the presence of arbitrary (possibly large and discontinuous) and unknown distribution shift, we can no longer rely on any of the extensions described above which consider deviations from exchangeability. In this setting, the work of Gibbs and Candès (2021) proposes an online variant of conformal prediction that offers a weaker guarantee—coverage is guaranteed on average over the stream of data. These ideas have been further extended by Gibbs and Candès (2024); Zaffran et al. (2022); Bhatnagar et al. (2023); Angelopoulos et al. (2023, 2024a), among others.

2 Background

In this section, we review background material on (standard) conformal prediction. We also review an alternative formulation of conformal prediction using the language of hypothesis testing, which will be used to construct our unified framework in the following sections. For more background on the ideas explored in this section see, e.g., Vovk et al. (2005); Lei et al. (2018); Angelopoulos et al. (2024b).

2.1 Conformal prediction for exchangeable data

Given training data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$, and a test covariate $X_{n+1} \in \mathcal{X}$, suppose we would like to predict the corresponding response value $Y_{n+1} \in \mathcal{Y}$. Conformal prediction is a framework for producing a prediction interval, or more generally, a prediction set $\mathcal{C}(X_{n+1})$ that aims to contain the test response Y_{n+1} with some prescribed coverage probability $1 - \alpha$. To run conformal prediction, we need to specify a score function:

$$\mathsf{s} : (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y})^{n+1} \to \mathbb{R},$$

which compares a data point (x, y) to a given dataset of points $z_i = (x_i, y_i)$, $i = 1, \ldots, n+1$. Larger values of the score indicate that (x, y) does not "conform" to the trends observed in the dataset. A canonical example is any score function of the form

$$\mathsf{s}\big((x, y), (z_1, \ldots, z_{n+1})\big) = \ell(y, \hat{f}(x)), \quad \text{for } \hat{f} = \mathcal{A}\big(z_1, \ldots, z_{n+1}\big), \tag{1}$$

where ℓ is a loss function on $\mathcal{Y} \times \mathcal{Y}$, and $\mathcal{A}$ is a regression algorithm (e.g., we could use least squares regression in the real-valued case, $\mathcal{Y} = \mathbb{R}$) which inputs a dataset and outputs a fitted model $\hat{f} : \mathcal{X} \to \mathcal{Y}$.

Given the score function s, denoting each training point by $Z_i = (X_i, Y_i)$, we define for arbitrary $y \in \mathcal{Y}$ an augmented dataset $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$, and define scores

$$S_i^y = s(Z_i^y, Z^y), \quad i \in [n+1]. \tag{2}$$

where here and throughout we write $[N] = \{1, \ldots, N\}$ for an integer $N \geq 1$. The conformal prediction (CP) set for Y_{n+1} is then given by

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \text{Quantile}_{1-\alpha}(S_1^y, \ldots, S_{n+1}^y) \right\}, \tag{3}$$

where $\text{Quantile}_\tau(P) = \inf\{x \in \mathbb{R} : \mathbb{P}\{V \leq x\} \geq \tau\}$ denotes the level-τ quantile of a random variable $V \sim P$, and we use $\text{Quantile}_\tau(v_1, \ldots, v_N) = \text{Quantile}_\tau(\frac{1}{N}\sum_{i=1}^N \delta_{v_i})$ to abbreviate the quantile of the empirical distribution associated with a vector $v \in \mathbb{R}^N$.

Next we state a well-known, finite-sample guarantee underlying conformal prediction.

Theorem 1 (Vovk et al. 2005). *Suppose $Z_1, \ldots, Z_{n+1}$ are exchangeable, and the score function s is* symmetric *in its second argument:*

$$s\big((x, y), (z_1, \ldots, z_{n+1})\big) = s\big((x, y), (z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})\big), \quad \text{for all } \sigma \in \mathcal{S}_{n+1}, \tag{4}$$

where $\mathcal{S}_{n+1}$ is the set of permutations on $[n+1]$. Then the CP set in (3) satisfies

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

To take a closer look at this symmetry condition on s, in the example of the loss-based score function as in (1), this is equivalent to requiring the algorithm $\mathcal{A}$ to be symmetric in the training data—that is, the function $\hat{f} = \mathcal{A}(z_1, \ldots, z_{n+1})$ is unchanged if we permute the data points $z_1, \ldots, z_{n+1}$ in the training set.

The bag of data. For a given $z = (z_1, \ldots, z_{n+1})$, we write $\lfloor z \rceil = \lfloor z_1, \ldots, z_{n+1} \rceil$ to denote the unordered "bag" of data, that is, the (unordered) multiset obtained from the (ordered) vector $z \in \mathcal{Z}^{n+1}$. For example, if $z = (1, 2, 1, 5)$, then $\lfloor z \rceil$ conveys the information that the dataset contains two 1s, one 2, and one 5, but does not specify the order in which these four values appear. Since Theorem 1 assumes $s((x, y), (z_1, \ldots, z_{n+1}))$ is invariant to permutations of the values $z_1, \ldots, z_{n+1}$, we can write the score function under this condition as

$$s((x, y), \lfloor z_1, \ldots, z_{n+1} \rceil).$$

Validity of CP. With the bag notation in place, we next review the proof of Theorem 1. Though there are by now several different ways of writing the proof, the key intuition is as follows: under exchangeability, conditional on observing the bag $\lfloor Z \rceil = \lfloor Z_1, \ldots, Z_n, Z_{n+1} \rceil$, the test point $Z_{n+1} = (X_{n+1}, Y_{n+1})$ is equally likely to be any one of the $n+1$ elements. (See also Dobriban and Lin (2023) for a discussion of CP from this conditional perspective.)

Proof of Theorem 1. Write $S_i = s(Z_i, \lfloor Z_1, \ldots, Z_{n+1} \rceil)$, for $i \in [n+1]$. Recalling (2), we can observe that $S_i = S_i^{Y_{n+1}}$ (taking $y = Y_{n+1}$). Hence, by construction of the CP set (3),

$$Y_{n+1} \in \mathcal{C}(X_{n+1}) \iff S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_{n+1}),$$

thus we need to prove that $\mathbb{P}\left\{S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_{n+1})\right\} \geq 1 - \alpha$. By definition of the quantile, we must have $\frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{1}\left\{S_i \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_{n+1})\right\} \geq 1 - \alpha$, and so

$$\frac{1}{n+1}\sum_{i=1}^{n+1} \mathbb{P}\left\{S_i \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_{n+1})\right\} \geq 1 - \alpha,$$

after taking an expectation. Our last step, then, is to verify that for each $i \in [n+1]$,

$$\mathbb{P}\left\{S_{n+1} \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_{n+1})\right\} = \mathbb{P}\left\{S_i \leq \text{Quantile}_{1-\alpha}(S_1, \ldots, S_{n+1})\right\},$$

We will prove this by showing that the vector of scores is exchangeable, i.e., for all $\sigma \in \mathcal{S}_{n+1}$,

$$(S_1, \ldots, S_{n+1}) \overset{\mathrm{d}}{=} (S_{\sigma(1)}, \ldots, S_{\sigma(n+1)}).$$

This is a consequence of the exchangeability of the data, together with the assumption of symmetry of s. To be more precise, define a function $f : \mathcal{Z}^{n+1} \to \mathbb{R}^{n+1}$ by

$$f(z) = \big(\mathsf{s}(z_1, \lfloor z \rfloor), \ldots, \mathsf{s}(z_{n+1}, \lfloor z \rfloor)\big).$$

Because the bag of data does not change when the data points are permuted, we can observe that f commutes with permutations—that is, for any $\sigma \in \mathcal{S}_{n+1}$,

$$(s_1, \ldots, s_{n+1}) = f(z_1, \ldots, z_{n+1}) \iff (s_{\sigma(1)}, \ldots, s_{\sigma(n+1)}) = f(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)}).$$

We therefore have for any $\sigma \in \mathcal{S}_{n+1}$,

$$(S_1, \ldots, S_{n+1}) = f(Z_1, \ldots, Z_{n+1}) \overset{\mathrm{d}}{=} f(Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)}) = (S_{\sigma(1)}, \ldots, S_{\sigma(n+1)}),$$

where the first step holds by definition of the scores, the second step holds by exchangeability of the data, and the last step holds since f commutes with permutations. $\square$

2.2 Reframing conformal via hypothesis testing

CP provides inference in the form of a prediction set, but the idea can be equivalently cast in the language of p-values and hypothesis testing. To develop this equivalence, we define what is known as a conformal p-value: for arbitrary $y \in \mathcal{Y}$, this is

$$p(y) = \frac{\sum_{i=1}^{n+1} \mathbb{1}\left\{S_i^y \geq S_{n+1}^y\right\}}{n+1}, \tag{5}$$

for $S_1^y, \ldots, S_{n+1}^y$ as defined in (2). Informally, $p(y)$ is often viewed as a p-value for testing the hypothesis $Y_{n+1} = y$, given exchangeable $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$. The following result explains the connection between conformal p-values and prediction sets.

Proposition 1 (Vovk et al. 2005). *The conformal prediction set $\mathcal{C}(X_{n+1})$ defined in (3) can be written in terms of the conformal p-value $p(y)$ defined in (5), as*

$$\mathcal{C}(X_{n+1}) = \big\{y \in \mathcal{Y} : p(y) > \alpha\big\}.$$

5

In order to prove the proposition, we need a result relating p-values and quantiles. Its proof is given Appendix A.1.

Lemma 1. *Let Q be a distribution on $\mathbb{R}$, which is supported on finitely many values. Then for any $\alpha \in [0, 1]$ and $x \in \mathbb{R}$,*

$$\mathbb{P}_Q \{X \geq x\} > \alpha \iff x \leq \text{Quantile}_{1-\alpha}(Q).$$

Proof of Proposition 1. By definition of the conformal prediction set, we have

$$y \in \mathcal{C}(X_{n+1}) \iff S_{n+1}^y \leq \text{Quantile}_{1-\alpha}\left(\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{S_i^y}\right).$$

where recall $\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{S_i^y}$ is the empirical distribution of the scores $S_1^y, \ldots, S_{n+1}^y$. Applying Lemma 1, we have that for any $s \in \mathbb{R}$,

$$s \leq \text{Quantile}_{1-\alpha}\left(\frac{1}{n+1}\sum_{i=1}^{n+1}\delta_{S_i^y}\right) \iff \frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{1}\{S_i^y \geq s\} > \alpha.$$

Taking $s = S_{n+1}^y$, and combining with the calculation above, we have

$$y \in \mathcal{C}(X_{n+1}) \iff \frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{1}\left\{S_i^y \geq S_{n+1}^y\right\} > \alpha,$$

which by definition of $p(y)$ in (5), completes the proof. $\qquad\square$

The coverage result in Theorem 1 can therefore be translated into the following statement about the conformal p-value.

Corollary 1 (Vovk et al. 2005)**.** *If $Z_1, \ldots, Z_{n+1}$ are exchangeable, and the score function is symmetric as in* (4)*, then the conformal p-value defined in* (5) *satisfies*

$$\mathbb{P}\{p(Y_{n+1}) \leq \alpha\} \leq \alpha, \quad \text{for all } \alpha \in [0, 1].$$

One way to interpret Proposition 1 and Corollary 1 is that the conformal prediction set is effectively given by inverting a permutation test, leveraging the fact that all $(n+1)!$ orderings of the dataset $Z = (Z_1, \ldots, Z_{n+1})$ are equally likely (due to exchangeability). To make this more explicit, observe that the conformal p-value in (5) is equivalently

$$p(y) = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}}\mathbb{1}\left\{\mathsf{s}(Z_{\sigma(n+1)}^y, Z_\sigma^y) \geq \mathsf{s}(Z_{n+1}^y, Z^y)\right\}}{(n+1)!},$$

where recall $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$, and we use $Z_\sigma^y = (Z_{\sigma(1)}^y, \ldots, Z_{\sigma(n+1)}^y)$ for the result of permuting its entries under σ. If we think of $\varphi(Z^y) = \mathsf{s}(Z_{n+1}^y, Z^y)$ as a test statistic of the dataset Z^y, then the above reveals that $p(y)$ is precisely a permutation p-value based on φ:

$$p(y) = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}}\mathbb{1}\{\varphi(Z_\sigma^y) \geq \varphi(Z^y)\}}{(n+1)!}.$$

6

The conformal prediction guarantee now becomes familiar: by the classical theory of permutation testing, we immediately know that $p(Y_{n+1})$ is a valid p-value since $Z^{Y_{n+1}} = Z$ has an exchangeable distribution.

We note that the connections between conformal prediction and hypothesis testing extend beyond standard CP to its many variants, such as WCP and NexCP. While these and other extensions are typically directly via prediction sets, they can also be represented through the language of conformal p-values. Our unified framework (Sections 3 and 5) will also use the language of hypothesis testing, to allow for a simple and clean exposition. When we apply the unified framework to derive specific variants of CP (Sections 4 and 6), we will verify the equivalence of its original formulation and p-value representation, in each case.

3 A unified framework

We are now ready to build our unified framework for conformal prediction methods. Recall that, in the standard exchangeable setting, the premise of conformal prediction is that after conditioning on the unordered bag of data $\lfloor Z \rceil$ we know the distribution of Z—it is simply uniform over all possible permutations of the elements in the bag. For our unified framework, we will generalize this idea in two ways: first, we will allow for settings that are more general than exchangeability, and second, we will allow for conditioning on more information than the bag of data $\lfloor Z \rceil$. To begin, we will need the following two ingredients:

- *Partial information* about the data, encoded by a random variable $U \in \mathcal{U}$ on which we will condition to perform inference. We will assume that U always contains sufficient information to reveal the unordered bag of data $\lfloor Z \rceil$, that is,

$$\lfloor Z \rceil = h(U) \text{ almost surely,}$$

for some function h. In many cases, we will simply have $U = \lfloor Z \rceil$, but in others U will contain additional information. Further, U may be equal to a deterministic function of Z (such as $U = \lfloor Z \rceil$), or may contain auxiliary sources of randomness.

- A *score function* $\mathsf{s} : \mathcal{Z}^{n+1} \times \mathcal{U} \to \mathbb{R}$. As in standard CP, the value of $\mathsf{s}(Z, U)$ is intended to reflect the nonconformity of the last data point Z_{n+1}, relative to the observed data. That is, a large value of $\mathsf{s}(Z, U)$ indicates that Z_{n+1} is likely an outlier.

We next need to determine the distribution of the statistic $\mathsf{s}(Z, U)$, in order to compute a p-value. Essentially, we proceed by approximating the conditional distribution of $Z \mid U$, and then computing a p-value for the observed statistic $\mathsf{s}(Z, U)$ using this conditional distribution. This requires a third ingredient:

- An *approximation* $Q_{Z|U}$ of the conditional distribution of $Z \mid U$. Notice that the true conditional distribution of $Z \mid U$ is supported on the finite set $\{z \in \mathcal{Z}^{n+1} : \lfloor z \rceil = h(U)\}$ (as $\lfloor Z \rceil = h(U)$, almost surely). We assume $Q_{Z|U}$ also has support contained in this set.

With these ingredients in place, we then define the p-value

$$p = p(Z, U) \quad \text{where} \quad p(z, u) = \mathbb{P}_{Q_{Z|U=u}} \{\mathsf{s}(Z, u) \geq \mathsf{s}(z, u)\}. \tag{6}$$

As explained in Section 2.2, conformal prediction methods can equivalently be written in terms of conformal p-values, thus predictive coverage guarantees for $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\}$ can be obtained by bounding $\mathbb{P}\{p \le \alpha\}$ for the p-value constructed in (6).

3.1 Main theorem

In order to examine the event $p \le \alpha$, it will be useful to define the quantity

$$\mathsf{t}_\alpha(u) = \mathrm{Quantile}_{1-\alpha}(Q_{S|U=u}),$$

where $Q_{S|U=u}$ denotes the conditional distribution of the score $\mathsf{s}(Z, u)$ for $Z \sim Q_{Z|U=u}$. By Lemma 1, since $Q_{S|U=u}$ is a distribution with finite support, it holds that

$$p(z, u) \le \alpha \iff \mathsf{s}(z, u) > \mathsf{t}_\alpha(u). \tag{7}$$

We now have the following guarantee for the p-value from the unified framework. Here and henceforth, $\mathrm{d}_{\mathrm{TV}}(\cdot, \cdot)$ denotes the total variation (TV) distance between two distributions.

Theorem 2. *Suppose $(Z, U) \sim P_{Z,U}$. Let $P_{S,T}$ be the induced joint distribution on $(S, T) = (\mathsf{s}(Z, U), \mathsf{t}_\alpha(U))$. Then the p-value defined in (6) satisfies*

$$\mathbb{P}\{p \le \alpha\} \le \alpha + \inf_{Q_U} \mathrm{d}_{\mathrm{TV}}(P_{S,T}, Q_{S,T}),$$

where the infimum is taken over all distributions Q_U on U, and where $Q_{S,T}$ denotes the joint distribution of $(S, T) = (\mathsf{s}(Z, U), \mathsf{t}_\alpha(U))$ induced by drawing $(Z, U) \sim Q_{Z,U} = Q_{Z|U} \times Q_U$.

Before giving the proof, we pause to comment on some alternative versions of this bound that are implied by the theorem.

Remark 1. As $(S, T) = (\mathsf{s}(Z, U), \mathsf{t}_\alpha(U))$, we must have $\mathrm{d}_{\mathrm{TV}}(P_{S,T}, Q_{S,T}) \le \mathrm{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U})$. This follows from a standard property of the total variation distance, which we will refer to as *information monotonicity*:

$$\mathrm{d}_{\mathrm{TV}}(P_{f(W)}, Q_{f(W)}) \le \mathrm{d}_{\mathrm{TV}}(P_W, Q_W) \text{ for any distributions } P_W, Q_W \text{ and any function } f.$$

(Here $P_{f(W)}$ denotes the distribution of $f(W)$ for $W \sim P$, and similarly for $Q_{f(W)}$.) Consequently, Theorem 2 implies a weaker bound,

$$\mathbb{P}\{p \le \alpha\} \le \alpha + \inf_{Q_U} \mathrm{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}).$$

The same argument holds more generally. For example, since (S, T) is a function of (S, U), we also have $\mathrm{d}_{\mathrm{TV}}(P_{S,T}, Q_{S,T}) \le \mathrm{d}_{\mathrm{TV}}(P_{S,U}, Q_{S,U})$. In the applications examined later, we will use various versions of this type of weaker bound, as convenient for each example.

Remark 2. Let $P_{Z|U}$ denote the conditional distribution associated with the joint $P_{Z,U}$. If $Q_{Z|U} = P_{Z|U}$ holds almost surely, i.e., our choice for the conditional distribution of $Z \mid U$ is exactly correct in implementing the unified conformal method, then by taking $Q_U = P_U$, we obtain $Q_{Z,U} = P_{Z,U}$ and therefore $\mathrm{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}) = 0$ which leads to

$$\mathbb{P}\{p \le \alpha\} \le \alpha.$$

3.2 Proof of Theorem 2

To build up toward the proof of the unified result, it is helpful to first recall some standard properties of hypothesis testing. Suppose that we observe data Z, and we have a prespecified test statistic—a function φ which maps data Z to a value $\varphi(Z) \in \mathbb{R}$ (with the convention that higher values indicate more evidence against the null). Fixing a null hypothesis $H_0 : Z \sim Q$ for a distribution Q, we may compute a p-value $p(Z)$, where $p(z) = \mathbb{P}_Q \{\varphi(Z) \geq \varphi(z)\}$ is the probability of observing a statistic at least as large as $\phi(z)$ under the null (for a one-sided test). By construction, the following holds:

$$\text{if } Z \sim Q \text{ then } \mathbb{P}\{p(Z) \leq \alpha\} \leq \alpha, \tag{8}$$

for all $\alpha \in [0, 1]$. But it can also be of interest to study the behavior of the p-value p under a different distribution, for which we have the following classical fact:

$$\text{if } Z \sim P \text{ then } \mathbb{P}\{p(Z) \leq \alpha\} \leq \alpha + \mathrm{d_{TV}}(P, Q). \tag{9}$$

The above fact is often interpreted as a statement about power, i.e., if P and Q are close in TV distance, then our test cannot have high power for rejecting Q in favor of the alternative P. However, we can also view (9) as a statement about robustness to model misspecification: if the null hypothesis of interest is $H_0 : Z \sim P$, and we only have access to an approximation Q of this null distribution, then the test may have inflated Type I error—but this inflation cannot be larger than the total variation distance $\mathrm{d_{TV}}(P, Q)$.

With this intuition in place, we are now ready to prove the theorem. At a high level, the proof can be viewed as a conditional version of the standard bound (9) above.

Proof of Theorem 2. By the standard fact (8) about p-values (applied with $Q_{Z|U=u}$ in place of Q, and the score function $\mathsf{s}(\cdot, u)$ in place of the test statistic φ), we have

$$\mathbb{P}_{Q_{Z|U=u}} \{p(Z, u) \leq \alpha\} \leq \alpha.$$

As this holds for each $u \in \mathcal{U}$, by averaging over $U \sim Q_U$, for any marginal distribution Q_U,

$$\mathbb{P}_{Q_{Z,U}} \{p(Z, U) \leq \alpha\} \leq \alpha.$$

Therefore, by definition of $Q_{S,T}$, along with the equivalence (7) relating p-values to thresholds,

$$\mathbb{P}_{Q_{S,T}} \{S > T\} = \mathbb{P}_{Q_{Z,U}} \{\mathsf{s}(Z, U) > \mathsf{t}_\alpha(U)\} = \mathbb{P}_{Q_{Z,U}} \{p(Z, U) \leq \alpha\} \leq \alpha.$$

Finally, observe that by (7) once again,

$$\begin{aligned}
\mathbb{P}_{P_{Z,U}} \{p(Z, U) \leq \alpha\} &= \mathbb{P}_{P_{Z,U}} \{\mathsf{s}(Z, U) > \mathsf{t}_\alpha(U)\} \\
&= \mathbb{P}_{P_{S,T}} \{S > T\} \\
&\leq \mathbb{P}_{Q_{S,T}} \{S > T\} + \mathrm{d_{TV}}(P_{S,T}, Q_{S,T}) \\
&\leq \alpha + \mathrm{d_{TV}}(P_{S,T}, Q_{S,T}),
\end{aligned}$$

where the next-to-last step holds by definition of total variation distance. $\square$

4 Special cases: known results

In this section, we examine a range of special cases in which the unified framework is able to reproduce known results in the literature. In each subsection below, we review the setting underlying a particular conformal method, the definition of the method itself, the associated theory, and then demonstrate how this can be viewed from the lens of the unified framework.

4.1 Standard conformal prediction

Standard conformal prediction (CP) (Vovk et al., 2005) works in the setting where the data $Z_1, \ldots, Z_{n+1}$ are exchangeable. This method was already described above in Section 2.1, but we briefly review it again here for completeness before describing its reformulation under the unified framework.

4.1.1 Method and theory

Fix any score function of the form $\mathsf{s}((x, y), \lfloor z \rceil)$, which assigns a value to a single data point (its first argument), based on a bag of data points (its second argument). For arbitrary $y \in \mathcal{Y}$, we define the augmented dataset $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$, and scores

$$S_i^y = \mathsf{s}(Z_i^y, \lfloor Z^y \rceil), \quad i \in [n+1].$$

We define the CP set at coverage level $1 - \alpha$ by

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \mathrm{Quantile}_{1-\alpha} \left(\frac{1}{n+1} \sum_{i=1}^{n+1} \delta_{S_i^y} \right) \right\}. \tag{10}$$

Recalling Theorem 1, if the data points $Z_1, \ldots, Z_{n+1}$ are exchangeable, then CP provides a marginal coverage guarantee $\mathbb{P} \{ Y_{n+1} \in \mathcal{C}(X_{n+1}) \} \geq 1 - \alpha$.

4.1.2 View from the unified framework

We now describe how CP fits into the unified conformal framework.

Choices of U and $Q_{Z|U}$. We define the partial information as $U = \lfloor Z \rceil$. In standard CP, given the bag $u = \lfloor z \rceil$, the score assigned to z depends only on z_{n+1}. Overloading notation, we can thus write the score function as

$$\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, \lfloor z \rceil).$$

Next we take the choice of the conditional distribution $Q_{Z|U}$ to be

$$Q_{Z|U=\lfloor z \rceil} = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \delta_{z_\sigma}.$$

This is the uniform distribution over all permutations of z, i.e., all vectors consistent with the observed bag $\lfloor z \rceil$ of data points.

P-value. Under these choices, the p-value in (6) is

$$p = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \mathbb{1} \left\{ \mathsf{s}(Z_{\sigma(n+1)}, \wr Z \wr) \geq \mathsf{s}(Z_{n+1}, \wr Z \wr) \right\}$$

$$= \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{1} \left\{ \mathsf{s}(Z_i, \wr Z \wr) \geq \mathsf{s}(Z_{n+1}, \wr Z \wr) \right\},$$

where the second line follows from the fact that, for each i, there are $n!$ many permutations $\sigma \in \mathcal{S}_{n+1}$ with $\sigma(n+1) = i$. As we can see, this is the conformal p-value $p(Y_{n+1})$ defined in (5). By Proposition 1, the event $p > \alpha$ is equivalent to $Y_{n+1} \in \mathcal{C}(X_{n+1})$, or, in other words, bounding $\mathbb{P}\{p \leq \alpha\}$ (as we will do next via the unified theory) is equivalent to providing a predictive coverage guarantee for CP.

Validity. We present an alternative derivation of the standard conformal prediction theory in Theorem 1, based on the unified result in Theorem 2.

Proof of Theorem 1 via the unified framework. Recall that $U = \wr Z \wr$. As have assumed the data is exchangeable, the true conditional distribution is given by

$$P_{Z|U=\wr z \wr} = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \delta_{z_\sigma},$$

i.e., after observing the unordered bag of data $\wr Z \wr = \wr z \wr$, each permutation of z is equally likely. Since $P_{Z|U} = Q_{Z|U}$, this proves (recalling Remark 2) that $\mathbb{P}\{p \leq \alpha\} \leq \alpha$. □

This template—providing a predictive coverage guarantee by bounding the Type I error of the associated p-value—is used in all variants of conformal prediction in this section.

4.2 Split conformal prediction

Split conformal prediction is a widely-used variant of standard conformal prediction where a data split is used in order to facilitate computation. Split CP was proposed by Papadopoulos et al. (2002) and studied in Lei et al. (2018), among many others. The problem setting that we consider here is precisely as in the last subsection: we assume exchangeable $Z_1, \ldots, Z_{n+1}$. (We note that split variants also exist for all conformal methods that will be discussed in the coming subsections, but for simplicity, we only study it as a variant of standard CP.)

4.2.1 Method and theory

We first partition the dataset $(Z_1, \ldots, Z_{n+1})$ into two parts, written as $Z_{(0)} = (Z_1, \ldots, Z_{n_0})$, and $Z_{(1)} = (Z_{n_0+1}, \ldots, Z_{n+1})$. Similarly, for any $z \in \mathcal{Z}^{n+1}$, we will write $z_{(0)} = (z_1, \ldots, z_{n_0})$, and $z_{(1)} = (z_{n_0+1}, \ldots, z_{n+1})$. The idea of split conformal prediction to train a model on $Z_{(0)}$, then compute scores for all remaining data points in order to build a prediction set for the test point. Concretely, we will work with a score function of the form

$$\mathsf{s}((x, y), z_{(0)}).$$

This accommodates choices of the form $\mathsf{s}((x,y), z_{(0)}) = \ell(y, \hat{f}(x))$, for $\hat{f} = \mathcal{A}(z_{(0)})$, as in (1). Notice that $\mathcal{A}$ here is not required to treat its input data points symmetrically. The split CP set at coverage level $1 - \alpha$ is then given by

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : \mathsf{s}((X_{n+1}, y), Z_{(0)}) \leq \text{Quantile}_{(1-\alpha)(1+\frac{1}{n_1})} \left(\frac{1}{n_1} \sum_{i=n_0+1}^{n} \delta_{\mathsf{s}(Z_i, Z_{(0)})} \right) \right\}, \quad (11)$$

where we write $n_1 = n - n_0$. Split CP is known to have training-conditional coverage under exchangeability.

Theorem 3 (Vovk et al. 2005). *If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then the split CP set given in (11) satisfies*

$$\mathbb{P} \left\{ Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid Z_{(0)} \right\} \geq 1 - \alpha,$$

almost surely. In particular, this implies $\mathbb{P} \left\{ Y_{n+1} \in \mathcal{C}(X_{n+1}) \right\} \geq 1 - \alpha$.

4.2.2 View from the unified framework

We now describe how split CP fits into the unified conformal framework.

Choices of U and $Q_{Z|U}$. Unlike in standard CP where we simply take $U = \wr Z \wr$, here we will use a different choice of partial information U: for split CP, we define $U = (\wr Z \wr, Z_{(0)})$. (Notice that an equivalent choice would be $U = (\wr Z_{(1)} \wr, Z_{(0)})$, since this would contain the same information about the data Z.) Overloading notation again, we write the score for any $u = (\wr z \wr, z_{(0)})$ as

$$\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, z_{(0)}).$$

That is, the score assigned to z depends only on comparing z_{n+1} (the last observation) to the first portion $z_{(0)}$ of the training data. Next we define

$$Q_{Z|U=(\wr z \wr, z_{(0)})} = \frac{1}{(n_1 + 1)!} \sum_{\sigma \in \mathcal{S}_{n_1+1}} \delta_{(z_{(0)}, [z_{(1)}]_\sigma)}.$$

This is the empirical distribution on all permutations of z consistent with $z_{(0)}$—that is, this distribution places equal weight on each vector of the form $(z_{(0)}, [z_{(1)}]_\sigma)$, which preserves the first n_0 entries of z (i.e., $z_{(0)}$) in their original order, and allows the remaining $n_1 + 1$ entries (i.e., $z_{(1)}$) to be permuted arbitrarily.

P-value. Under these choices, the p-value in (6) is

$$p = \frac{1}{(n_1 + 1)!} \sum_{\sigma \in \mathcal{S}_{n_1+1}} \mathbb{1} \left\{ \mathsf{s}([Z_{(1)}]_{\sigma(n_1+1)}, Z_{(0)}) \geq \mathsf{s}(Z_{n+1}, Z_{(0)}) \right\}$$

$$= \frac{1}{n_1 + 1} \sum_{i=n_0+1}^{n+1} \mathbb{1} \left\{ \mathsf{s}(Z_i, Z_{(0)}) \geq \mathsf{s}(Z_{n+1}, Z_{(0)}) \right\},$$

where the second line follows similarly to the simplification used in standard conformal.

To see that this corresponds to split CP, recall that by Lemma 1,

$$p > \alpha \iff \mathsf{s}(Z_{n+1}, Z_{(0)}) \leq \text{Quantile}_{1-\alpha}\left(\frac{1}{n_1 + 1} \sum_{i=n_0+1}^{n+1} \delta_{\mathsf{s}(Z_i, Z_{(0)})}\right).$$

Next we will need an elementary fact about quantiles: for any values $v_1, \ldots, v_{m+1} \in \mathbb{R}$,

$$v_{m+1} \leq \text{Quantile}_{1-\alpha}(v_1, \ldots, v_{m+1}) \iff v_{m+1} \leq \text{Quantile}_{(1-\alpha)(1+1/m)}(v_1, \ldots, v_m).$$

This implies that

$$p > \alpha \iff \mathsf{s}(Z_{n+1}, Z_{(0)}) \leq \text{Quantile}_{(1-\alpha)(1+\frac{1}{n_1})}\left(\frac{1}{n_1} \sum_{i=n_0+1}^{n} \delta_{\mathsf{s}(Z_i, Z_{(0)})}\right).$$

The right-hand side above can be directly seen to be equivalent to the event $Y_{n+1} \in \mathcal{C}(X_{n+1})$ for the split CP set in (11).

Validity. We prove Theorem 3 using the unified result in Theorem 2.

Proof of Theorem 3 via the unified framework. Since $Z = (Z_1, \ldots, Z_{n+1})$ is exchangeable, it also holds that $Z_{(1)} \mid Z_{(0)} = z_{(0)}$ is exchangeable, for almost every $z_{(0)}$. Fix any $z_{(0)}$ for which this conditional exchangeability holds. Let P_Z denote the distribution of $Z \mid Z_{(0)} = z_{(0)}$, and recalling that $U = (\langle Z \rangle, Z_{(0)})$, write $P_{Z,U}$ as the induced joint distribution on (Z, U). Then we can calculate the true conditional distribution as

$$P_{Z \mid U = (\langle z \rangle, z_{(0)})} = \frac{1}{(n_1 + 1)!} \sum_{\sigma \in \mathcal{S}_{n_1+1}} \delta_{(z_{(0)}, [z_{(1)}]_\sigma)},$$

because the distribution of $Z_{(1)}$ conditional on $Z_{(0)} = z_{(0)}$ is exchangeable. We therefore see that $P_{Z \mid U} = Q_{Z \mid U}$, which proves (recalling Remark 2) that $\mathbb{P}\{p \leq \alpha\} \leq \alpha$. $\square$

4.3 Weighted conformal prediction

Weighted conformal prediction (WCP) works in a problem setting where $Z_1, \ldots, Z_n$ are i.i.d. from F, whereas the test point Z_{n+1} is drawn independently from G.[1] While F and G are generally unknown, we additionally assume that we have knowledge of the *distribution shift* relating the likelihood of G to F,

$$w^*(x, y) = \frac{\mathsf{d}G}{\mathsf{d}F}(x, y). \tag{12}$$

As our knowledge of this shift might be only approximate, we will work with a user-specified weight function w that approximates w^*. WCP was introduced by Tibshirani et al. (2019), who focused on covariate shift (where $w^*(x, y)$ depends only on x). It has also been studied in other settings, such as label shift (Podkopaev and Ramdas, 2021), causal inference (Lei and Candès, 2021), and survival analysis (Candès et al., 2023).

---

[1]The results for WCP hold under a weaker assumption than what is stated here, called *weighted exchangeability*, as defined in Tibshirani et al. (2019). For simplicity, we work with independent data in our treatment here, but we note that the unified framework can also encompass the weighted exchangeable setting.

4.3.1 Method and theory

Fix any score function of the form $\mathsf{s}((x, y), \lfloor z \rfloor)$ (as in standard CP, the score here treats the dataset z symmetrically). For arbitrary $y \in \mathcal{Y}$, define as usual $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$, and scores

$$S_i^y = \mathsf{s}(Z_i^y, \lfloor Z^y \rfloor), \quad i \in [n+1].$$

Given the user-specified weight function $w : \mathcal{Z} \to (0, \infty)$, we define the WCP set at coverage level $1 - \alpha$ by

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \text{Quantile}_{1-\alpha}\left(\frac{\sum_{i=1}^n w(Z_i) \cdot \delta_{S_i^y} + w(X_{n+1}, y) \cdot \delta_{S_{n+1}^y}}{\sum_{i=1}^n w(Z_i) + w(X_{n+1}, y)} \right) \right\}. \quad (13)$$

WCP is known to have exact coverage if $w = w^*$, and approximate coverage if $w \approx w^*$.

Theorem 4. *For independent $Z_1, \ldots, Z_n \sim F$ and $Z_{n+1} \sim G$, where $w^* = \mathsf{d}G/\mathsf{d}F$, the WCP set in* (13) *satisfies the following:*

(a) (Tibshirani et al., 2019). *If $w = w^*$, then $\mathbb{P}\left\{ Y_{n+1} \in \mathcal{C}(X_{n+1}) \right\} \geq 1 - \alpha$.*

(b) (Lei and Candès, 2021). *Assuming $\mathbb{E}_F\left[w(X, Y)\right] < \infty$, and defining a normalized version of w by $\bar{w}(x, y) = w(x, y)/\mathbb{E}_F\left[w(X, Y)\right]$, we have*

$$\mathbb{P}\left\{ Y_{n+1} \in \mathcal{C}(X_{n+1}) \right\} \geq 1 - \alpha - \frac{1}{2}\mathbb{E}_F\left[|\bar{w}(X, Y) - w^*(X, Y)| \right].$$

4.3.2 View from the unified framework

We describe how WCP fits into the unified conformal framework.

Choices of U and $Q_{Z|U}$. Define $U = \lfloor Z \rfloor$, and overloading notation again, write the score function for $u = \lfloor z \rfloor$ as

$$\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, \lfloor z \rfloor).$$

Next define

$$Q_{Z|U=\lfloor z \rfloor} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} w(z_{\sigma(n+1)}) \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} w(z_{\sigma(n+1)})},$$

which is the weighted empirical distribution that places weight proportional to $w(z_{\sigma(n+1)})$ on each permutation z_σ.

P-value. Under these choices, the p-value in (6) is

$$\begin{aligned} p &= \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} w(Z_{\sigma(n+1)}) \cdot \mathbb{1}\left\{ \mathsf{s}(Z_{\sigma(n+1)}, \lfloor Z \rfloor) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rfloor) \right\}}{\sum_{\sigma \in \mathcal{S}_{n+1}} w(Z_{\sigma(n+1)})} \\ &= \frac{\sum_{i=1}^{n+1} w(Z_i) \cdot \mathbb{1}\left\{ \mathsf{s}(Z_i, \lfloor Z \rfloor) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rfloor) \right\}}{\sum_{i=1}^{n+1} w(Z_i)}, \end{aligned}$$

where the second line follows from the fact that, for each i, there are $n!$ many permutations $\sigma \in \mathcal{S}_{n+1}$ with $\sigma(n+1) = i$. To see that this corresponds to WCP, note that by Lemma 1,

$$p > \alpha \iff \mathsf{s}(Z_{n+1}, \wr Z \wr) \leq \text{Quantile}_{1-\alpha}\left(\frac{\sum_{i=1}^{n+1} w(Z_i) \cdot \delta_{\mathsf{s}(Z_i, \wr Z \wr)}}{\sum_{i=1}^{n+1} w(Z_i)}\right).$$

The right-hand side above can be directly seen to be equivalent to the event $Y_{n+1} \in \mathcal{C}(X_{n+1})$, for the WCP set in (13).

Validity. We prove Theorem 4 using the unified result in Theorem 2.

Proof of Theorem 4 via the unified framework. Recalling $U = \wr Z \wr$, define Q_U to be the distribution of $\wr Z \wr$ when $Z \sim H$, where

$$\mathsf{d}H(z) = \frac{\sum_{i=1}^{n+1} \bar{w}(z_i)}{n+1} \cdot \mathsf{d}F^{n+1}(z).$$

For intuition, note that we can interpret this as a mixture: we sample an index $i \in [n+1]$ uniformly at random, then sample Z_i from a distribution $F \circ \bar{w}$, defined by $\mathsf{d}(F \circ \bar{w})(x, y) = w(x, y) \cdot \mathsf{d}F(x, y)$, whereas the remaining n data points are drawn as $Z_j \overset{\text{iid}}{\sim} F$, for $j \neq i$. With $Q_{Z,U} = Q_{Z|U} \times Q_U$, as usual, we examine the corresponding marginal Q_Z on Z: by definition of Q_U and $Q_{Z|U}$, we have

$$
\begin{aligned}
Q_Z(A) &= \mathbb{E}_{Z \sim H}\left[\frac{\sum_{\sigma \in \mathcal{S}_{n+1}} w(Z_{\sigma(n+1)}) \cdot \mathbb{1}\{Z_\sigma \in A\}}{\sum_{\sigma \in \mathcal{S}_{n+1}} w(Z_{\sigma(n+1)})}\right] \\
&= \mathbb{E}_{Z \sim F^{n+1}}\left[\frac{\sum_{i=1}^{n+1} \bar{w}(Z_i)}{n+1} \cdot \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} w(Z_{\sigma(n+1)}) \cdot \mathbb{1}\{Z_\sigma \in A\}}{\sum_{\sigma \in \mathcal{S}_{n+1}} w(Z_{\sigma(n+1)})}\right] \\
&= \mathbb{E}_{Z \sim F^{n+1}}\left[\frac{\sum_{\sigma \in \mathcal{S}_{n+1}} \bar{w}(Z_{\sigma(n+1)})}{(n+1)!} \cdot \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} \bar{w}(Z_{\sigma(n+1)}) \cdot \mathbb{1}\{Z_\sigma \in A\}}{\sum_{\sigma \in \mathcal{S}_{n+1}} \bar{w}(Z_{\sigma(n+1)})}\right] \\
&= \mathbb{E}_{Z \sim F^{n+1}}\left[\frac{1}{(n+1)!}\sum_{\sigma \in \mathcal{S}_{n+1}} \bar{w}(Z_{\sigma(n+1)}) \cdot \mathbb{1}\{Z_\sigma \in A\}\right] \\
&= \mathbb{E}_{Z \sim F^{n+1}}\left[\bar{w}(Z_{n+1}) \cdot \mathbb{1}\{Z \in A\}\right] \\
&= \mathbb{E}_{Z \sim F^n \times (F \circ \bar{w})}\left[\mathbb{1}\{Z \in A\}\right],
\end{aligned}
$$

where the next-to-last line holds by exchangeability of F^{n+1}.

Meanwhile, the true distribution on Z is $P_Z = F^n \times G = F^n \times (F \circ w^*)$. Theorem 2 thus gives the Type I error bound,

$$\mathbb{P}\{p \leq \alpha\} \leq \alpha + \mathsf{d}_{\text{TV}}(P_{S,T}, Q_{S,T}) \leq \alpha + \mathsf{d}_{\text{TV}}(P_Z, Q_Z),$$

where (recalling information monotonicity, in Remark 1) the last step uses the fact that (S, T) is a function of Z.

15

To complete the proof, we only need to bound $\mathrm{d}_{\mathrm{TV}}(P_Z, Q_Z)$. We have

$$
\begin{aligned}
\mathrm{d}_{\mathrm{TV}}(P_Z, Q_Z) &= \mathrm{d}_{\mathrm{TV}}(F^n \times (F \circ w^*), F^n \times (F \circ \bar{w})) \\
&= \mathrm{d}_{\mathrm{TV}}(F \circ w^*, F \circ \bar{w}) \\
&= \frac{1}{2} \int_{\mathcal{Z}} |w^*(x,y) - \bar{w}(x,y)| \ \mathrm{d}F(x,y) \\
&= \frac{1}{2} \mathbb{E}_F \left[|\bar{w}(X,Y) - w^*(X,Y)| \right],
\end{aligned}
$$

where the next-to-last line applies the well-known L^1 representation for TV distance. This recovers the result in part (b). In the special case where $w = w^*$ (and so $\bar{w} = w^*$), we have $\mathbb{E}_F \left[|\bar{w}(X,Y) - w^*(X,Y)| \right] = 0$, which recovers the result in part (a). $\qquad\square$

4.4   Nonexchangeable conformal prediction

Nonexchangeable conformal prediction (NexCP) works in a setting where $Z_1, \ldots, Z_{n+1}$ are drawn from an arbitrary joint distribution, but there is some underlying structure (e.g., the samples are indexed by time or space) allowing us to posit a guess as to which of $Z_1, \ldots, Z_n$ are "closer" in distribution to Z_{n+1}. NexCP was introduced by Barber et al. (2023).

4.4.1   Method and theory

Fix any score function $\mathsf{s}((x,y), z)$, which assigns a value to a data point (its first argument) based on an *ordered* dataset (its second argument). Notice that this score function is not required to be symmetric in z, and accommodates choices of the form $\mathsf{s}((x,y), z) = \ell(y, \hat{f}(x))$, for a fitted model $\hat{f} = \mathcal{A}(z)$, as in (1), where the algorithm $\mathcal{A}$ does not need to treat its input data points symmetrically.

Given user-specified weights $w_1, \ldots, w_{n+1} \geq 0$, with $\sum_{k=1}^{n+1} w_k = 1$, we sample an index

$$
K \sim \sum_{k=1}^{n+1} w_k \cdot \delta_k. \tag{14}
$$

We then define, for arbitrary $y \in \mathcal{Y}$, the augmented dataset $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$ as usual, and scores

$$
S_i^{y,K} = \mathsf{s}(Z_i^y, (Z^y)^K), \quad i \in [n+1],
$$

where for any $z \in \mathbb{R}^{n+1}$ and $k \in [n+1]$, we use z^k to denote the vector obtained by swapping the entries in positions k and $n+1$ of z, i.e.,

$$
z^k = (z_1, \ldots, z_{k-1}, z_{n+1}, z_{k+1}, \ldots, z_n, z_k) \in \mathcal{Z}^{n+1}
$$

(and, to handle the case $k = n+1$, we simply define $z^{n+1} = z$). We now define the NexCP set at coverage level $1 - \alpha$ by

$$
\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^{y,K} \leq \mathrm{Quantile}_{1-\alpha}\left(\sum_{i=1}^{n+1} w_i \cdot \delta_{S_i^{y,K}} \right) \right\}. \tag{15}
$$

NexCP is known to provide a coverage guarantee for arbitrary joint distributions.

Theorem 5 (Barber et al. 2023). *For $Z_1, \ldots, Z_{n+1}$ of arbitrary joint distribution, if $w_{n+1} \geq w_i$ for all $i \in [n+1]$, then the NexCP set in (15) satisfies*

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha - \sum_{k=1}^{n+1} w_k \cdot d_{\text{TV}}(g(Z), g(Z^k)),$$

where the function $g : \mathcal{Z}^{n+1} \to \mathbb{R}^{n+1}$ is defined as

$$g(z) = \big(\mathsf{s}(z_1, z), \ldots, \mathsf{s}(z_{n+1}, z)\big).$$

In particular, by information monotonicity, this implies that $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha - \sum_{k=1}^{n+1} w_k \cdot d_{\text{TV}}(Z, Z^k)$.

This result can be interpreted as a guaranteeing that coverage holds at approximately the desired level $1 - \alpha$, as long as the dataset Z—or rather, its corresponding vector of scores, $g(Z)$—is approximately exchangeable.

4.4.2 View from the unified framework

We describe how NexCP fits into the unified conformal framework.

Choices of U and $Q_{Z|U}$. Set $U = Z^K$, where K is drawn as in (14). Notice that the choice of the partial information U in this setting contains additional information beyond the bag of data. Overloading notation, we write the score function as

$$\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, u).$$

Next define

$$Q_{Z|U=u} = \sum_{k=1}^{n+1} w_k \cdot \delta_{u^k},$$

placing weight w_k on each swapped vector u^k.

P-value. Under these choices, the p-value in (6) is

$$p = \sum_{k=1}^{n+1} w_k \cdot \mathbb{1}\left\{\mathsf{s}((Z^K)_k, Z^K) \geq \mathsf{s}(Z_{n+1}, Z^K)\right\}.$$

On the event $K \in [n]$, then we calculate

$$p = \sum_{k \in [n]\setminus\{K\}} w_k \cdot \mathbb{1}\left\{\mathsf{s}(Z_k, Z^K) \geq \mathsf{s}(Z_{n+1}, Z^K)\right\} + w_K + w_{n+1} \cdot \mathbb{1}\left\{\mathsf{s}(Z_K, Z^K) \geq \mathsf{s}(Z_{n+1}, Z^K)\right\},$$

using the fact that $(Z^K)_k = Z_k$ for all $k \in [n]\setminus\{K\}$, whereas $(Z^K)_K = Z_{n+1}$ and $(Z^K)_{n+1} = Z_K$. Since $w_{n+1} \geq w_K$ by assumption, we therefore have

$$p \leq \sum_{k \in [n]\setminus\{K\}} w_k \cdot \mathbb{1}\left\{\mathsf{s}(Z_k, Z^K) \geq \mathsf{s}(Z_{n+1}, Z^K)\right\} + w_{n+1} + w_K \cdot \mathbb{1}\left\{\mathsf{s}(Z_K, Z^K) \geq \mathsf{s}(Z_{n+1}, Z^K)\right\}$$

$$= \sum_{k=1}^{n+1} w_k \cdot \mathbb{1}\left\{\mathsf{s}(Z_k, Z^K) \geq \mathsf{s}(Z_{n+1}, Z^K)\right\} =: p^*.$$

If instead $K = n + 1$, then by definition we simply have $p = p^*$. Based on Lemma 1, we can see that $Y_{n+1} \in \mathcal{C}(X_{n+1})$ for NexCP (15) is equivalent to $p^* > \alpha$. Since we have shown that $p \leq p^*$, note that $\mathbb{P}\{p^* \leq \alpha\} \leq \mathbb{P}\{p \leq \alpha\}$ and thus a Type I error bound on p will translate into one for the NexCP p-value p^*.

Validity. We prove Theorem 5 using the unified result in Theorem 2.

Proof of Theorem 5 via the unified framework. By definition of $Q_{Z|U}$, observe that

$$\mathsf{t}_\alpha(U) = \mathrm{Quantile}_{1-\alpha}\left(\sum_{k=1}^{n+1} w_k \cdot \delta_{\mathsf{s}((U^k)_{n+1},U)}\right) = \mathrm{Quantile}_{1-\alpha}\left(\sum_{k=1}^{n+1} w_k \cdot \delta_{\mathsf{s}(U_k,U)}\right),$$

and thus, by definition of g,

$$\mathsf{t}_\alpha(U) = \mathrm{Quantile}_{1-\alpha}\left(\sum_{k=1}^{n+1} w_k \cdot \delta_{g(U)_k}\right).$$

Similarly, for any $k \in [n+1]$,

$$\mathsf{s}(U^k, U) = \mathsf{s}((U^k)_{n+1}, U) = \mathsf{s}(U_k, U) = g(U)_k.$$

For convenience, define

$$h_k(v) = \left(v_k, \mathrm{Quantile}_{1-\alpha}\left(\sum_{k'=1}^{n+1} w_{k'} \cdot \delta_{v_{k'}}\right)\right).$$

By construction, then, for any $k \in [n+1]$, we have

$$(\mathsf{s}(U^k, U), \mathsf{t}_\alpha(U)) = h_k(g(U)).$$

Therefore, for $U = Z^K$, we have

$$(S, T) = (\mathsf{s}(Z, U), \mathsf{t}_\alpha(U)) = (\mathsf{s}(U^K, U), \mathsf{t}_\alpha(U)) = h_K(g(U)) = h_K(g(Z^K)).$$

Since by construction we sample $K \sim \sum_{k=1}^{n+1} w_k \cdot \delta_k$ independently of Z, note that under the joint distribution $(Z, U) \sim P_{Z,U}$, we have

$$P_{S,T} = \sum_{k=1}^{n+1} w_k \cdot P_{h_k(g(Z^k))},$$

where $P_{h_k(g(Z^k))}$ denotes the distribution of $h_k(g(Z^k))$ induced by $Z \sim P_Z$.

Next, we choose to define $Q_U = P_Z$, and now we need to calculate the distribution $Q_{S,T}$. Since $Q_{Z|U} = \sum_{k=1}^{n+1} w_k \cdot \delta_{U^k}$, we can therefore calculate that under $Q_{Z|U}$,

$$(S, T) = (\mathsf{s}(Z, U), \mathsf{t}_\alpha(U)) \sim \sum_{k=1}^{n+1} w_k \cdot \delta_{(\mathsf{s}(U^k,U),\mathsf{t}_\alpha(U))} = \sum_{k=1}^{n+1} w_k \cdot \delta_{h_k(g(U))}.$$

18

In other words, we have shown

$$Q_{S,T} = \sum_{k=1}^{n+1} w_k \cdot Q_{h_k(g(U))},$$

where $Q_{h_k(g(U))}$ denotes the distribution of $h_k(g(U))$ induced by $U \sim Q_U$.

We therefore calculate

$$
\begin{aligned}
\mathrm{d}_{\mathrm{TV}}(P_{S,T}, Q_{S,T}) &= \mathrm{d}_{\mathrm{TV}}\left(\sum_{k=1}^{n+1} w_k \cdot P_{h_k(g(Z^k))}, \sum_{k=1}^{n+1} w_k \cdot Q_{h_k(g(U))} \right) \\
&\leq \sum_{k=1}^{n+1} w_k \cdot \mathrm{d}_{\mathrm{TV}}\left(P_{h_k(g(Z^k))}, Q_{h_k(g(U))} \right) \\
&\leq \sum_{k=1}^{n+1} w_k \cdot \mathrm{d}_{\mathrm{TV}}\left(P_{g(Z^k)}, Q_{g(U)} \right) \\
&= \sum_{k=1}^{n+1} w_k \cdot \mathrm{d}_{\mathrm{TV}}\left(P_{g(Z^k)}, P_{g(Z)} \right),
\end{aligned}
$$

where the third step uses the information monotonicity property of total variation distance (recall Remark 1), while the last step uses $Q_U = P_Z$. Therefore,

$$\mathbb{P}\left\{ p \leq \alpha \right\} \leq \alpha + \mathrm{d}_{\mathrm{TV}}(P_{S,T}, Q_{S,T}) \leq \alpha + \sum_{k=1}^{n+1} w_k \cdot \mathrm{d}_{\mathrm{TV}}(P_{g(Z^k)}, P_{g(Z)}),$$

which completes the proof. $\qquad\square$

4.5   Randomly-localized conformal prediction

Localized conformal methods work in a setting where $Z_1, \ldots, Z_{n+1}$ are exchangeable but we seek to go beyond the marginal guarantee $\mathbb{P}\left\{ Y_{n+1} \in \mathcal{C}(X_{n+1}) \right\} \geq 1 - \alpha$ offered by standard CP. A stronger coverage guarantee that holds conditionally on the value of the test feature, as in $\mathbb{P}\left\{ Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} = x \right\} \geq 1 - \alpha$, would be a highly desirable (and aspirational) goal. Unfortunately, when the covariate X is continuously distributed, well-known negative results by Vovk (2012); Lei and Wasserman (2014) show this is not possible to achieve in a distribution-free manner, excepting trivial (infinitely wide) prediction intervals.

Researchers have therefore turned to developing procedures with *approximate* conditional coverage over small neighborhoods in the feature space $\mathcal{X}$. Guan (2023) introduced localized-conformal prediction (LCP), which computes a prediction set by assigning higher weight to data points that lie closer to the test point X_{n+1}. LCP gives marginal coverage, and leads to approximate conditional coverage in practice, but does not have accompanying finite-sample theory. To address this, Hore and Barber (2023) derived a variant called randomly-localized conformal prediction (RLCP), which we study in this subsection.

4.5.1 Method and theory

Fix any score function of the form $\mathsf{s}((x,y),\wr z\wr)$. For arbitrary $y \in \mathcal{Y}$, we now define as usual $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$, and scores

$$S_i^y = \mathsf{s}(Z_i^y, \wr Z^y \wr), \quad i \in [n+1],$$

The user specifies a "localizing" kernel $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, e.g., in the case $\mathcal{X} = \mathbb{R}^d$, we can use the Gaussian kernel $H(x, x') = e^{-\|x-x'\|_2^2/2\sigma^2}/(2\pi\sigma^2)^{d/2}$. We assume that, for any $x \in \mathcal{X}$, $H(x, \cdot)$ is a density (relative to some base measure), with $H(x, x) > 0$. We then sample

$$\tilde{X}_{n+1} \mid X_{n+1} \sim H(X_{n+1}, \cdot),$$

which is then used to define weights in the RLCP set at coverage level $1 - \alpha$, defined by[2]

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \mathrm{Quantile}_{1-\alpha}\left(\frac{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1}) \cdot \delta_{S_i^y}}{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1})} \right) \right\}. \qquad (16)$$

As $\tilde{X}_{n+1}$ is sampled to lie near X_{n+1}, we can view it as a proxy for the test point, ensuring that RLCP will generally place higher weights on data points near the test point.

RLCP is known to have marginal coverage under exchangeability, and an approximate form of conditional coverage.[3]

Theorem 6 (Hore and Barber 2023). *If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then the RLCP set in (16) satisfies a marginal coverage guarantee $\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1})\right\} \geq 1 - \alpha$. Moreover, for any $B \subseteq \mathcal{X}$, RLCP satisfies an approximate conditional coverage guarantee,*

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in B\right\} \geq 1 - \alpha$$
$$- \frac{\inf_{\epsilon > 0}\left\{\mathbb{P}\{\|X_{n+1} - \tilde{X}_{n+1}\| > \epsilon\} + \mathbb{P}\{X_{n+1} \in \mathrm{bd}_{2\epsilon}(B)\}\right\}}{\mathbb{P}\left\{X_{n+1} \in B\right\}},$$

where $\|\cdot\|$ is any norm on $\mathcal{X}$, and where the r-boundary of a set B is defined as

$$\mathrm{bd}_r(B) = \left\{x \in B : \inf_{x' \in B^c} \|x - x'\| \leq r\right\}.$$

To interpret the second result, we see that as long as the kernel H is strongly localizing (i.e., $\|X_{n+1} - \tilde{X}_{n+1}\|$ is likely to be small), an approximate coverage guarantee holds when we condition on $X_{n+1} \in B$ for any set B that satisfies some regularity conditions (so that its boundary does not contain too much mass). In fact, the proof given later will show that the unified framework allows us to establish a strictly stronger result:

$$\mathbb{P}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in B\right\} \geq 1 - \alpha - \frac{\mathbb{E}\left[\mathrm{Var}(\mathbb{1}\left\{X_{n+1} \in B\right\} \mid \tilde{X}_{n+1})\right]}{\mathbb{P}\left\{X_{n+1} \in B\right\}}.$$

---

[2]Hore and Barber (2023) describe a split implementation of RLCP, whereas what we describe here is the corresponding full-data version.

[3]Hore and Barber (2023) also derive a result on robustness to unknown covariate shift; this is possible to obtain via the unified framework as well but for a concise presentation we do not address this here.

4.5.2 View from the unified framework

We describe how RLCP fits into the unified conformal framework.

Choices of U and $Q_{Z|U}$. Define $U = (\wr Z \wr, \tilde{X}_{n+1})$. Note that U here contains additional information beyond the bag of data. Overloading notation, we write the score function for $u = (\wr z \wr, \tilde{x})$ as

$$\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, \wr z \wr).$$

Next define, for $z = ((x_1, y_1), \ldots, (x_{n+1}, y_{n+1}))$,

$$Q_{Z|U=(\wr z \wr, \tilde{x})} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} H(x_{\sigma(n+1)}, \tilde{x}) \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} H(x_{\sigma(n+1)}, \tilde{x})}.$$

This is the weighted empirical distribution which places weight proportional to $H(x_{\sigma(n+1)}, \tilde{x})$ on each z_σ (i.e., higher weight if $\tilde{X}_{n+1} = \tilde{x}$ has a higher likelihood given $X_{n+1} = x_{\sigma(n+1)}$).

P-value. Under these choices, the p-value in (6) is

$$p = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} H(X_{\sigma(n+1)}, \tilde{X}_{n+1}) \cdot \mathbb{1}\left\{\mathsf{s}(Z_{\sigma(n+1)}, \wr Z \wr) \geq \mathsf{s}(Z_{n+1}, \wr Z \wr)\right\}}{\sum_{\sigma \in \mathcal{S}_{n+1}} H(X_{\sigma(n+1)}, \tilde{X}_{n+1})}$$

$$= \frac{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1}) \cdot \mathbb{1}\left\{\mathsf{s}(Z_i, \wr Z \wr) \geq \mathsf{s}(Z_{n+1}, \wr Z \wr)\right\}}{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1})},$$

where the second line follows from a similar simplification as in the WCP case. To see that this corresponds to RLCP, observe that by Lemma 1,

$$p > \alpha \iff \mathsf{s}(Z_{n+1}, \wr Z \wr) \leq \text{Quantile}_{1-\alpha}\left(\frac{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1}) \cdot \delta_{\mathsf{s}(Z_i, \wr Z \wr)}}{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1})}\right).$$

The right-hand side above can be directly seen to be equivalent to the event $Y_{n+1} \in \mathcal{C}(X_{n+1})$ for the RLCP set in (16).

Validity. We prove Theorem 6 using the unified result in Theorem 2.

Proof of Theorem 6 via the unified framework. Note that the marginal guarantee is implied by the approximate conditional guarantee: we can simply take $B = \mathcal{X}$ and let $\epsilon \to \infty$. Thus from this point on, we focus on the conditional guarantee.

First let P_Z^* be the distribution of $Z = ((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}))$ (which we assume is exchangeable), and let P_Z be the distribution of Z conditional on the event $X_{n+1} \in B$. In other words, the conditional coverage probability can be written as

$$\mathbb{P}_{P_Z^*}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid X_{n+1} \in B\right\} = \mathbb{P}_{P_Z}\left\{Y_{n+1} \in \mathcal{C}(X_{n+1})\right\}.$$

Let $u = (\wr z \wr, \tilde{x})$, and write $z = ((x_1, y_1), \ldots, (x_{n+1}, y_{n+1}))$. Then

$$P_{Z|U=(\wr z \wr, \tilde{x})} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} H(x_{\sigma(n+1)}, \tilde{x}) \cdot \mathbb{1}\left\{x_{\sigma(n+1)} \in B\right\} \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} H(x_{\sigma(n+1)}, \tilde{x}) \cdot \mathbb{1}\left\{x_{\sigma(n+1)} \in B\right\}},$$

by using the fact that $\mathsf{d}P_Z(z_\sigma)/\mathsf{d}P_Z(z) = \mathbb{1}\left\{x_{\sigma(n+1)} \in B\right\}$ (for $x_{n+1} \in B$), and recalling that $\tilde{X}_{n+1}$ is drawn from a distribution with density $H(X_{n+1}, \cdot)$. Therefore,

$$
\begin{aligned}
\mathsf{d}_{\mathrm{TV}}&(P_{Z|U=(\wr z\backslash,\tilde{x})}, Q_{Z|U=(\wr z\backslash,\tilde{x})}) \\
&\leq \sum_{\sigma \in \mathcal{S}_{n+1}} \left(\frac{H(x_{\sigma(n+1)}, \tilde{x}) \cdot \mathbb{1}\left\{x_{\sigma(n+1)} \in B\right\}}{\sum_{\sigma' \in \mathcal{S}_{n+1}} H(x_{\sigma'(n+1)}, \tilde{x}) \cdot \mathbb{1}\left\{x_{\sigma'(n+1)} \in B\right\}} - \frac{H(x_{\sigma(n+1)}, \tilde{x})}{\sum_{\sigma' \in \mathcal{S}_{n+1}} H(x_{\sigma'(n+1)}, \tilde{x})} \right)_+ \\
&= \sum_{i=1}^{n+1} \left(\frac{H(x_i, \tilde{x}) \cdot \mathbb{1}\left\{x_i \in B\right\}}{\sum_{i'=1}^{n+1} H(x_{i'}, \tilde{x}) \cdot \mathbb{1}\left\{x_{i'} \in B\right\}} - \frac{H(x_i, \tilde{x})}{\sum_{i'=1}^{n+1} H(x_{i'}, \tilde{x})} \right)_+ \\
&= \sum_{i=1}^{n+1} H(x_i, \tilde{x}) \cdot \mathbb{1}\left\{x_i \in B\right\} \cdot \left(\frac{1}{\sum_{i'=1}^{n+1} H(x_{i'}, \tilde{x}) \cdot \mathbb{1}\left\{x_{i'} \in B\right\}} - \frac{1}{\sum_{i'=1}^{n+1} H(x_{i'}, \tilde{x})} \right) \\
&= 1 - \frac{\sum_{i=1}^{n+1} H(x_i, \tilde{x}) \cdot \mathbb{1}\left\{x_i \in B\right\}}{\sum_{i=1}^{n+1} H(x_i, \tilde{x})} = \frac{\sum_{i=1}^{n+1} H(x_i, \tilde{x}) \cdot \mathbb{1}\left\{x_i \notin B\right\}}{\sum_{i=1}^{n+1} H(x_i, \tilde{x})}.
\end{aligned}
$$

Next let P_U be the marginal distribution of $U = (\wr Z\backslash, \tilde{X}_{n+1})$ under $P_{Z,U}$, and let $Q_U = P_U$. To bound $\mathsf{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U})$, as $P_U = Q_U$, we have $\mathsf{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}) = \mathbb{E}_{P_U}\left[\mathsf{d}_{\mathrm{TV}}(P_{Z|U}, Q_{Z|U})\right]$, and so

$$
\begin{aligned}
\mathsf{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}) &\leq \mathbb{E}_{P_Z}\left[\frac{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1}) \cdot \mathbb{1}\left\{X_i \notin B\right\}}{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1})} \right] \\
&= \mathbb{E}_{P_Z^*}\left[\frac{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1}) \cdot \mathbb{1}\left\{X_i \notin B\right\}}{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1})} \;\middle|\; X_{n+1} \in B \right],
\end{aligned}
$$

by definition of P_Z as compared to P_Z^*. We can rewrite this as

$$
\begin{aligned}
\mathsf{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}) &= \frac{\mathbb{E}_{P_Z^*}\left[\frac{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1}) \cdot \mathbb{1}\left\{X_i \notin B\right\}}{\sum_{i=1}^{n+1} H(X_i, \tilde{X}_{n+1})} \cdot \mathbb{1}\left\{X_{n+1} \in B\right\} \right]}{\mathbb{P}_{P_Z^*}\left\{X_{n+1} \in B\right\}} \\
&= \frac{\mathbb{E}_{P_Z^*}\left[\mathbb{P}_{P_Z^*}\left\{X_{n+1} \notin B \;\middle|\; \wr Z\backslash, \tilde{X}_{n+1}\right\} \cdot \mathbb{P}_{P_Z^*}\left\{X_{n+1} \in B \;\middle|\; \wr Z\backslash, \tilde{X}_{n+1}\right\} \right]}{\mathbb{P}_{P_Z^*}\left\{X_{n+1} \in B\right\}},
\end{aligned}
$$

where the last step holds using similar calculations to the above to characterize $Z \mid U$, this time applied to $X_{n+1} \mid U$. In particular, using the variance of a Bernoulli random variable, we can see that

$$
\begin{aligned}
\mathsf{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}) &= \frac{\mathbb{E}_{P_Z^*}\left[\mathrm{Var}_{P_Z^*}\left(\mathbb{1}\left\{X_{n+1} \in B\right\} \mid \wr Z\backslash, \tilde{X}_{n+1}\right) \right]}{\mathbb{P}_{P_Z^*}\left\{X_{n+1} \in B\right\}} \\
&\leq \frac{\mathbb{E}_{P_Z^*}\left[\mathrm{Var}_{P_Z^*}\left(\mathbb{1}\left\{X_{n+1} \in B\right\} \mid \tilde{X}_{n+1}\right) \right]}{\mathbb{P}_{P_Z^*}\left\{X_{n+1} \in B\right\}},
\end{aligned}
$$

where the inequality holds by the law of total variance.

To complete the proof, we need to show that this last expression results in the bound claimed in the theorem. First, on the event $\tilde{X}_{n+1} \in B^c \cup \mathrm{bd}_\epsilon(B)$,

$$\mathbb{P}_{P_Z^*}\left\{X_{n+1} \in B \mid \tilde{X}_{n+1}\right\} \leq \mathbb{P}_{P_Z^*}\left\{X_{n+1} \in \mathrm{bd}_{2\epsilon}(B) \text{ or } \|X_{n+1} - \tilde{X}_{n+1}\| > \epsilon \mid \tilde{X}_{n+1}\right\}.$$

On the other hand, on the event $\tilde{X}_{n+1} \in B \setminus \mathrm{bd}_\epsilon(B)$,

$$\mathbb{P}_{P_Z^*}\left\{X_{n+1} \notin B \mid \tilde{X}_{n+1}\right\} \leq \mathbb{P}_{P_Z^*}\left\{\|X_{n+1} - \tilde{X}_{n+1}\| > \epsilon \mid \tilde{X}_{n+1}\right\}.$$

Using $p(1-p) \leq \min\{p, 1-p\}$, and combining the above bounds,

$$\mathrm{Var}_{P_Z^*}\left(\mathbb{1}\left\{X_{n+1} \in B\right\} \mid \tilde{X}_{n+1}\right)$$
$$\leq \mathbb{P}_{P_Z^*}\left\{X_{n+1} \in \mathrm{bd}_{2\epsilon}(B) \text{ or } \|X_{n+1} - \tilde{X}_{n+1}\| > \epsilon \mid \tilde{X}_{n+1}\right\}$$
$$\leq \mathbb{P}_{P_Z^*}\left\{X_{n+1} \in \mathrm{bd}_{2\epsilon}(B) \mid \tilde{X}_{n+1}\right\} + \mathbb{P}_{P_Z^*}\left\{\|X_{n+1} - \tilde{X}_{n+1}\| > \epsilon \mid \tilde{X}_{n+1}\right\}.$$

and marginalizing over $\tilde{X}_{n+1}$ completes the proof. $\qquad\square$

4.6 Generalized weighted conformal prediction

In two of the settings considered above—standard CP and WCP—the basic premise is that, after conditioning on the unordered bag of data $\lfloor Z \rceil$, we know the distribution of the data Z itself. For standard CP, under exchangeability, this distribution is uniform over the $(n+1)!$ possible permutations of data vectors, while for WCP the distribution over permutations is weighted due to distribution shift. A natural generalization is to consider a setting where we allow the likelihood ratio between any two permutations of the same dataset to be arbitrary, but still known; this leads to a generalization of WCP, studied by Prinster et al. (2024).

Specifically, suppose that $Z \in \mathcal{Z}^{n+1}$ has joint density $f : \mathcal{Z}^{n+1} \to \mathbb{R}_+$, with respect to any exchangeable base measure. While f itself may be unknown, suppose that for any $z \in \mathcal{Z}^{n+1}$, we can calculate the ratio

$$f(z_\sigma)/f(z).$$

Returning to the earlier examples, under exchangeability this ratio is simply equal to 1, while in the WCP setting (12) this ratio is equal to $w^*(z_{\sigma(n+1)})/w^*(z_{n+1})$.

As studied in Nair and Janson (2023); Prinster et al. (2024) (building on earlier work by Fannjiang et al. (2022)), an important application of this generalized view is the problem of feedback covariate shift (FCS), where data is collected sequentially. At time t, the observed data from past times $1, \ldots, t-1$ determines a distribution from which the next covariate X_t is sampled—for example, this can arise if our aim is to identify regions of feature space that are most likely to lead to, say, higher response values. To make this concrete, suppose

$$X_1 \sim g_{X,1}, \; Y_1 \mid X_1 \sim P_{Y|X=X_1},$$

at time $t = 1$, and then for each time $t \geq 2$,

$$X_t \mid ((X_i, Y_i))_{i=1}^{t-1} \sim \mathcal{A}\big(((X_i, Y_i))_{i=1}^{t-1}\big),$$
$$Y_t \mid \big(((X_i, Y_i))_{i=1}^{t-1}, X_t\big) \sim P_{Y|X=X_t},$$

23

where $\mathcal{A}$ is some algorithm mapping the observed data $((X_i, Y_i))_{i=1}^{t-1}$ to a density for the next covariate X_t, relative to some base measure on $\mathcal{X}$. We can calculate the likelihood ratio by

$$\frac{f(z_\sigma)}{f(z)} = \frac{g_{X,1}(x_{\sigma(1)}) \cdot \prod_{t=2}^{n+1} \left[\mathcal{A}(z_{\sigma(1)}, \ldots, z_{\sigma(t-1)})\right](x_{\sigma(t)})}{g_{X,1}(x_1) \cdot \prod_{t=2}^{n+1} \left[\mathcal{A}(z_1, \ldots, z_{t-1})\right](x_t)}, \tag{17}$$

which does not depend on the unknown conditional distribution $P_{Y|X}$. In other words, here $f(z_\sigma)/f(z)$ is known, and depends only on choices made by the user.

4.6.1 Method and theory

For arbitrary $y \in \mathcal{Y}$, we define as usual $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$, and scores

$$S_i^y = \mathsf{s}(Z_i^y, \wr Z^y \wr), \quad i \in [n+1],$$

for a score function of the form $\mathsf{s}((x, y), \wr z \wr)$. Next define[4]

$$W_i^y = \sum_{\sigma:\sigma(n+1)=i} \frac{f(Z_\sigma^y)}{f(Z^y)},$$

noting that by our assumptions above, each of these ratios is known even if the joint density f is unknown. The generalized WCP set at coverage level $1 - \alpha$ is then defined by

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \mathrm{Quantile}_{1-\alpha}\left(\frac{\sum_{i=1}^{n+1} W_i^y \cdot \delta_{S_i^y}}{\sum_{i=1}^{n+1} W_i^y} \right) \right\}. \tag{18}$$

(To reiterate, both standard CP and WCP can be seen as special cases of this construction.) When $f(z_\sigma)/f(z)$ is known exactly, for any z and σ, generalized WCP has exact coverage.

Theorem 7 (Prinster et al. 2024). *If $Z \in \mathcal{Z}^{n+1}$ has density f with respect to an exchangeable base measure, then the generalized WCP set in* (18) *satisfies $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha$.*

4.6.2 View from the unified framework

We describe how generalized WCP fits into the unified conformal framework.

Choices of U and $Q_{Z|U}$. Define $U = \wr Z \wr$, and overloading notation again, write the score function for $u = \wr z \wr$ as

$$\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, \wr z \wr).$$

Next define

$$Q_{Z|U=\wr z \wr} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} f(z_\sigma) \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} f(z_\sigma)}.$$

Note that this can equivalently be written as

$$Q_{Z|U=\wr z \wr} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} f(z_\sigma)/f(z) \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} f(z_\sigma)/f(z)}.$$

which can be computed by the analyst, since each ratio $f(z_\sigma)/f(z)$ is assumed to be known.

---

[4]For simplicity, we will assume that the density f is always positive so that the ratio is well-defined, but the method can be modified in straightforward ways to avoid this condition.

P-value. Under these choices, the p-value in (6) is

$$p = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_\sigma)/f(Z) \cdot \mathbb{1}\left\{\mathsf{s}(Z_{\sigma(n+1)}, \lfloor Z \rfloor) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rfloor)\right\}}{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_\sigma)/f(Z)}$$

$$= \frac{\sum_{i=1}^{n+1} \sum_{\sigma:\sigma(n+1)=i} f(Z_\sigma)/f(Z) \cdot \mathbb{1}\left\{\mathsf{s}(Z_i, \lfloor Z \rfloor) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rfloor)\right\}}{\sum_{i=1}^{n+1} \sum_{\sigma:\sigma(n+1)=i} f(Z_\sigma)/f(Z)}$$

$$= \frac{\sum_{i=1}^{n+1} W_i^{Y_{n+1}} \cdot \mathbb{1}\left\{\mathsf{s}(Z_i, \lfloor Z \rfloor) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rfloor)\right\}}{\sum_{i=1}^{n+1} W_i^{Y_{n+1}}}.$$

To see that this corresponds to generalized WCP, observe that by Lemma 1,

$$p > \alpha \iff \mathsf{s}(Z_{n+1}, \lfloor Z \rfloor) \leq \mathrm{Quantile}_{1-\alpha}\left(\frac{\sum_{i=1}^{n+1} W_i^{Y_{n+1}} \cdot \delta_{\mathsf{s}(Z_i, \lfloor Z \rfloor)}}{\sum_{i=1}^{n+1} W_i^{Y_{n+1}}}\right).$$

The right-hand side above can be directly seen to be equivalent to the event $Y_{n+1} \in \mathcal{C}(X_{n+1})$, for the generalized WCP set in (18).

Validity. We prove Theorem 7 using the unified result in Theorem 2.

Proof of Theorem 7 via the unified framework. Recall that $U = \lfloor Z \rfloor$. Since we have assumed that Z has joint density f with respect to some exchangeable base measure, this means that $P_{Z|U} = Q_{Z|U}$, by construction. Therefore (recalling Remark 2), we have $\mathbb{P}\{p \leq \alpha\} \leq \alpha$. □

5   Extensions of the unified framework

In this section, we develop several extensions of the unified framework. These extensions will be used to derive some of the new results on conformal given in the next section. All proofs for this section are deferred until the appendix.

5.1   Monte Carlo p-values

For our first extension, we develop Monte Carlo versions of Theorem 2, which can be useful in settings where the computational cost of conformal procedures is prohibitive. Specifically, in using Monte Carlo, when computing the p-value p we can avoid integration with respect to $Q_{Z|U}$ (which may be computationally hard) by instead sampling with respect to $Q_{Z|U}$ (which may be easier).

5.1.1   Simple Monte Carlo

Given the conditional distribution $Q_{Z|U}$, the p-value p defined in (6) for the unified method requires integration with respect to $Q_{Z|U}$: we can rewrite this p-value as

$$p = \int_{\mathcal{Z}} \mathbb{1}\left\{\mathsf{s}(z, U) \geq \mathsf{s}(Z, U)\right\} \, \mathrm{d}Q_{Z|U}(z).$$

To avoid calculating such an integral (i.e., to avoid calculating a probability with respect to $Q_{Z|U}$), we can instead take a simple Monte Carlo approximation,

$$\tilde{p} = \frac{1}{M+1} \sum_{m=0}^{M} \mathbb{1}\left\{ \mathsf{s}(Z^{(m)}, U) \geq \mathsf{s}(Z, U) \right\}, \tag{19}$$

where $Z^{(1)}, \ldots, Z^{(M)}$ are i.i.d. samples from $Q_{Z|U}$, with $(Z^{(1)}, \ldots, Z^{(M)}) \perp\!\!\!\perp Z \mid U$, and where we let $Z^{(0)} = Z$ for notational simplicity. (Observe that, since $Q_{Z|U}$ is supported on the set $\{z \in \mathcal{Z} : \lfloor z \rceil = h(U) = \lfloor Z \rceil\}$, each $Z^{(m)}$ must be equal to some permutation of Z.)

Next we give our guarantee for this Monte Carlo p-value $\tilde{p}$, similar to that in Theorem 2 for the original p-value p in (6). This result is proved in Appendix A.2.

Theorem 8. *Suppose $(Z, U) \sim P_{Z,U}$. Then the Monte Carlo p-value defined in (19) satisfies*

$$\mathbb{P}\{\tilde{p} \leq \alpha\} \leq \alpha + \inf_{Q_U} \mathrm{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}),$$

where the infimum is taken over all distributions Q_U on U, and where $Q_{Z,U} = Q_{Z|U} \times Q_U$.

Remark 3. The guarantee above, with Type I error inflation bounded by $\mathrm{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U})$, is exactly the same as that derived in Remark 1, which is a corollary to Theorem 2. Indeed the original result of Theorem 2, with Type I error inflation bounded by $\mathrm{d}_{\mathrm{TV}}(P_{S,T}, Q_{S,T})$, is strictly stronger: as (S, T) may potentially contain much less information than (Z, U), this TV distance can be substantially smaller. An analogous refinement can also be derived for the Monte Carlo case, but we defer this to the appendix; see Appendix A.4.

Remark 4. Just as in Remark 2, if $Q_{Z|U} = P_{Z|U}$ holds almost surely (i.e., in implementing the Monte Carlo method we correctly specify the conditional distribution of $Z \mid U$), then by taking $Q_U = P_U$ we obtain exact Type I error control,

$$\mathbb{P}\{\tilde{p} \leq \alpha\} \leq \alpha.$$

5.1.2 Importance sampling

When sampling from $Q_{Z|U}$ is itself a difficult task, we can also employ alternate Monte Carlo schemes based on importance sampling. Let $R_{Z|U}$ be any (conditional) proposal distribution. We will assume that $Q_{Z|U}, R_{Z|U}$ are absolutely continuous with respect to each other, almost surely. Given a number of samples $M \geq 1$, and letting $Z^{(0)} = Z$ as before, we now study an importance sampling scheme to define the p-value:

$$\tilde{p} = \frac{\sum_{m=0}^{M} \frac{\mathrm{d}Q_{Z|U}(Z^{(m)})}{\mathrm{d}R_{Z|U}(Z^{(m)})} \cdot \mathbb{1}\left\{ \mathsf{s}(Z^{(m)}, U) \geq \mathsf{s}(Z, U) \right\}}{\sum_{m=0}^{M} \frac{\mathrm{d}Q_{Z|U}(Z^{(m)})}{\mathrm{d}R_{Z|U}(Z^{(m)})}}, \tag{20}$$

where now $Z^{(1)}, \ldots, Z^{(M)}$ are i.i.d. samples from $R_{Z|U}$, and $(Z^{(1)}, \ldots, Z^{(M)}) \perp\!\!\!\perp Z \mid U$. In the Monte Carlo literature, this type of construction is often called *self-normalized* importance sampling. Notice that this can be viewed as a generalization of the Monte Carlo procedure above, as the p-value $\tilde{p}$ in (19) can be obtained by simply letting $R_{Z|U} = Q_{Z|U}$.

Next we present our theoretical guarantee for the importance sampling p-value $\tilde{p}$. This result is proved in Appendix A.3.

Theorem 9. *Suppose $(Z, U) \sim P_{Z,U}$. Then the importance sampling p-value defined in (20) satisfies*

$$\mathbb{P}\{\tilde{p} \le \alpha\} \le \alpha + \inf_{Q_U} d_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}),$$

where the infimum is taken over all distributions Q_U on U, and where $Q_{Z,U} = Q_{Z|U} \times Q_U$.

Remark 5. Analogous to Remark 3 for the simple Monte Carlo setting, refinements of this result are again possible, with an inflation in Type I error bounded by a TV distance between the (joint) distributions of scores and threshold, but we omit the details.

Remark 6. As in Remark 4 for simple Monte Carlo, if $Q_{Z|U} = P_{Z|U}$, almost surely, then we obtain exact Type I error control,

$$\mathbb{P}\{\tilde{p} \le \alpha\} \le \alpha.$$

5.2 Unnormalized $Q_{Z|U}$

For our second extension, we will develop an unnormalized version of Theorem 2. Returning to the construction of the p-value, we now allow $Q_{Z|U=u}$ to be measure on $\mathcal{Z}$, for each $u \in \mathcal{U}$; the difference here is that $Q_{Z|U=u}$ is no longer required to be a distribution—that is, we may have $Q_{Z|U=u}(\mathcal{Z}) \ne 1$. However, as before, we assume that $Q_{Z|U=u}$ is supported on the finite set $\{z \in \mathcal{Z}^{n+1} : \lfloor z \rceil = h(u)\}$, for each u.

We then define

$$p = p(Z, U) \quad \text{where} \quad p(z, u) = \int_{\mathcal{Z}} \mathbb{1}\{s(z', u) \ge s(z, u)\} \, dQ_{Z|U=u}(z'). \tag{21}$$

If $Q_{Z|U=u}$ is a distribution for each $u \in \mathcal{U}$, then note this is the same as the original p-value constructed in (6). In the general case, we still refer to (21) as a p-value. We will also define a generalized threshold

$$t_\alpha(u) = \inf\left\{s \in \mathbb{R} : \int_{\mathcal{Z}} \mathbb{1}\{s(z, u) > s\} \, dQ_{Z|U=u}(z) \le \alpha\right\},$$

which reduces to $t_\alpha(u) = \mathrm{Quantile}_{1-\alpha}(Q_{S|U=u})$ when $Q_{Z|U=u}$ is a distribution. Lastly, given a distribution Q_U on $\mathcal{U}$, we write $Q_{Z,U} = Q_{Z|U} \times Q_U$ for the measure on $(Z, U) \in \mathcal{Z} \times \mathcal{U}$ that is defined as[5]

$$Q_{Z,U}(A) = \mathbb{E}_{Q_U}\left[\int_{\mathcal{Z}} \mathbb{1}\{(z, U) \in A\} \, dQ_{Z|U}(z)\right],$$

for measurable $A \subseteq \mathcal{Z} \times \mathcal{U}$. In the normalized case, where $Q_{Z|U}$ is chosen to be a conditional distribution, note that this is simply the joint distribution with marginal Q_U and conditional $Q_{Z|U}$, as before.

We are now ready to present the unnormalized generalization of Theorem 2. Its proof is similar to that of Theorem 2, and is given in Appendix A.5.

---

[5]Formally, in order for this to be well-defined, we require that the function $u \mapsto Q_{Z|U=u}(\{z : (z, u) \in A\})$ is measurable, for each measurable $A \subseteq \mathcal{Z} \times \mathcal{U}$; this condition is naturally satisfied in the normalized case, where $Q_{Z|U}$ is a conditional distribution.

Theorem 10. *Suppose $(Z, U) \sim P_{Z,U}$. Let $P_{S,T}$ be the induced joint distribution on $(S, T) = (\mathsf{s}(Z, U), \mathsf{t}_\alpha(U))$. Then the p-value defined in (21) satisfies*

$$\mathbb{P}\{p \leq \alpha\} \leq \alpha + \inf_{Q_U} \sup_{A \subseteq \mathbb{R} \times \mathbb{R}} \left\{ P_{S,T}(A) - Q_{S,T}(A) \right\},$$

where the infimum is taken over all distributions Q_U on U, and where $Q_{S,T}$ denotes the joint measure on $(S, T) = (\mathsf{s}(Z, U), \mathsf{t}_\alpha(U)) \in \mathbb{R} \times \mathbb{R}$ induced by the joint measure $Q_{Z,U} = Q_{Z|U} \times Q_U$ on $(Z, U) \in \mathcal{Z} \times \mathcal{U}$.

When $Q_{Z|U}$ is a conditional distribution, this reduces to the bound in Theorem 2, since then $Q_{S,T}$ is a distribution and so we have $\mathrm{d}_{\mathrm{TV}}(P_{S,T}, Q_{S,T}) = \sup_{A \subseteq \mathbb{R} \times \mathbb{R}} \{P_{S,T}(A) - Q_{S,T}(A)\}$. Before moving on, we make one further remark about information monotonicity.

Remark 7. Just as before in Remark 1, information monotonicity implies simpler but coarser versions of the bound in Theorem 10. For example, because (S, T) is a function of (Z, U), we have the weaker bound

$$\mathbb{P}\{p \leq \alpha\} \leq \alpha + \inf_{Q_U} \sup_{A \subseteq \mathcal{Z} \times \mathcal{U}} \left\{ P_{Z,U}(A) - Q_{Z,U}(A) \right\}.$$

6 Special cases: new results

We now present new results which can be derived as special cases of the unified framework for conformal prediction. We will focus on developing extensions of the conformal approaches and theory in Section 4; our intention to provide a flavor of the types of extensions possible and not to derive an exhaustive set of new results. Some of these examples will rely on the extensions of the unified theory from Section 5.

6.1 WCP with unnormalized weights

We return to weighted conformal prediction (WCP) as considered in Section 4.3, for handling distribution shift. As before, we will assume that the training data $Z_1, \ldots, Z_n$ are i.i.d. from F, the test point is drawn independently from G, and we have access to a weight function w that approximates the likelihood ratio in (12). Recall, the WCP set (13) is defined by placing a weight on point i which is proportional to $w(Z_i)$, where the weights are normalized to sum to 1. Here we study an alternative construction, where the weights are applied without such a normalization step, and we will derive a validity guarantee using the unnormalized theory from Section 5.2.

6.1.1 Method and theory

Fix any score function $\mathsf{s}((x, y), \wr z \wr)$. Defining the data points Z_i^y and scores S_i^y exactly as in Section 4.3, the unnormalized WCP set at level $1 - \alpha$ is defined by

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \inf \left\{ s \in \mathbb{R} : \frac{\sum_{i=1}^{n+1} w(Z_i^y) \cdot \mathbb{1}\{S_i^y > s\}}{n+1} \leq \alpha \right\} \right\}. \quad (22)$$

28

To compare to the usual (normalized) WCP method, we can observe that the original WCP set (13) can equivalently be defined as

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \inf \left\{ s \in \mathbb{R} : \frac{\sum_{i=1}^{n+1} w(Z_i^y) \cdot \mathbb{1}\{S_i^y > s\}}{\sum_{i=1}^{n+1} w(Z_i^y)} \leq \alpha \right\} \right\}, \qquad (23)$$

since this infimum is simply equal to the quantile,

$$\inf \left\{ s \in \mathbb{R} : \frac{\sum_{i=1}^{n+1} w(Z_i^y) \cdot \mathbb{1}\{S_i^y > s\}}{\sum_{i=1}^{n+1} w(Z_i^y)} \leq \alpha \right\}$$

$$= \inf \left\{ s \in \mathbb{R} : \frac{\sum_{i=1}^{n+1} w(Z_i^y) \cdot \mathbb{1}\{S_i^y \leq s\}}{\sum_{i=1}^{n+1} w(Z_i^y)} \geq 1 - \alpha \right\} = \mathrm{Quantile}_{1-\alpha}\left(\frac{\sum_{i=1}^{n+1} w(Z_i^y) \cdot \delta_{S_i^y}}{\sum_{i=1}^{n+1} w(Z_i^y)} \right).$$

The only difference between the unnormalized WCP method in (22), and the original WCP method as expressed in (23), is that the former divides each weight $w(Z_i^y)$ by $n+1$ rather than $\sum_{j=1}^{n+1} w(Z_j^y)$. To motivate this, notice that for the true likelihood ratio $w^* = \mathrm{d}G/\mathrm{d}F$, we have $\mathbb{E}\left[w^*(Z_i)\right] = 1$ for each $i \in [n]$, and so the original denominator $\sum_{j=1}^{n+1} w(Z_j^y)$ has expected value approximately $n + 1$.

The unnormalized WCP method has guarantees analogous to that for WCP.

Theorem 11. *For independent $Z_1, \ldots, Z_n \sim F$ and $Z_{n+1} \sim G$, where as before $w^* = \mathrm{d}G/\mathrm{d}F$, the unnormalized WCP set in (22) satisfies the following:*

(a) *If $w = w^*$, then $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha$.*

(b) *For any w, we have $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha - \mathbb{E}_F\left[(w^*(X, Y) - w(X, Y))_+\right]$.*

To take a closer look at part (b) of this result, note also that if $\mathbb{E}_F\left[w(X, Y)\right] = 1$ (the user-specified likelihood ratio estimate w is normalized at the population level), then

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha - \frac{1}{2}\mathbb{E}_F\left[|w^*(X, Y) - w(X, Y)|\right],$$

since, for any random variable $V \in \mathbb{R}$ with $\mathbb{E}[V] = 0$, we have $\mathbb{E}[(V)_+] = \mathbb{E}[(V)_-] = \frac{1}{2}\mathbb{E}[|V|]$. This result now appears identical to the guarantee of Theorem 4 part (b), for the normalized case. But there is a subtle difference: here w must already be normalized (at the population), whereas for normalized WCP, the result always holds (since it applies to $\bar{w}$ in place of w).

6.1.2 View from the unified framework

We describe how unnormalized WCP fits into the unified conformal framework.

Choices of U and $Q_{Z|U}$. Define $U = \lfloor Z \rceil$, and write the score function for $u = \lfloor z \rceil$ as

$$\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, \lfloor z \rceil).$$

Next define

$$Q_{Z|U=\lfloor z \rceil} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} w(z_{\sigma(n+1)}) \cdot \delta_{z_\sigma}}{(n+1)!},$$

which is a measure (not necessarily a distribution), supported on permutations of z.

P-value. Under these choices, the p-value in (21) is

$$p = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} w(Z_{\sigma(n+1)}) \cdot \mathbb{1}\left\{\mathsf{s}(Z_{\sigma(n+1)}, \lfloor Z \rceil) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rceil)\right\}}{(n+1)!}$$

$$= \frac{\sum_{i=1}^{n+1} w(Z_i) \cdot \mathbb{1}\left\{\mathsf{s}(Z_i, \lfloor Z \rceil) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rceil)\right\}}{n+1},$$

where the second line follows from the fact that, for each i, there are $n!$ many permutations $\sigma \in \mathcal{S}_{n+1}$ with $\sigma(n+1) = i$. To see that this corresponds to the unnormalized WCP set, we have

$$p > \alpha \iff \mathsf{s}(Z_{n+1}, \lfloor Z \rceil) \leq \inf\left\{s \in \mathbb{R} : \frac{\sum_{i=1}^{n+1} w(Z_i) \cdot \mathbb{1}\left\{\mathsf{s}(Z_i, \lfloor Z \rceil) > s\right\}}{n+1} \leq \alpha\right\}.$$

by Lemma 3, which is an unnormalized analogue of Lemma 1, and is given in Appendix A.5. The right-hand side above can be directly seen to be equivalent to the event $Y_{n+1} \in \mathcal{C}(X_{n+1})$, for the unnormalized WCP set in (22).

Validity. We prove Theorem 11 using the unified result in Theorem 10.

Proof of Theorem 11 via the unified framework. As in the proof of Theorem 4, part (a) is a special case of part (b), so we only prove part (b). First we define Q_U as the distribution of $U = \lfloor Z \rceil$ when $Z \sim F^{n+1}$. We can then calculate the measure $Q_{Z,U}$ as

$$Q_{Z,U}(A) = \mathbb{E}_{Q_U}\left[\int_{\mathcal{Z}} \mathbb{1}\left\{(z, U) \in A\right\} \mathsf{d}Q_{Z|U}(z)\right]$$

$$= \mathbb{E}_{F^{n+1}}\left[\frac{\sum_{\sigma \in \mathcal{S}_{n+1}} w(Z_{\sigma(n+1)}) \cdot \mathbb{1}\left\{(Z_\sigma, \lfloor Z \rceil) \in A\right\}}{(n+1)!}\right]$$

$$= \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} \mathbb{E}_{F^{n+1}}\left[w(Z_{\sigma(n+1)}) \cdot \mathbb{1}\left\{(Z_\sigma, \lfloor Z \rceil) \in A\right\}\right]}{(n+1)!}$$

$$= \mathbb{E}_{F^{n+1}}\left[w(Z_{n+1}) \cdot \mathbb{1}\left\{(Z, \lfloor Z \rceil) \in A\right\}\right],$$

where the second step holds by definition of Q_U and $Q_{Z|U}$, and the last step holds since F^{n+1} is an exchangeable distribution. On the other hand, since $P_Z = F^n \times (F \circ w^*)$, we have

$$P_{Z,U}(A) = \mathbb{P}_{F^n \times (F \circ w^*)}\left\{(Z, \lfloor Z \rceil) \in A\right\} = \mathbb{E}_{F^{n+1}}\left[w^*(Z_{n+1}) \cdot \mathbb{1}\left\{(Z, \lfloor Z \rceil) \in A\right\}\right].$$

Then

$$\sup_{A \subseteq \mathcal{Z} \times \mathcal{U}} \left\{P_{Z,U}(A) - Q_{Z,U}(A)\right\} = \sup_A \mathbb{E}_{F^{n+1}}\left[(w^*(Z_{n+1}) - w(Z_{n+1})) \cdot \mathbb{1}\left\{(Z, \lfloor Z \rceil) \in A\right\}\right]$$

$$\leq \sup_A \mathbb{E}_{F^{n+1}}\left[(w^*(Z_{n+1}) - w(Z_{n+1}))_+\right].$$

Applying Theorem 10 (together with Remark 7) completes the proof. $\qquad\square$

6.2 WCP under distribution drift

When we studied WCP previously in Section 4.3, and then again in Section 6.1, our model for the data was that the training points $Z_1, \ldots, Z_n$ are drawn i.i.d. from some distribution F, while the test point Z_{n+1} is instead drawn from G. In this subsection, we will extend to a more challenging setting: we will allow for the training data to exhibit distribution drift, as well. To formalize this, suppose $Z_i \sim F_i$, $i \in [n]$, and $Z_{n+1} \sim G$, all independently. In other words, there is distribution shift between training and test sets, and in addition, the training set consists of independent but not necessarily identically distributed samples.

6.2.1 Method and theory

The method is exactly the same as before, in Section 4.3—fixing any score function of the form $\mathsf{s}((x,y), \wr z \wr)$, we will consider the WCP set in (13), with a prespecified weight function $w(x,y)$. At a high level, the idea is that WCP should achieve approximate coverage as long as the training data are approximately i.i.d. (i.e., $F_1, \ldots, F_n$ are similar each other), and w approximates the distribution shift (i.e., for each i, we have $w \approx \mathsf{d}G/\mathsf{d}F_i$).

As our next result shows, WCP has approximate coverage in this setting with training distribution drift.

Theorem 12. *For independent $Z_1 \sim F_1, \ldots, Z_n \sim F_n, Z_{n+1} \sim G$, assuming $\mathbb{E}_{\bar{F}}[w(X,Y)] < \infty$ under the mixture $\bar{F} = \frac{1}{n} \sum_{i=1}^{n} F_i$, the WCP set in (13) satisfies*

$$\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha - \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{F_i}\left[\left(w_i^*(X,Y) - \bar{w}(X,Y)\right)_+\right] + \mathsf{d}_{\mathrm{TV}}(F_i, \bar{F})\right\},$$

where we define

$$w_i^*(x,y) = \frac{\mathsf{d}G}{\mathsf{d}F_i}(x,y), \quad i \in [n],$$

and where $\bar{w}(x,y) = w(x,y)/\mathbb{E}_{\bar{F}}[w(x,y)]$.

Comparing this to the original result for WCP, we can see that the bound in Theorem 4 can be derived as a special case with $F_i = F$ and consequently $w_i^* = w$, for each i.

6.2.2 View from the unified framework

We describe how WCP under training drift fits into the unified conformal framework.

Choices of U and $Q_{Z|U}$. The unified view here is the same in Section 4.3, for WCP. For concreteness, we set $U = \wr Z \wr$, and define

$$Q_{Z|U=\wr z \wr} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} w(z_{\sigma(n+1)}) \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} w(z_{\sigma(n+1)})}.$$

P-value. Since the choice of U and $Q_{Z|U}$ is exactly the same as in Section 4.3, we again have that $p > \alpha$ if and only if $Y_{n+1} \in \mathcal{C}(X_{n+1})$, for the WCP set defined in (13).

Validity. While the construction of the method is identical to that in Section 4.3 (i.e., we are simply using the same WCP method), the proof is now different because our assumption on the data is more general, allowing for drift.

Proof of Theorem 12 via the unified framework. We choose Q_U to be the distribution of $U = \lfloor Z \rceil$ when $Z \sim H$, where

$$\mathsf{d}H(z) = \frac{1}{n+1} \sum_{i=1}^{n+1} \bar{w}(z_i) \cdot \mathsf{d}(F_1 \times \cdots \times F_n \times \bar{F})(z).$$

(Note that H is a distribution—i.e., the above definition integrates to 1—by definition of $\bar{F}$ and $\bar{w}$.) The result of Theorem 2 (together with Remark 1) tells us that

$$\mathbb{P}\{p \le \alpha\} \le \alpha + \mathrm{d}_{\mathrm{TV}}(P_{Z_{n+1},U}, Q_{Z_{n+1},U}),$$

because (S,T) can be expressed as a function of (S,U), which in turn can be expressed as a function of (Z_{n+1}, U). To be clear, here $Q_{Z_{n+1},U}$ denotes the distribution of (Z_{n+1}, U) induced by the joint distribution $Q_{Z,U} = Q_{Z|U} \times Q_U$.

We now need to bound the total variation distance, which we can express as

$$\mathrm{d}_{\mathrm{TV}}(P_{Z_{n+1},U}, Q_{Z_{n+1},U}) = \sup_{A \in \mathcal{Z} \times \mathcal{U}} \left\{ P_{Z_{n+1},U}(A) - Q_{Z_{n+1},U}(A) \right\}.$$

First, observe that by definition of Q_U and $Q_{Z|U}$,

$$Q_{Z_{n+1},U}(A) = \mathbb{P}_{Q_{Z,U}}\{(Z_{n+1}, U) \in A\}$$

$$= \mathbb{E}_H \left[\sum_{\sigma \in \mathcal{S}_{n+1}} \frac{w(Z_{\sigma(n+1)})}{\sum_{\sigma' \in \mathcal{S}_{n+1}} w(Z_{\sigma'(n+1)})} \mathbb{1}\left\{ (Z_{\sigma(n+1)}, \lfloor Z \rceil) \in A \right\} \right]$$

$$= \mathbb{E}_H \left[\sum_{i=1}^{n+1} \frac{w(Z_i)}{\sum_{j=1}^{n+1} w(Z_j)} \mathbb{1}\left\{ (Z_i, \lfloor Z \rceil) \in A \right\} \right]$$

$$= \mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} \left[\frac{\sum_{i=1}^{n+1} \bar{w}(Z_i)}{n+1} \cdot \sum_{i=1}^{n+1} \frac{w(Z_i)}{\sum_{j=1}^{n+1} w(Z_j)} \mathbb{1}\left\{ (Z_i, \lfloor Z \rceil) \in A \right\} \right]$$

$$= \mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} \left[\frac{1}{n+1} \sum_{i=1}^{n+1} \bar{w}(Z_i) \cdot \mathbb{1}\left\{ (Z_i, \lfloor Z \rceil) \in A \right\} \right],$$

where the next-to-last step uses the definition of H. Therefore, we can split the TV bound into two types of terms:

$$\mathrm{d}_{\mathrm{TV}}(P_{Z_{n+1},U}, Q_{Z_{n+1},U}) \le$$

$$\frac{1}{n+1} \sum_{i=1}^{n} \underbrace{\sup_A \left\{ P_{Z_{n+1},U}(A) - \mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} [\bar{w}(Z_i) \cdot \mathbb{1}\{(Z_i, \lfloor Z \rceil) \in A\}] \right\}}_{\text{Term } i}$$

$$+ \frac{1}{n+1} \underbrace{\sup_A \left\{ P_{Z_{n+1},U}(A) - \mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} [\bar{w}(Z_{n+1}) \cdot \mathbb{1}\{(Z_{n+1}, \lfloor Z \rceil) \in A\}] \right\}}_{\text{Term } n+1}.$$

In Appendix A.6, we will verify that

$$\text{Term } i \leq \mathbb{E}_{F_i}\left[\left(w_i^*(X,Y) - \bar{w}(X,Y)\right)_+\right] + \text{d}_{\text{TV}}(F_i, \bar{F}), \tag{24}$$

for each $i \in [n]$, and

$$\text{Term } n+1 \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{F_i}\left[\left(w_i^*(X,Y) - \bar{w}(X,Y)\right)_+\right]. \tag{25}$$

Combining these calculations, we then have

$$\text{d}_{\text{TV}}(P_{Z_{n+1},U}, Q_{Z_{n+1}|U} \times Q_U) \leq \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{F_i}\left[\left(w_i^*(X,Y) - \bar{w}(X,Y)\right)_+\right] + \text{d}_{\text{TV}}(F_i, \bar{F})\right\},$$

which completes the proof. □

6.3 RLCP with feature resampling

We now return to randomly-localized conformal prediction (RLCP), introduced in Section 4.5. Recall, in that method, after sampling a noisy version of the test feature $\tilde{X}_{n+1} \sim H(X_{n+1}, \cdot)$, the conformal set is defined by placing weight $\propto H(X_i, \tilde{X}_{n+1})$ on each data point $i \in [n+1]$, placing higher weight on data lying near the test point. Here we introduce a different version of the method: the weights will be centered at X_K for some random choice of $K \in [n+1]$, rather than at a synthetic value $\tilde{X}_{n+1}$. This version of RLCP may be computationally easier if sampling from the density $H(X_{n+1}, \cdot)$ is difficult.

Furthermore, we may prefer a version of RLCP which samples among existing observed covariate values since this may be more interpretable in certain settings. For example, if our feature points lie on a lattice, then directly sampling $\tilde{X}_{n+1}$ may not be a feasible value for the test feature—RLCP with a Gaussian kernel H leads to a value $\tilde{X}_{n+1}$ that lies off the lattice. In contrast, the modified form of RLCP presented here will always reuse an existing data point instead of sampling a new value, and therefore avoids this issue.

6.3.1 Method and theory

Fix any score function of the form $\mathsf{s}((x,y), \{z\})$, and a localizing kernel $H : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, as in RLCP in Section 4.5. Now, given the training features $X_1, \ldots, X_n$ and test feature X_{n+1}, we sample a random index $K \in [n+1]$ by

$$K \mid X_1, \ldots, X_{n+1} \sim \sum_{k=1}^{n+1}\bar{H}_{n+1,k} \cdot \delta_k,$$

where for $i, k \in [n+1]$, we introduce the weight

$$\bar{H}_{ik} = \frac{H(X_i, X_k)}{\sum_{j=1}^{n+1} H(X_i, X_j)}.$$

Defining Z^y and S_i^y, $i \in [n+1]$ as before, we then define the modified RLCP set as

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{n+1}^y \leq \text{Quantile}_{1-\alpha}\left(\frac{\sum_{i=1}^{n+1} \bar{H}_{iK} \cdot \delta_{S_i^y}}{\sum_{i=1}^{n+1} \bar{H}_{iK}} \right) \right\}. \tag{26}$$

In other words, this is the same as the prediction set for the RLCP method (16) except with X_K in place of $\tilde{X}_{n+1}$.

Like the RLCP method, this modified construction again offers marginal coverage under exchangeability. (Approximate conditional coverage results can also be established, analogous to the results of Theorem 6 for RLCP, but for brevity we omit these here.)

Theorem 13. *If $Z_1, \ldots, Z_{n+1}$ are exchangeable, then the modified RLCP set in (26) satisfies a marginal coverage guarantee,*

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

6.3.2 View from the unified framework

We describe how our modified RLCP method fits into the unified framework.

Choices of U and $Q_{Z|U}$. Analogous to RLCP in Section 4.5, we let $U = (\lfloor Z \rceil, X_K)$. Now, define $Q_{Z|U}$ as follows: at $u = (\lfloor z \rceil, x)$, with $z = ((x_1, y_1), \ldots, (x_{n+1}, y_{n+1}))$,

$$Q_{Z|U=(\lfloor z \rceil, x)} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} \frac{H(x_{\sigma(n+1)}, x)}{\sum_{k=1}^{n+1} H(x_{\sigma(n+1)}, x_k)} \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} \frac{H(x_{\sigma(n+1)}, x)}{\sum_{k=1}^{n+1} H(x_{\sigma(n+1)}, x_k)}}$$

We again write the score function as $\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, \lfloor z \rceil)$ for $u = \lfloor z \rceil$.

P-value. Under these choices, the p-value in (6) is

$$\begin{aligned}
p &= \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} \frac{H(X_{\sigma(n+1)}, X_K)}{\sum_{k=1}^{n+1} H(X_{\sigma(n+1)}, X_k)} \cdot \mathbb{1}\left\{ \mathsf{s}(Z_{\sigma(n+1)}, \lfloor Z \rceil) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rceil) \right\}}{\sum_{\sigma \in \mathcal{S}_{n+1}} \frac{H(X_{\sigma(n+1)}, X_K)}{\sum_{k=1}^{n+1} H(X_{\sigma(n+1)}, X_k)}} \\
&= \frac{\sum_{i=1}^{n+1} \frac{H(X_i, X_K)}{\sum_{k=1}^{n+1} H(X_i, X_k)} \cdot \mathbb{1}\left\{ \mathsf{s}(Z_i, \lfloor Z \rceil) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rceil) \right\}}{\sum_{i=1}^{n+1} \frac{H(X_i, X_K)}{\sum_{k=1}^{n+1} H(X_i, X_k)}} \\
&= \frac{\sum_{i=1}^{n+1} \bar{H}_{iK} \cdot \mathbb{1}\left\{ \mathsf{s}(Z_i, \lfloor Z \rceil) \geq \mathsf{s}(Z_{n+1}, \lfloor Z \rceil) \right\}}{\sum_{i=1}^{n+1} \bar{H}_{iK}}.
\end{aligned}$$

similarly to the RLCP calculations. And again as for RLCP, by Lemma 1, this corresponds to the modified RLCP set defined in (26), since

$$p > \alpha \iff \mathsf{s}(Z_{n+1}, \lfloor Z \rceil) \leq \text{Quantile}_{1-\alpha}\left(\frac{\sum_{i=1}^{n+1} \bar{H}_{iK} \cdot \delta_{\mathsf{s}(Z_i, \lfloor Z \rceil)}}{\sum_{i=1}^{n+1} \bar{H}_{iK}} \right).$$

Validity. We prove Theorem 13 using the unified result in Theorem 2.

Proof of Theorem 13 via the unified framework. Fix any $z = ((x_1, y_1), \ldots, (x_{n+1}, y_{n+1}))$. Then by definition of the distribution of the random index K, we have

$$U \mid Z = z \sim \frac{\sum_{k=1}^{n+1} H(x_{n+1}, x_k) \cdot \delta_{(\lfloor z \rceil, x_k)}}{\sum_{k=1}^{n+1} H(x_{n+1}, x_k)}.$$

Similarly, for any permutation $\sigma \in \mathcal{S}_{n+1}$, we can calculate (using $\lfloor z \rceil = \lfloor z_\sigma \rceil$) that

$$U \mid Z = z_\sigma \sim \frac{\sum_{k=1}^{n+1} H(x_{\sigma(n+1)}, x_k) \cdot \delta_{(\lfloor z \rceil, x_k)}}{\sum_{k=1}^{n+1} H(x_{\sigma(n+1)}, x_k)}.$$

Comparing these two calculations, since Z is assumed to be exchangeable, we therefore have

$$P_{Z \mid U = (\lfloor z \rceil, x_i)} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} \frac{H(x_{\sigma(n+1)}, x_i)}{\sum_{k=1}^{n+1} H(x_{\sigma(n+1)}, x_k)} \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} \frac{H(x_{\sigma(n+1)}, x_i)}{\sum_{k=1}^{n+1} H(x_{\sigma(n+1)}, x_k)}}.$$

We therefore see that $P_{Z \mid U} = Q_{Z \mid U}$, which proves that $\mathbb{P}\{p \leq \alpha\} \leq \alpha$ (recall Remark 2). $\quad\square$

6.4 Generalized WCP with a nonsymmetric score

We revisit the generalized WCP setting from Section 4.6. Recall here the dataset $Z \in \mathcal{Z}^{n+1}$ has an arbitrary joint density f, and the generalized WCP set (18) is constructed by placing weight $W_i^y = \sum_{\sigma:\sigma(n+1)=i} f(Z_\sigma^y)/f(Z^y)$ on point i. The previous treatment assumed the score function is of the form $\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, \lfloor z \rceil)$, i.e., the score assigned to z_{n+1} is not allowed to depend on the ordering of the remaining n points in z. However, in this subsection we will now see that such a restriction is actually unnecessary, from the perspective of the unified framework. Relaxing it, as we will do below, leads to a generalized WCP set that allows for *any* score function, without a symmetry assumption.

6.4.1 Method and theory

The score function now takes the form $\mathsf{s}((x, y), z)$, comparing a data point (x, y) to an *ordered* dataset $z \in \mathcal{Z}^{n+1}$ (note that s is no longer required to be symmetric in its second argument). This accommodates, e.g., a score that can depend on the time-ordering of data collected in the FCS application (with likelihood ratio in (17)). For arbitrary $y \in \mathcal{Y}$, we define as usual $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$, and scores that are now indexed by permutations,

$$S_\sigma^y = \mathsf{s}(Z_{\sigma(n+1)}^y, Z_\sigma^y), \quad \sigma \in \mathcal{S}_{n+1}.$$

The generalized WCP set at coverage level $1 - \alpha$ is defined by

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{\mathsf{id}}^y \leq \mathrm{Quantile}_{1-\alpha}\left(\frac{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_\sigma^y)/f(Z^y) \cdot \delta_{S_\sigma^y}}{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_\sigma^y)/f(Z^y)} \right) \right\}, \qquad (27)$$

where id denotes the identity permutation. This construction with a nonsymmetric score has the same coverage guarantee, assuming exact knowledge of the density ratio.

Theorem 14. *If $Z \in \mathcal{Z}^{n+1}$ has density f with respect to an exchangeable base measure, then for any score function (not necessarily symmetric), the generalized WCP set in (27) satisfies $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha$.*

6.4.2 View from the unified framework

We describe how this extension of generalized WCP fits into the unified framework.

Choices of U and $Q_{Z|U}$. As for generalized WCP in Section 4.6, we define $U = \wr Z \wr$, but we now write the score function as

$$\mathsf{s}(z, u) = \mathsf{s}(z_{n+1}, z).$$

(Compare this to $\mathsf{s}(z_{n+1}, \wr z \wr)$ for the symmetric setting of Section 4.6). Define

$$Q_{Z|U=\wr z \wr} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} f(z_\sigma) \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} f(z_\sigma)} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} f(z_\sigma)/f(z) \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} f(z_\sigma)/f(z)}.$$

P-value. Under these choices, the p-value in (6) is

$$p = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_\sigma)/f(Z) \cdot \mathbb{1}\left\{\mathsf{s}(Z_{\sigma(n+1)}, Z_\sigma) \geq \mathsf{s}(Z_{n+1}, Z)\right\}}{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_\sigma)/f(Z)}.$$

To see that this corresponds to generalized WCP, recall that by Lemma 1,

$$p > \alpha \iff \mathsf{s}(Z_{n+1}, Z) \leq \mathrm{Quantile}_{1-\alpha}\left(\frac{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_\sigma)/f(Z) \cdot \delta_{\mathsf{s}(Z_{\sigma(n+1)}, Z_\sigma)}}{\sum_{\sigma \in \mathcal{S}_{n+1}} f(Z_\sigma)/f(Z)}\right).$$

The right-hand side above can be directly seen to be equivalent to the event $Y_{n+1} \in \mathcal{C}(X_{n+1})$, for the generalized WCP set in (27).

Validity. We prove Theorem 14 using the unified result in Theorem 2.

Proof of Theorem 14 via the unified framework. The proof is identical to that of Theorem 7: since we have assumed that Z has joint density f with respect to some exchangeable base measure, this means that $P_{Z|U} = Q_{Z|U}$, by construction, and therefore (recalling Remark 2) we have $\mathbb{P}\{p \leq \alpha\} \leq \alpha$. $\square$

6.5 Generalized WCP with Monte Carlo p-values

In the generalized WCP setting once again, an important bottleneck that can arise in practice is that the generalized WCP set (18) can be very expensive to compute, due to the calculation required for each weight W_i^y: indeed, each W_i^y here is a sum over $n!$ terms, where each term requires computing the likelihood ratio $f(Z_\sigma^y)/f(Z^y)$, which can itself be highly nontrivial. Of course, computational tractability is only made worse in the extension (27) to nonsymmetric scores given in the last subsection. We present in what follows a Monte Carlo approximation

to this generalized WCP set, which will improve computational efficiency while preserving the same validity guarantee as the original construction.

In this subsection, we will specialize to a setting where the covariates have a known but potentially complex dependence:

$$X = (X_1, \ldots, X_{n+1}) \sim P_X,$$

while $Y_i \mid X$, for $i \in [n+1]$, are independent from a common conditional distribution $P_{Y\mid X}$. If f denotes the joint density of Z and f_X the corresponding density of X, then observe that under this model, for any $z = ((x_1, y_1), \ldots, (x_{n+1}, y_{n+1}))$,

$$\frac{f(z_\sigma)}{f(z)} = \frac{f_X(x_{\sigma(1)}, \ldots, x_{\sigma(n+1)})}{f_X(x_1, \ldots, x_{n+1})}, \tag{28}$$

where this simplification is due to our assumption on the distribution of $(Y_1, \ldots, Y_{n+1}) \mid X$.

6.5.1 Method and theory

For arbitrary $y \in \mathcal{Y}$, we define $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$ as usual, and just as in the last subsection, define scores that are indexed by permutations,

$$S_\sigma^y = \mathsf{s}\big(Z_{\sigma(n+1)}^y, Z_\sigma^y\big), \quad \sigma \in \mathcal{S}_{n+1}.$$

Fixing $M \geq 1$, let g_X be a user-chosen proposal density, where f_X, g_X are assumed to be absolutely continuous with respect to one another. Conditional on $\lceil X \rfloor$, we can think of g_X as inducing a distribution on permutations, which places probability

$$\frac{g_X(X_\sigma)}{\sum_{\sigma' \in \mathcal{S}_{n+1}} g_X(X_{\sigma'})}$$

on $\sigma \in \mathcal{S}_{n+1}$. Let $\sigma_1, \ldots, \sigma_M$ be i.i.d. samples from this distribution, and set $\sigma_0 = \mathsf{id}$ to be the identity permutation, for convenience. Then the importance sampling approximation to the generalized WCP set at coverage level $1 - \alpha$ is defined by

$$\mathcal{C}(X_{n+1}) = \left\{ y \in \mathcal{Y} : S_{\mathsf{id}}^y \leq \mathrm{Quantile}_{1-\alpha}\left(\frac{\sum_{m=0}^M \frac{f_X(X_{\sigma_m})/f_X(X)}{g_X(X_{\sigma_m})/g_X(X)} \cdot \delta_{S_{\sigma_m}^y}}{\sum_{m=0}^M \frac{f_X(X_{\sigma_m})/f_X(X)}{g_X(X_{\sigma_m})/g_X(X)}} \right) \right\}. \tag{29}$$

This Monte Carlo set, while computationally cheaper, still has an exact validity guarantee.

Theorem 15. *Suppose $Z \in \mathcal{Z}^{n+1}$ has density f with respect to an exchangeable base measure, with f satisfying (28). Let g_X be an arbitrary proposal density (such that f_X, g_X are absolutely continuous with respect to each other). Then for any choice $M \geq 1$, the importance sampling approximation to the generalized WCP set in (29) satisfies $\mathbb{P}\{Y_{n+1} \in \mathcal{C}(X_{n+1})\} \geq 1 - \alpha$.*

6.5.2 View from the unified framework

We describe how this importance sampling version of generalized WCP fits into the unified framework.

Choices of U, $Q_{Z|U}$, and $R_{Z|U}$. We define U and $Q_{Z|U}$ precisely as in generalized WCP for a nonsymmetric score, as in Section 6.4. Moreover, we take $R_{Z|U}$ to be the conditional distribution corresponding to joint density g_X on $X = (X_1, \ldots, X_{n+1})$, namely, for $u = \lfloor z \rceil$ where $z = ((x_1, y_1), \ldots, (x_{n+1}, y_{n+1}))$,

$$R_{Z|U=\lfloor z \rceil} = \frac{\sum_{\sigma \in \mathcal{S}_{n+1}} g_X(x_{\sigma(1)}, \ldots, x_{\sigma(n+1)}) \cdot \delta_{z_\sigma}}{\sum_{\sigma \in \mathcal{S}_{n+1}} g_X(x_{\sigma(1)}, \ldots, x_{\sigma(n+1)})}.$$

P-value. Under these choices, the importance sampling p-value in (20) is

$$\tilde{p} = \frac{\sum_{m=0}^{M} \frac{\mathsf{d}Q_{Z|U}(Z_{\sigma_m})}{\mathsf{d}R_{Z|U}(Z_{\sigma_m})} \cdot \mathbb{1}\left\{\mathsf{s}(Z_{\sigma_m(n+1)}, Z_{\sigma_m}) \geq \mathsf{s}(Z_{n+1}, Z)\right\}}{\sum_{m=0}^{M} \frac{\mathsf{d}Q_{Z|U}(Z_{\sigma_m})}{\mathsf{d}R_{Z|U}(Z_{\sigma_m})}},$$

where we are writing $Z^{(m)} = Z_{\sigma_m}$, for some permutation σ_m. Plugging in our choices of $Q_{Z|U}$ and $R_{Z|U}$, then,

$$\tilde{p} = \frac{\sum_{m=0}^{M} \frac{f_X(X_{\sigma_m})/f_X(X)}{g_X(X_{\sigma_m})/g_X(X)} \cdot \mathbb{1}\left\{\mathsf{s}(Z_{\sigma_m(n+1)}, Z_{\sigma_m}) \geq \mathsf{s}(Z_{n+1}, Z)\right\}}{\sum_{m=0}^{M} \frac{f_X(X_{\sigma_m})/f_X(X)}{g_X(X_{\sigma_m})/g_X(X)}}.$$

To see that this corresponds to generalized WCP, observe that by Lemma 1,

$$\tilde{p} > \alpha \iff \mathsf{s}(Z_{n+1}, Z) \leq \text{Quantile}_{1-\alpha}\left(\frac{\sum_{m=0}^{M} \frac{f_X(X_{\sigma_m})/f_X(X)}{g_X(X_{\sigma_m})/g_X(X)} \cdot \delta_{\mathsf{s}(Z_{\sigma_m(n+1)}, Z_{\sigma_m})}}{\sum_{m=0}^{M} \frac{f_X(X_{\sigma_m})/f_X(X)}{g_X(X_{\sigma_m})/g_X(X)}}\right).$$

The right-hand side above can be directly seen to be equivalent to the event $Y_{n+1} \in \mathcal{C}(X_{n+1})$, for the importance sampling generalized WCP set in (29).

Validity. We prove Theorem 15 using the unified result in Theorem 9.

Proof of Theorem 15 via the unified framework. The proof is similar to that of Theorem 14: since we have assumed that Z has joint density f with respect to some exchangeable base measure, this means that $P_{Z|U} = Q_{Z|U}$, by construction, and thus (now using Remark 6) we have $\mathbb{P}\{p \leq \alpha\} \leq \alpha$. $\qquad\square$

7 Discussion

In this work, we presented a unified treatment of different existing methods and theorems in the conformal prediction literature which handle departures from the standard assumption of exchangeability. At its core, our unified view of conformal methods is that they are based on inference about the joint distribution of data $Z = (Z_1, \ldots, Z_{n+1})$ (where each $Z_i = (X_i, Y_i)$), given partial information U about the data. The partial information U contains at least the information in $\lfloor Z \rceil$ (the unordered bag of points in Z) and it may contain more. We find that many existing results can be framed as follows: the user specifies a conditional distribution

$Q_{Z|U}$ as a model for the true distribution of $Z \mid U$, and then uses $Q_{Z|U}$ to form a prediction set. Our theory provides a coverage guarantee that depends on the total variation distance between $Q_{Z|U}$ and the true distribution $P_{Z|U}$.

We also showed that new conformal prediction results follow from the unified framework. Many others should also be possible to derive. For example, two existing methods, WCP and NexCP, are both weighted variants of conformal prediction but they appear quite different in their assumptions and technical motivations. Our work provides a unified explanation for why they both work; this suggests the possibility of combining the two styles of weights, and raises an open question of determining problem settings in which such a combination might be useful. As another example, the results we presented for generalized WCP assume that the likelihood ratio between different possible data permutations is known exactly; robustness results (such as those for WCP) would be important to develop, for applications where this likelihood ratio is unknown or cannot be computed exactly due to computational complexity. The unified framework not only offers a route towards developing these possible extensions, but may also enable the development of completely new methods, for settings not currently covered by the footprint of existing conformal methodology.

Acknowledgements

References

Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I. Jordan. Uncertainty sets for image classifiers using conformal prediction. In *Proceedings of the International Conference on Learning Representations*, 2021.

Anastasios N. Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *Advances in Neural Information Processing Systems*, 2023.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *Proceedings of the International Conference on Machine Learning*, 2024a.

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024b.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *Proceedings of the International Conference on Machine Learning*, 2023.

Emmanuel J. Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *Journal of the Royal Statistical Society: Series B*, 85(1):24–45, 2023.

Edgar Dobriban and Zhanran Lin. Joint coverage regions: Simultaneous confidence and prediction sets. *arXiv preprint arXiv:2303.00203*, 2023.

Clara Fannjiang, Stephen Bates, Anastasios N. Angelopoulos, Jennifer Listgarten, and Michael I. Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, 2021.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Laszlo Gyôrfi and Harro Walk. Nearest neighbor based conformal prediction. *Annales de L'Institut de Statistique de L'Université de Paris*, 63(2–3):173–190, 2019.

Matthew T. Harrison. Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika*, 99(1):57–69, 2012.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: Randomization enables local guarantees. *arXiv preprint arXiv:2310.07850*, 2023.

Rafael Izbicki, Gilson Shimizu, and Rafael B. Stern. CD-split and HPD-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87):1–32, 2022.

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

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

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B*, 83(5):911–938, 2021.

Yash Nair and Lucas Janson. Randomization tests for adaptively collected data. *arXiv preprint arXiv:2301.05365*, 2023.

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

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2021.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *Proceedings of the International Conference on Machine Learning*, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 2019.

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

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

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *Proceedings of the International Conference on Machine Learning*, 2022.

A   Additional proofs and technical results

A.1   Proof of Lemma 1

We can assume $\alpha < 1$ since otherwise the claim is trivial (we would have $\text{Quantile}_{1-\alpha}(Q) = -\infty$). Let $x_1 < \cdots < x_m \in \mathbb{R}$ be the values in the support of Q, and $p_i = \mathbb{P}_Q\{X = x_i\} > 0$, for each $i \in [m]$. Then by definition of the quantile,

$$\text{Quantile}_{1-\alpha}(Q) = \inf\{t \in \mathbb{R} : \mathbb{P}_Q\{X \leq t\} \geq 1 - \alpha\} = x_{k^*},$$

where

$$k^* = \min\left\{k \in [m] : \sum_{i \leq k} p_i \geq 1 - \alpha\right\}.$$

Therefore, for any $x \in \mathbb{R}$, if $x \leq x_{k^*}$ then

$$\mathbb{P}_Q\{X \geq x\} = \sum_{i=1}^m p_i \mathbb{1}\{x_i \geq x\} \geq \sum_{i=1}^m p_i \mathbb{1}\{x_i \geq x_{k^*}\} = \sum_{i \geq k^*} p_i = 1 - \sum_{i < k^*} p_i > \alpha,$$

since $\sum_{i < k^*} p_i < 1 - \alpha$ by definition of k^*. And similarly, if $x > x_{k^*}$ then

$$\mathbb{P}_Q\{X \geq x\} = \sum_{i=1}^m p_i \mathbb{1}\{x_i \geq x\} \leq \sum_{i=1}^m p_i \mathbb{1}\{x_i > x_{k^*}\} = \sum_{i > k^*} p_i = 1 - \sum_{i \leq k^*} p_i \leq \alpha,$$

since $\sum_{i \leq k^*} p_i \geq 1 - \alpha$ by definition of k^*.

A.2   Proof of Theorem 8

This result is a consequence of Theorem 9 for the importance sampling setting, obtained by simply choosing the proposal distribution as $R_{Z|U} = Q_{Z|U}$.

A.3   Proof of Theorem 9

Although we have presented this result as a guarantee for $\tilde{p}$, which is approximation of the original p-value p that was defined in (6) and analyzed in Theorem 2, we will now see that Theorem 9 can in fact be derived (exactly) as a special case of Theorem 2.

We will first need to reset some of our notation. Specifically, we need a new definition of partial information: instead of U on its own, the partial information will now be given by

$$\tilde{U} = \big(U, \{Z^{(0)}, \dots, Z^{(M)}\}\big),$$

and overloading notation, we will let $\mathsf{s}(Z, \tilde{U}) = \mathsf{s}(Z, U)$. We will also define the conditional distribution

$$Q_{Z|\tilde{U}=(u,\{z^{(0)},\dots,z^{(M)}\})} = \frac{\sum_{m=0}^{M} \frac{\mathrm{d}Q_{Z|U=u}}{\mathrm{d}R_{Z|U=u}}\big(z^{(m)}\big) \cdot \delta_{z^{(m)}}}{\sum_{m=0}^{M} \frac{\mathrm{d}Q_{Z|U=u}}{\mathrm{d}R_{Z|U=u}}\big(z^{(m)}\big)}.$$

Observe that the p-value $\tilde{p}$ defined in (20) is exactly the same as the original p-value p in (6) if we use the new conditional distribution $Q_{Z|\tilde{U}}$ in place of the original one $Q_{Z|U}$. Applying Theorem 2 (together with Remark 1), we therefore have the guarantee

$$\mathbb{P}\{\tilde{p} \leq \alpha\} \leq \alpha + \inf_{Q_{\tilde{U}}} \mathrm{d}_{\mathrm{TV}}(P_{Z,\tilde{U}}, Q_{Z,\tilde{U}}),$$

where the infimum is taken over all distributions $Q_{\tilde{U}}$ on $\tilde{U} = (U, \{Z^{(0)}, \dots, Z^{(M)}\})$, and where $Q_{Z,\tilde{U}} = Q_{Z|\tilde{U}} \times Q_{\tilde{U}}$.

Now we need to work with this remaining TV term. First, we can observe that $P_{Z,\tilde{U}}$ is the distribution on $(Z, \tilde{U}) = (Z, U, \{Z, Z^{(1)}, \dots, Z^{(M)}\})$ induced by the joint distribution

$$(Z, U, Z^{(1)}, \dots, Z^{(M)}) \sim P_{Z,U} \times (Q_{Z|U})^M,$$

where this notation indicates that we first draw $(Z, U) \sim P_{Z,U}$, and then conditional on this draw, we sample $Z^{(1)}, \dots, Z^{(M)} \overset{\mathrm{iid}}{\sim} Q_{Z|U}$. Next, fix any marginal Q_U on $U \in \mathcal{U}$, sample

$$(Z, U, Z^{(1)}, \dots, Z^{(M)}) \sim Q_{Z,U} \times (Q_{Z|U})^M,$$

and define $Q_{\tilde{U}}$ as the corresponding distribution of $\tilde{U} = (U, \{Z, Z^{(1)}, \dots, Z^{(M)}\})$. Note that by exchangeability of $Z, Z^{(1)}, \dots, Z^{(M)}$ under this joint distribution, the induced distribution of $(Z, \tilde{U})$ is thus equivalent to $Q_{Z,\tilde{U}} = Q_{Z|\tilde{U}} \times Q_{\tilde{U}}$. And by information monotonicity, we have

$$\mathrm{d}_{\mathrm{TV}}(P_{Z,\tilde{U}}, Q_{Z,\tilde{U}}) \leq \mathrm{d}_{\mathrm{TV}}\big(P_{Z,U} \times (Q_{Z|U})^M, Q_{Z,U} \times (Q_{Z|U})^M\big) = \mathrm{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}).$$

In other words, we have verified that

$$\inf_{Q_{\tilde{U}}} \mathrm{d}_{\mathrm{TV}}(P_{Z,\tilde{U}}, Q_{Z,\tilde{U}}) \leq \inf_{Q_U} \mathrm{d}_{\mathrm{TV}}(P_{Z,U}, Q_{Z,U}),$$

which completes the proof.

A.4 Refinement of Theorem 8

Theorem 16. *Suppose $(Z, U) \sim P_{Z,U}$. Fix any $M \geq 1$ with $\alpha(M + 1) \geq 1$. Let $P_{S,\tilde{T}}$ be the induced joint distribution on $(S, \tilde{T}) = (\mathsf{s}(Z, U), \mathsf{t}_A(U))$, where $A \sim \mathrm{Beta}_{M,\alpha}$ is independent of (Z, U), for*

$$\mathrm{Beta}_{M,\alpha} = \mathrm{Beta}\Big(\lfloor \alpha(M+1) \rfloor, \lceil (1-\alpha)(M+1) \rceil\Big).$$

Then the Monte Carlo p-value defined in (19) *satisfies*

$$\mathbb{P}\{\tilde{p} \leq \alpha\} \leq \alpha + \inf_{Q_U} \mathrm{d}_{\mathrm{TV}}(P_{S,\tilde{T}}, Q_{S,\tilde{T}}),$$

where the infimum is taken over all distributions Q_U on U, and where $Q_{S,\tilde{T}}$ denotes the joint distribution of $(S, \tilde{T}) = (\mathsf{s}(Z, U), \mathsf{t}_A(U))$ induced by drawing $(Z, U, A) \sim Q_{Z,U} \times \mathrm{Beta}_{M,\alpha}$, for $Q_{Z,U} = Q_{Z|U} \times Q_U$.

Proof. Let $Q_{S|U=u}$ denote the induced distribution on $S = \mathsf{s}(Z, u)$, under $Z \sim Q_{Z|U=u}$. Since $S \in \mathbb{R}$, we can generate samples from this distribution by using its quantile function—that is, if $V \sim \mathrm{Unif}[0, 1]$, then defining

$$S = \mathrm{Quantile}_V(Q_{S|U=u}),$$

this generates a random draw from $Q_{S|U=u}$.

Now let $V_1, \ldots, V_M \overset{\text{iid}}{\sim} \mathrm{Unif}[0, 1]$, independent of (Z, U). Defining $S_m = \mathrm{Quantile}_{V_m}(Q_{S|U})$ for each $m \in [M]$, we can therefore equivalently define the p-value $\tilde{p}$ as

$$\tilde{p} = \frac{1 + \sum_{m=1}^M \mathbb{1}\{S_m \geq \mathsf{s}(Z, U)\}}{M + 1}.$$

By construction, we can observe that, for $k = \lceil (1-\alpha)(M+1) \rceil \in [M]$,

$$\tilde{p} \leq \alpha \iff \mathsf{s}(Z, U) > S_{(k)},$$

where $S_{(1)} \leq \cdots \leq S_{(M)}$ are the order statistics of $S_1, \ldots, S_M$. Since $v \mapsto \mathrm{Quantile}_v(Q_{S|U})$ is a monotone nondecreasing function, it holds that $S_{(k)} = \mathrm{Quantile}_{V_{(k)}}(Q_{S|U})$. Therefore,

$$\tilde{p} \leq \alpha \iff \mathsf{s}(Z, U) > \mathrm{Quantile}_{V_{(k)}}(Q_{S|U}) = \mathsf{t}_{1-V_{(k)}}(U),$$

where the last step uses the definition of the threshold $\mathsf{t}_a(u) = \mathrm{Quantile}_{1-a}(Q_{S|U=u})$. Further, by construction, we have $V_{(k)} \perp (Z, U)$, and $1 - V_{(k)} \sim \mathrm{Beta}(M + 1 - k, k) = \mathrm{Beta}_{M,\alpha}$.

In other words, setting $\tilde{T} = \mathsf{t}_A(U)$ where we define $A = 1 - V_{(k)} \sim \mathrm{Beta}_{M,\alpha}$, so far we we have shown that

$$PpP_{Z,U} \times (Q_{Z|U})^M \tilde{p} \leq \alpha = \mathbb{P}_{P_{S,\tilde{T}}}\Big\{S > \tilde{T}\Big\}.$$

But by an identical argument, it also holds that

$$\mathbb{P}_{Q_{Z,U} \times (Q_{Z|U})^M}\{\tilde{p} \leq \alpha\} = \mathbb{P}_{Q_{S,\tilde{T}}}\Big\{S > \tilde{T}\Big\}.$$

We therefore have

$$\mathbb{P}_{P_{Z,U} \times (Q_{Z|U})^M} \{\tilde{p} \leq \alpha\} = \mathbb{P}_{P_{S,\tilde{T}}} \left\{ S > \tilde{T} \right\}$$
$$\leq \mathbb{P}_{Q_{S,\tilde{T}}} \left\{ S > \tilde{T} \right\} + \mathrm{d}_{\mathrm{TV}}(P_{S,\tilde{T}}, Q_{S,\tilde{T}})$$
$$= \mathbb{P}_{Q_{Z,U} \times (Q_{Z|U})^M} \{\tilde{p} \leq \alpha\} + \mathrm{d}_{\mathrm{TV}}(P_{S,\tilde{T}}, Q_{S,\tilde{T}})$$
$$\leq \alpha + \mathrm{d}_{\mathrm{TV}}(P_{S,\tilde{T}}, Q_{S,\tilde{T}}),$$

where the last step holds since, under the joint distribution $Q_{Z,U} \times (Q_{Z|U})^M$, by construction it holds that $Z, Z^{(1)}, \ldots, Z^{(M)}$ are exchangeable conditional on U (because, conditional on U, these random variables comprise $M + 1$ i.i.d. draws from $Q_{Z|U}$), and hence $\tilde{p}$ is superuniform under this distribution. $\qquad \square$

A.5 Proof of Theorem 10

We will need two additional supporting lemmas. The first is due to Lemma A1 from Harrison (2012), and we transcribe it to our notation here for concreteness.

Lemma 2 (Harrison 2012). *Let Q be a measure on a space $\mathcal{Z}$, which is supported on finitely many values. Let $\varphi : \mathcal{Z} \to \mathbb{R}$ be a measurable function. Define*[6]

$$p(z) = Q\{\varphi(Z) \geq \varphi(z)\}.$$

Then for all $\alpha \geq 0$,

$$Q\{p(Z) \leq \alpha\} \leq \alpha.$$

When Q is a distribution (i.e., $Q(\mathcal{Z}) = 1$) this lemma reduces to the standard fact in (8): a p-value constructed for distribution Q is valid (i.e., superuniform) for data drawn from Q. The next lemma is similar to Lemma 1, and its proof is similar to the proof of this lemma in Appendix A.1, hence omitted.

Lemma 3. *Let Q be a measure on $\mathbb{R}$, which is supported on finitely many values. Then for any $\alpha \geq 0$ and $x \in \mathbb{R}$,*

$$Q\{X \geq x\} > \alpha \iff x \leq \inf\{t \in \mathbb{R} : Q\{X > t\} \leq \alpha\}.$$

When Q is a distribution this lemma reduces to the result in Lemma 1. We now give the proof of Theorem 10.

Proof of Theorem 10. By Lemma 2 (applied with $Q_{Z|U=u}$ in place of Q, and the score $\mathsf{s}(\cdot, u)$ in place of the test statistic φ), we have

$$Q_{Z|U=u}\{p(Z, u) \leq \alpha\} \leq \alpha.$$

---

[6]For conciseness, throughout Appendix A.5, for the unnormalized setting we are using capital letters, such as Z, to define events—for example, $\{p(Z) \leq \alpha\}$ should be interpreted as the subset $\{z \in \mathcal{Z} : p(z) \leq \alpha\} \subseteq \mathcal{Z}$, and consequently, for a measure Q on $\mathcal{Z}$, the notation $Q\{p(Z) \leq \alpha\}$ should be interpreted as $Q(\{z \in \mathcal{Z} : p(z) \leq \alpha\})$.

Since this holds for each $u \in \mathcal{U}$, we thus have for any marginal distribution Q_U,

$$\mathbb{E}_{Q_U}\left[Q_{Z|U}\{p(Z,U) \leq \alpha\}\right] \leq \alpha.$$

Next, let $Q_{S|U=u}$ denote the measure on $\mathsf{s}(Z,u)$ induced by the measure $Q_{Z|U=u}$ on $\mathcal{Z}$. Then by Lemma 3 (applied with $Q_{S|U=u}$ in place of Q), we have for any $x \in \mathbb{R}$,

$$Q_{S|U=u}\{S \geq x\} > \alpha \iff x \leq \inf\left\{t \in \mathbb{R} : Q_{S|U=u}\{S > t\} \leq \alpha\right\}.$$

Equivalently, plugging in the definition of $Q_{S|U=u}$ and $\mathsf{t}_\alpha(u)$, we can write

$$Q_{Z|U=u}\{\mathsf{s}(Z,u) \geq x\} > \alpha \iff x \leq \mathsf{t}_\alpha(u).$$

Fixing any $z \in \mathcal{Z}$ and plugging in $x = \mathsf{s}(z,u)$, we therefore have

$$Q_{Z|U=u}\{\mathsf{s}(Z,u) \geq \mathsf{s}(z,u)\} > \alpha \iff \mathsf{s}(z,u) \leq \mathsf{t}_\alpha(u).$$

But since $p(z,u) = Q_{Z|U=u}\{\mathsf{s}(Z,u) \geq \mathsf{s}(z,u)\}$ by definition of the p-value in the unnormalized setting, given in (21), we equivalently have

$$p(z,u) > \alpha \iff \mathsf{s}(z,u) \leq \mathsf{t}_\alpha(u).$$

Therefore,

$$\mathbb{E}_{Q_U}\left[Q_{Z|U}\{\mathsf{s}(Z,U) > \mathsf{t}_\alpha(U)\}\right] = \mathbb{E}_{Q_U}\left[Q_{Z|U}\{p(Z,U) \leq \alpha\}\right] \leq \alpha,$$

and hence

$$
\begin{aligned}
\mathbb{P}_{P_{Z,U}}\{p(Z,U) \leq \alpha\} &= \mathbb{P}_{P_{Z,U}}\{\mathsf{s}(Z,U) > \mathsf{t}_\alpha(U)\} \\
&\leq \alpha + \mathbb{P}_{P_{Z,U}}\{\mathsf{s}(Z,U) > \mathsf{t}_\alpha(U)\} - \mathbb{E}_{Q_U}\left[Q_{Z|U}\{\mathsf{s}(Z,U) > \mathsf{t}_\alpha(U)\}\right],
\end{aligned}
$$

The proof is complete by noting that the difference on the right-hand side is bounded by the supremum in the theorem statement (taking $A = \{(z,u) \in \mathcal{Z} \times \mathcal{U} : \mathsf{s}(z,u) > \mathsf{t}_\alpha(u)\}$). $\qquad\square$

A.6 Additional calculations for the proof of Theorem 12

First we verify the bound (24) on Term i. By definition of P, we calculate

$$P_{Z_{n+1},U}(A) = \mathbb{P}_{P_{Z,U}}\{(Z_{n+1},U) \in A\} = \mathbb{P}_{F_1 \times \cdots \times F_n \times G}\{(Z_{n+1}, \lfloor Z \rceil) \in A\}.$$

We can bound this last expression as

$$
\begin{aligned}
&\mathbb{P}_{F_1 \times \cdots \times F_n \times G}\{(Z_{n+1}, \lfloor Z \rceil) \in A\} \\
&\leq \mathbb{P}_{F_1 \times \cdots \times F_{i-1} \times \bar{F} \times F_{i+1} \times \cdots \times F_n \times G}\{(Z_{n+1}, \lfloor Z \rceil) \in A\} + \mathrm{d}_{\mathrm{TV}}(F_i, \bar{F}) \\
&= \mathbb{P}_{F_1 \times \cdots \times F_{i-1} \times G \times F_{i+1} \times \cdots \times F_n \times \bar{F}}\{(Z_i, \lfloor Z \rceil) \in A\} + \mathrm{d}_{\mathrm{TV}}(F_i, \bar{F}) \\
&= \mathbb{E}_{F_1 \times \cdots \times F_{i-1} \times F_i \times F_{i+1} \times \cdots \times F_n \times \bar{F}}\left[w_i^*(Z_i) \cdot \mathbb{1}\{(Z_i, \lfloor Z \rceil) \in A\}\right] + \mathrm{d}_{\mathrm{TV}}(F_i, \bar{F}),
\end{aligned}
$$

where the second step holds by permuting the variables, and the last step holds by definition of w_i^*. Returning to the definition of Term i, we then calculate

$$\text{Term } i = \sup_A \left\{ P_{Z_{n+1}, U}(A) - \mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} \left[\bar{w}(Z_i) \cdot \mathbb{1}\left\{ (Z_i, \lfloor Z \rceil) \in A \right\} \right] \right\}$$

$$\leq \sup_A \left\{ \left(\mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} \left[w_i^*(Z_i) \cdot \mathbb{1}\left\{ (Z_i, \lfloor Z \rceil) \in A \right\} \right] + \mathrm{d}_{\mathrm{TV}}(F_i, \bar{F}) \right) \right.$$

$$\left. - \mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} \left[\bar{w}(Z_i) \cdot \mathbb{1}\left\{ (Z_i, \lfloor Z \rceil) \in A \right\} \right] \right\}$$

$$\leq \mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} \left[\left(w_i^*(Z_i) - \bar{w}(Z_i) \right)_+ \right] + \mathrm{d}_{\mathrm{TV}}(F_i, \bar{F})$$

$$= \mathbb{E}_{F_i} \left[\left(w_i^*(X, Y) - \bar{w}(X, Y) \right)_+ \right] + \mathrm{d}_{\mathrm{TV}}(F_i, \bar{F}),$$

which establishes the desired bound (24).

Next we verify the bound (25) on Term $n + 1$. By definition of $\bar{F}$,

$$\mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} \left[\bar{w}(Z_{n+1}) \cdot \mathbb{1}\left\{ (Z_{n+1}, \lfloor Z \rceil) \in A \right\} \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{F_1 \times \cdots \times F_n \times F_i} \left[\bar{w}(Z_{n+1}) \cdot \mathbb{1}\left\{ (Z_{n+1}, \lfloor Z \rceil) \in A \right\} \right].$$

And, for any $i \in [n]$, by definition of w_i^* we can calculate

$$P_{Z_{n+1}, U}(A) = \mathbb{P}_{F_1 \times \cdots \times F_n \times G} \left\{ (Z_{n+1}, \lfloor Z \rceil) \in A \right\}$$

$$= \mathbb{E}_{F_1 \times \cdots \times F_n \times F_i} \left[w_i^*(Z_{n+1}) \cdot \mathbb{1}\left\{ (Z_{n+1}, \lfloor Z \rceil) \in A \right\} \right],$$

and so taking an average,

$$P_{Z_{n+1}, U}(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{F_1 \times \cdots \times F_n \times F_i} \left[w_i^*(Z_{n+1}) \cdot \mathbb{1}\left\{ (Z_{n+1}, \lfloor Z \rceil) \in A \right\} \right].$$

Therefore, returning to the definition of Term $n + 1$, we have

$$\text{Term } n + 1 = \sup_A \left\{ P_{Z_{n+1}, U}(A) - \mathbb{E}_{F_1 \times \cdots \times F_n \times \bar{F}} \left[\bar{w}(Z_{n+1}) \cdot \mathbb{1}\left\{ (Z_{n+1}, \lfloor Z \rceil) \in A \right\} \right] \right\}$$

$$= \sup_A \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{F_1 \times \cdots \times F_n \times F_i} \left[w_i^*(Z_{n+1}) \cdot \mathbb{1}\left\{ (Z_{n+1}, \lfloor Z \rceil) \in A \right\} \right] \right.$$

$$\left. - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{F_1 \times \cdots \times F_n \times F_i} \left[\bar{w}(Z_{n+1}) \cdot \mathbb{1}\left\{ (Z_{n+1}, \lfloor Z \rceil) \in A \right\} \right] \right\}$$

$$= \sup_A \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{F_1 \times \cdots \times F_n \times F_i} \left[\left(w_i^*(Z_{n+1}) - \bar{w}(Z_{n+1}) \right) \cdot \mathbb{1}\left\{ (Z_{n+1}, \lfloor Z \rceil) \in A \right\} \right] \right\}$$

$$\leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{F_1 \times \cdots \times F_n \times F_i} \left[\left(w_i^*(Z_{n+1}) - \bar{w}(Z_{n+1}) \right)_+ \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{F_i} \left[\left(w_i^*(X, Y) - \bar{w}(X, Y) \right)_+ \right],$$

which completes our proof of the bound (25).