Conformal Prediction Through the Lens of Hypothesis Testing: Universality, Impossibility, and Optimality

Ryan J. Tibshirani[1], Rina Foygel Barber[2], Aaditya Ramdas[3]

[1]University of California, Berkeley, [2]University of Chicago, [3]Stanford University

Abstract

The connections between conformal prediction and permutation tests are already widely-known in the literature. Some authors motivate conformal prediction by saying that it computes a permutation p-value for the hypothesis $H_0 : Y_{n+1} = y$, and then inverts this to form a prediction set for Y_{n+1} (i.e., accepts all values y into the prediction set for which the p-value is large). In this paper, we examine an alternative view, which is less well-known: we again cast conformal prediction via the inversion of a permutation test, but for the null of exchangeability of the joint distribution of the $n + 1$ samples. This change in perspective, while simple, adheres more closely to traditional formalization in hypothesis testing, which offers several benefits. First, we use the duality between conformal sets and testing to show that foundational universality and impossibility results in the conformal prediction literature can be reproduced directly using classical hypothesis testing theory (due to Neyman, Lehmann, Scheffé, Kraft, Le Cam, and others). Furthermore, we show that an optimality result for conformal prediction can be derived using standard Neyman-Pearson theory: for any joint distribution of the covariates and response X, Y, and any sample size, the optimal method for prediction sets—delivering the most efficient set among all methods with valid coverage for exchangeable distributions—is a conformal predictor whose score is the inverse conditional density of $Y|X$.

1 Introduction

Given independent and identically distributed (i.i.d.) samples $Z_1, Z_2, \ldots, Z_{n+1}$, where each $Z_i = (X_i, Y_i)$, many powerful tools have been developed in statistics and machine learning over the years that enable us to predict Y_{n+1} based on X_{n+1}, and the first n samples $Z_1, \ldots, Z_n$. Conformal prediction goes one step further: it is a framework for producing a prediction set for Y_{n+1} based on X_{n+1} and $Z_1, \ldots, Z_n$. That is, conformal prediction is a general recipe for using X_{n+1} and $Z_1, \ldots, Z_n$ to produce a set $C_n(X_{n+1})$ which satisfies, for a prespecified error level $\alpha \in [0, 1]$,

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

Note that the probability statement in (1) is over everything that is random: $Z_1, \ldots, Z_{n+1}$, and assumes nothing about the joint distribution of each (X_i, Y_i).

Conformal prediction is typically at its best when the practitioner already has a prediction model in mind that (they believe) works well for the problem at hand, and a metric in mind which can be used to measure performance. In other words, conformal prediction does not attempt to re-solve the prediction problem from scratch, but rather, it is a framework that is fundamentally about leveraging existing predictive capabilities, and carefully re-orienting them in order to address the goal of uncertainty quantification.

To be more precise, let A be an algorithm which takes a sequence $(z_1, \ldots, z_k)$, of arbitrary length $k \geq 1$, and returns a predictor $\hat{f} = A(z_1, \ldots, z_k)$ such that $\hat{f}(x)$ is a prediction of the value of Y_i we expect to see when $X_i = x$. Also, let ℓ be a loss function, where a lower value of $\ell(y, \hat{f}(x))$ indicates that $\hat{f}(x)$ is a more accurate prediction of y. With this notation in place, conformal prediction works as follows: we first pick a *trial value* y for Y_{n+1}, and we then train A on the synthetically-augmented sequence:

$$Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y)),$$

to produce a predictor $A(Z^y)$. The loss ℓ is used to measure the quality of these predictions, i.e., we compute what are known as conformal scores,

$$s_i^y = \begin{cases} \ell(Y_i, A(Z^y)(X_i)) & \text{if } i \le n, \\ \ell(y, A(Z^y)(X_{n+1})) & \text{if } i = n+1. \end{cases} \tag{2}$$

The conformal set $C_n(X_{n+1})$ is loosely defined as the set of all trial values y for which the last score s_{n+1}^y is not "too large" relative to the other scores; when this happens the score s_{n+1}^y can be intuitively understood to "conform" to the behavior of the other scores. Precisely,

$$C_n(X_{n+1}) = \Big\{ y : s_{n+1}^y \le \text{the } \lceil (n+1)(1-\alpha) \rceil^{\text{th}} \text{ smallest of } s_1^y, \ldots, s_{n+1}^y \Big\}. \tag{3}$$

The standard guarantee backing conformal prediction is given next. It assumes an exchangeable sequence of data, which is less restrictive than assuming an i.i.d. sequence, as we have done above.

Theorem 1 (Validity)**.** *Let A be an algorithm which does not depend on the order of its input: namely, for any $z_1, \ldots, z_{n+1}$, it holds that*

$$A(z_1, \ldots, z_{n+1}) = A(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)}), \quad \text{for any } \sigma \in \mathcal{S}_{n+1}, \tag{4}$$

where $\mathcal{S}_{n+1}$ is the set of permutations of $1, \ldots, n+1$. Let ℓ be an arbitrary loss function. If $(Z_1, \ldots, Z_{n+1})$ has an exchangeable distribution, then the set defined in (2), (3) satisfies (1).

Remark 1. *The scores do not need to be defined using an algorithm and a loss function, as in (2). In fact, we can consider scores more generally defined by*

$$s_i^y = \begin{cases} s((X_i, Y_i); Z^y) & \text{if } i \le n, \\ s((X_{n+1}, y); Z^y) & \text{if } i = n+1, \end{cases} \tag{5}$$

for a score function s satisfying a generalization of (4): for any $z = (z_1, \ldots, z_{n+1})$ and any (x, y),

$$s((x, y); z) = s((x, y); z_\sigma), \quad \text{for any } \sigma \in \mathcal{S}_{n+1}, \tag{6}$$

where we denote $z_\sigma = (z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})$. Then under the symmetry property (6), the conformal set defined using (3), (5) achieves coverage (1). We will provide a proof of this in Section 2.2.

Conformal prediction was pioneered by Vladimir Vovk and collaborators in the 1990s. In the last ten or so years, there has been an explosion of interest in conformal prediction from the statistics and machine learning communities. The literature on conformal and related topics is now vast, and we do not attempt to provide an overview. Instead, we refer to the books by Vovk et al. (2022); Angelopoulos et al. (2025), which provide an excellent treatment of the foundations as well as many of the newer developments. In recent work (Barber and Tibshirani, 2026), we showed that conformal prediction and a number of recent extensions can be cast in a unified framework which uses the language of hypothesis testing. This helps us understand how and why these methods work, and fluidly leads to new combinations and generalizations.

The current paper is not about generalizations of conformal prediction, but about the basic method itself: we return to the connection between conformal prediction and hypothesis testing from first principles, and examine what the latter can teach us about the former. We find this to be fruitful in three directions. We show that well-known results about *universality* of conformal prediction (Section 3), and *impossibility* of conditional coverage (Section 5), are consequences of classical results in hypothesis testing (due to Neyman, Lehmann, Scheffé, Kraft, Le Cam, and others). Furthermore, the connections to hypothesis testing enable us to draw conclusions about *optimality* (Section 6): as an application of Neyman-Pearson theory, we show that for any joint distribution of the covariates and response X, Y, and any sample size n, the optimal method—delivering the most efficient prediction set in an average sense, among all methods which have valid coverage over exchangeable distributions—is a conformal predictor whose score function is the inverse conditional density of $Y|X$.

Before covering these results, we begin by describing the general duality between prediction sets and hypothesis tests, and explain how this connects conformal prediction to permutation tests in particular.

2 Prediction sets and hypothesis tests

The duality between confidence sets and hypothesis tests, which dates back to Neyman (1937), is a standard and widely-used result in classical statistical inference. Less well-known perhaps is the parallel duality between prediction sets and hypothesis tests. This dates back to at least Cox (1975), but related ideas were around much earlier, such as tolerance regions (Wilks, 1941) and fiducial inference (Fisher, 1935).

To fix notation, let $Z = (Z_1, \ldots, Z_{n+1})$, and consider a null hypothesis $H_0 : Z \sim P$. Let ϕ be a valid level α test for H_0, i.e., it satisfies $\phi(z) \in \{0, 1\}$ for each z, and

$$\mathbb{E}_P[\phi(Z)] = \mathbb{P}_P\{\phi(Z) = 1\} \leq \alpha.$$

We can invert this to form a prediction set for Z_{n+1} based on $Z_1, \ldots, Z_n$, with coverage under P: letting

$$C_n = \{z : \phi(Z_1, \ldots, Z_n, z) = 0\}, \tag{7}$$

it follows that $\mathbb{P}_P\{Z_{n+1} \in C_n\} = \mathbb{P}_P\{\phi(Z) = 0\} \geq 1 - \alpha$. The same construction extends to the setting in which each $Z_i = (X_i, Y_i)$, and we seek a prediction set for Y_{n+1} based on X_{n+1} and $Z_1, \ldots, Z_n$: letting

$$C_n(X_{n+1}) = \{y : \phi(Z_1, \ldots, Z_n, (X_{n+1}, y)) = 0\}, \tag{8}$$

it again follows that $\mathbb{P}_P\{Y_{n+1} \in C_n(X_{n+1})\} = \mathbb{P}_P\{\phi(Z) = 0\} \geq 1 - \alpha$. We describe a few more extensions below and gather these into a proposition for easy reference.

Proposition 1. *Let $\mathcal{P}$ be a class of distributions, and consider the composite null hypothesis $H_0 : Z \sim P$ for $P \in \mathcal{P}$. If ϕ is valid level α test for H_0,*

$$\mathbb{E}_P[\phi(Z)] \leq \alpha, \quad \textit{for all } P \in \mathcal{P},$$

then C_n in (7) and $C_n(X_{n+1})$ in (8) are both valid prediction sets with coverage $1 - \alpha$ over the class $\mathcal{P}$,

$$\mathbb{P}_P\{Z_{n+1} \in C_n\} \geq 1 - \alpha, \quad \textit{for all } P \in \mathcal{P},$$
$$\mathbb{P}_P\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha, \quad \textit{for all } P \in \mathcal{P}.$$

Finally, if ϕ is an exact test, satisfying $\mathbb{E}_P[\phi(Z)] = \alpha$ for each $P \in \mathcal{P}$, then the inequalities above become equalities: these sets have exact coverage $1 - \alpha$ across $\mathcal{P}$.

Next, we work through some examples.

2.1 Ordinary least squares regression and t-tests

To demonstrate the above duality, we turn to one of the most familiar settings in statistics: ordinary least squares regression. Let $Z = (Z_1, \ldots, Z_{n+1}) \sim P$ have i.i.d. entries, where each $Z_i = (X_i, Y_i)$ and

$$X_i \sim P_X, \quad Y_i | X_i \sim N(X_i^\mathsf{T}\beta, \sigma^2). \tag{9}$$

Following the usual notation in regression, we denote by $\mathsf{X} \in \mathbb{R}^{n \times d}$ and $\mathsf{Y} \in \mathbb{R}^n$ the covariate matrix and response vector, respectively, collected over the first n samples. Assuming that $n > d$ and $\mathsf{X}^\mathsf{T}\mathsf{X}$ is almost surely invertible, we denote the ordinary least squares coefficients by

$$\hat{\beta} = (\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}\mathsf{X}^\mathsf{T}\mathsf{Y}.$$

Furthermore, denote $\hat{\sigma}^2 = \|\mathsf{Y} - \mathsf{X}\hat{\beta}\|_2^2/(n - d)$, and define

$$T(Z) = \frac{Y_{n+1} - X_{n+1}^\mathsf{T}\hat{\beta}}{\hat{\sigma}\sqrt{1 + X_{n+1}^\mathsf{T}(\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}X_{n+1}}}. \tag{10}$$

By a standard calculation, this is a pivotal statistic, and has distribution $T(Z) \sim t_{n-d}$, where t_{n-d} denotes the t-distribution with $n - d$ degrees of freedom. Define the test

$$\phi(Z) = 1\{|T(Z)| > t_{n-d,1-\alpha/2}\},$$

where $t_{n-d,1-\alpha/2}$ denotes the level $1-\alpha/2$ quantile of t_{n-d}, and let $\mathcal{P} = \{P_{X,Y}^{n+1} : P_{X,Y} \text{ is of the form (9)}\}$. Then by construction ϕ is an exact level α test over $\mathcal{P}$, i.e., $\mathbb{E}_P[\phi(Z)] = \alpha$, for all $P \in \mathcal{P}$.

Pursuing the inversion scheme (8) leads to

$$\begin{aligned} C_n(X_{n+1}) &= \Big\{ y : \big|T(Z_1, \ldots, Z_n, (X_{n+1}, y))\big| > t_{n-d,1-\alpha/2} \Big\} \\ &= \big[X_{n+1}^\mathsf{T}\hat{\beta} - \hat{s}_{X_{n+1}} \cdot t_{n-d,1-\alpha/2},\, X_{n+1}^\mathsf{T}\hat{\beta} + \hat{s}_{X_{n+1}} \cdot t_{n-d,1-\alpha/2} \big], \end{aligned} \tag{11}$$

where we abbreviate $\hat{s}^2_{X_{n+1}} = \hat{\sigma}^2(1 + X_{n+1}^\mathsf{T}(\mathsf{X}^\mathsf{T}\mathsf{X})^{-1}X_{n+1})$. The set in (11) is the textbook ordinary least squares prediction interval for Y_{n+1} based on $X_{n+1}, Z_1, \ldots, Z_n$. Its validity comes directly from the duality between prediction sets and testing in Proposition 1: since $Y_{n+1} \in C_n(X_{n+1}) \iff \phi(Z) = 0$, it holds that $\mathbb{P}_P\{Y_{n+1} \in C_n(X_{n+1})\} = 1-\alpha$, for all $P \in \mathcal{P}$.

More can be said. The statistic $T(Z)$ in (10) is in fact *conditionally* distributed as t_{n-d} given $X_1, \ldots, X_{n+1}$, which means that ϕ is conditionally valid: $\mathbb{E}_P[\phi(Z) \,|\, X_1, \ldots, X_{n+1}] = \alpha$, for all $P \in \mathcal{P}$. Thus, by the same logic, the prediction interval in (11) is also conditionally valid:

$$\mathbb{P}_P\Big\{ Y_{n+1} \in \big[x^\mathsf{T}\hat{\beta} - \hat{s}_x \cdot t_{n-d,1-\alpha/2},\, x^\mathsf{T}\hat{\beta} + \hat{s}_x \cdot t_{n-d,1-\alpha/2} \big] \,\Big|\, \mathsf{X}, X_{n+1} = x \Big\} = 1-\alpha, \tag{12}$$

for all x, and $P \in \mathcal{P}$. This turns out to be a special feature of the parametric model in (9), which admits a pivotal statistic for fixed design (i.e., fixed X and X_{n+1}). In contrast, as we will see later on (Section 5), analogous conditional coverage statements cannot hold for more general methods which are valid over all i.i.d. sequences, except for those producing trivially inefficient prediction sets.

2.2 Conformal prediction and permutation tests

Now we turn to a nonparametric setting, which will remain our focus for the rest of the paper. Let $Z = (Z_1, \ldots, Z_{n+1}) \sim P$, where P is exchangeable. Given a statistic T, define the permutation p-value

$$p = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} 1\{T(Z_\sigma) \geq T(Z)\}, \tag{13}$$

where recall $\mathcal{S}_{n+1}$ is the set of all permutations of $1, \ldots, n+1$, and we write $Z_\sigma = (Z_{\sigma(1)}, \ldots, Z_{\sigma(n+1)})$. The level α permutation test is then defined by

$$\phi(Z) = 1\{p \leq \alpha\}.$$

Let $\mathcal{P}$ denote the class of all exchangeable distributions. Then by a standard result in permutation testing (e.g., Theorem 17.2.1 in Lehmann and Romano 2022), this is a valid level α test over $\mathcal{P}$, i.e., $\mathbb{E}_P[\phi(Z)] \leq \alpha$, for all $P \in \mathcal{P}$. (This test can be made exact by auxiliary randomization.)

Let us inspect the inversion scheme (8). Writing as before $Z^y = (Z_1, \ldots, Z_n, (X_{n+1}, y))$, this yields

$$C_n(X_{n+1}) = \Big\{ y : \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} 1\{T(Z^y_\sigma) \geq T(Z^y)\} > \alpha \Big\}. \tag{14}$$

Suppose further that we set $T(z) = s(z_{n+1}; z)$ for a score function s satisfying the symmetry condition (6). Then $T(Z_\sigma) = s((X_i, Y_i); Z)$ for any permutation σ such that $\sigma(n+1) = i$. Recalling the notation in (5) for conformal scores, the set in (14) therefore reduces to

$$C_n(X_{n+1}) = \Big\{ y : \frac{1}{n+1} \sum_{i=1}^{n+1} 1\{s_i^y \geq s_{n+1}^y\} > \alpha \Big\}. \tag{15}$$

This is a well-known equivalent form for the conformal prediction set in (3) (e.g., Chapter 2.2.3 of Vovk et al. 2022, or Chapter 3.5.1 of Angelopoulos et al. 2025). Consequently, the validity of conformal prediction in Theorem 1 (and the generalization in Remark 1 to arbitrary symmetric scores) can again be

derived from the duality in Proposition 1: as $Y_{n+1} \in C_n(X_{n+1}) \iff \phi(Z) = 0$, it holds that $\mathbb{P}_P\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha$, for all exchangeable distributions P.

Of course, this duality carries over even in the general case in which T does not exhibit any special symmetry properties. (After all, validity of the permutation test does not require any assumptions about the statistic T.) Setting $T(z) = s(z_{n+1}; z)$, without any symmetry condition on s, the set in (14) becomes

$$C_n(X_{n+1}) = \bigg\{ y : \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} 1\big\{ s(Z^y_{\sigma(n+1)}; Z^y_\sigma) \geq s((X_{n+1}, y); Z) \big\} > \alpha \bigg\}. \tag{16}$$

Proposition 1 again implies this set has coverage: $\mathbb{P}_P\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha$, for all exchangeable P. Already, this says something interesting about conformal prediction: symmetry of the score function is *not* necessary for validity; it is instead a computational device used to simplify the combinatorial nature of the permutation p-value. In other words, exchangeability of the data (not exchangeability of the scores) is all that is needed.

2.3 What is new (and what is not)?

The duality between hypothesis tests and prediction sets described at the start of this section is not new. Although it is perhaps underappreciated, this idea dates back at least 50 years to Cox (1975), or 85 years to Wilks (1941), depending on the interpretation. The connections between conformal prediction and hypothesis testing are also not new. Vovk and coauthors have traditionally defined conformal prediction sets by inverting p-values; see, e.g., Chapter 2.2.3 of Vovk et al. (2022), and also Vovk et al. (1999), which marks the inception of the literature. Furthermore, these authors frequently motivate conformal prediction by drawing connections to the work of Gosset, Fisher, Neyman, and others in the early to mid 1900s; see, e.g., Chapter 13.3 of Vovk et al. (2022), and also Shafer and Vovk (2008) for related discussion.

The statistics literature on conformal prediction has also made clear reference to hypothesis testing, with Lei et al. (2014); Lei and Wasserman (2014); Lei et al. (2018) being early instances of this. Indeed, in both Lei and Wasserman (2014); Lei et al. (2018), it is explicitly stated that the conformal prediction set in (3) can be viewed as inverting a test for the null hypothesis $H_0 : Y_{n+1} = y$. Connections between conformal prediction and permutation/randomization testing are also drawn in Dobriban and Lin (2023); Hoff (2023); Zhang and Zhao (2023); Dobriban and Yu (2025); Koning and van Meer (2025); Barber and Tibshirani (2026); Nair and Janson (2026), and Chapter 3.5.1 of Angelopoulos et al. (2025), among others.

What is new in the current paper? To build toward our results in the coming sections, we start by clarifying the formal connection between conformal prediction and permutation testing (Section 2.2), emphasizing that conformal prediction inverts a permutation test for the null hypothesis of *exchangeability of the full data* $Z = (Z_1, \ldots, Z_{n+1})$. To state the connection once again: the conformal set (3) can be equivalently written as

$$C_n(X_{n+1}) = \Big\{ y : \text{the level } \alpha \text{ permutation test with statistic } T(z) = s(z_{n+1}; z) \text{ applied to} \\ \text{data } (Z_1, \ldots, Z_n, (X_{n+1}, y)) \text{ does not reject the null of exchangeability} \Big\}.$$

With P denoting the distribution of Z, and $\mathcal{P}$ the class of exchangeable distributions, the underlying null hypothesis here is $H_0 : P \sim \mathcal{P}$, which maps cleanly onto traditional formalization in statistical inference.

Our main contribution is to follow this connection to its logical conclusion, and ask what hypothesis testing theory can teach us about conformal prediction. This yields simple, streamlined proofs of results on universality and impossibility (Sections 3 and 5), which are well-known in the conformal literature, as well as on optimality (Section 6), which is comparatively new and still emerging in the literature. Moreover, beyond these particular results, our paper suggests a shift in perspective. Most of the conformal literature referenced above treats facts about hypothesis testing as motivating principles, and treats conformal theory as separate, but analogous. We present a different view: these are not just analogies, and core facts about hypothesis testing *directly* translate to conformal prediction. Hence, the former provides a language which *precisely* helps us understand the latter.

3 Universality

Let ϕ be a valid level α test for a composite null hypothesis $H_0 : Z \sim P$, $P \in \mathcal{P}$, i.e., let ϕ satisfy

$$\mathbb{E}_P[\phi(Z)] \leq \alpha, \quad \text{for all } P \in \mathcal{P}. \tag{17}$$

Let $U = U(Z)$ be a sufficient statistic for $\mathcal{P}$ (which means the distribution of $Z|U$ does not depend on P). For $P \in \mathcal{P}$, let P_U denote the induced distribution on U, and similarly, let $\mathcal{P}_U = \{P_U : P \in \mathcal{P}\}$ denote the induced class. Then any test which satisfies[1]

$$\mathbb{E}[\phi(Z)|U = u] \leq \alpha, \quad \text{for } \mathcal{P}_U\text{-almost every } u, \tag{18}$$

clearly satisfies (17). Tests satisfying (18) are said to have *Neyman structure* with respect to U (see, e.g., Chapter 4.3 of Lehmann and Romano 2022). A seminal result, which can be traced back to the work of Neyman (1937); Ghosh (1948); Hoel (1948); Lehmann and Scheffé (1950, 1955); Basu (1955): if the class $\mathcal{P}$ is rich enough, then Neyman structure (18) is not only sufficient for validity (17), but also necessary.

Theorem 2 (Adaptation of Theorem 4.3.2 in Lehmann and Romano, 2022)**.** *Let U be a sufficient statistic for $\mathcal{P}$. Assume that U is* boundedly complete *for $\mathcal{P}$, which means that for any bounded function f,*

$$\mathbb{E}_{P_U}[f(U)] \leq 0 \text{ for all } P_U \in \mathcal{P}_U \quad \Longrightarrow \quad f(u) \leq 0 \text{ for } \mathcal{P}_U\text{-almost every } u. \tag{19}$$

Then a necessary and sufficient condition for a test to be level α over $\mathcal{P}$ (17) is that it is of Neyman structure (18). That is, marginal validity (17) and U-conditional validity (18) over $\mathcal{P}$ are equivalent.

To be clear, our treatment is a seemingly minor yet nontrivial adaptation of the results in Chapter 4.3 of Lehmann and Romano (2022). They focus on exact level α tests, whereas we focus on valid (but possibly conservative) level α tests. This requires modifying the definitions of Neyman structure and bounded completeness. Notably, our definition (19) is stronger than the traditional definition of bounded completeness, but is suitable for our purposes in what follows. A simple proof of Theorem 2 is given in Appendix A.1. A discussion of the differences in notions of completeness is given in Appendix A.2.

To study the application of this theory to conformal prediction, we inspect the case in which $\mathcal{P}$ is the class of exchangeable distributions. First, note that a sufficient statistic for this class is the multiset or "bag" of elements of $Z = (Z_1, \ldots, Z_{n+1})$, denoted $U = \{\!|Z|\!\}$.[2] For $Z \sim P$, with P exchangeable, and $u = \{\!|z|\!\}$,

$$Z|U = u \sim \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \delta_{z_\sigma}. \tag{20}$$

The right-hand side is the uniform distribution over all permutations z_σ of the given $n+1$ elements, and does not depend on P, which verifies sufficiency. Next, we claim $U = \{\!|Z|\!\}$ is boundedly complete (19) for the class $\mathcal{P}$ of exchangeable distributions. This is straightforward to check: let f be such that $\mathbb{E}_{P_U}[f(U)] \leq 0$ for all exchangeable P, fix $u = \{\!|z|\!\}$, and inspect $P = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \delta_{z_\sigma}$. This choice is exchangeable, and the induced distribution on U is $P_U = \delta_u$. By construction $\mathbb{E}_{P_U}[f(U)] = f(u)$, and so $f(u) \leq 0$.

Therefore, Theorem 2 implies that any test ϕ which is level α over the class of exchangeable distributions must have Neyman structure with respect to $U = \{\!|Z|\!\}$. Recalling (20), we have for $u = \{\!|z|\!\}$,

$$\mathbb{E}[\phi(Z)|U = u] = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \phi(z_\sigma),$$

and so Neyman structure (18) in the current setting translates to

$$\frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \phi(z_\sigma) \leq \alpha, \quad \text{for all } z. \tag{21}$$

[1] For a class $\mathcal{P}$, a statement is said to hold $\mathcal{P}$-almost everywhere if it holds except on a set S with $P(S) = 0$ for all $P \in \mathcal{P}$.

[2] When $Z \in \mathbb{R}^{n+1}$, an equivalent formulation for the sufficient statistic here is to define $U = (Z_{(1)}, \ldots, Z_{(n+1)})$, the vector of order statistics. This is more common in the classical literature on statistical inference. We use the multiset $U = \{\!|Z|\!\}$, as in the conformal literature, because this extends seamlessly to arbitrary unordered spaces.

We can verify that the above condition is equivalent to ϕ having permutation structure, i.e.,

$$\phi \text{ satisfies (21)} \quad \Longleftrightarrow \quad \phi(Z) = 1\{p \leq \alpha\} \text{ for a p-value } p \text{ of the form (13).} \tag{22}$$

This follows from a standard argument in the literature on permutation testing, given in Appendix A.3 for completeness. As valid prediction sets are the result of inverting valid tests, and conformal prediction sets are the result of inverting permutation tests, we have thus reproduced a well-known conformal universality result (cf. Proposition 2.9 of Vovk et al. 2022, or Theorem 9.7 of Angelopoulos et al. 2025).

Theorem 3 (Universality)**.** *Let $Z = (Z_1, \ldots, Z_{n+1}) \sim P$, and consider a mapping which takes X_{n+1} and $Z_1, \ldots, Z_n$ to a prediction set $C_n(X_{n+1})$. If $\mathbb{P}_P\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha$ for all exchangeable P, then this prediction set must be of conformal type, i.e., of the form (16) for a score function s.*

Furthermore, if the set $C_n(X_{n+1})$ is invariant to the ordering of $Z_1, \ldots, Z_n$, then it is of the form (15) for scores (5) and a symmetric score function s.

Proof. Write $C_n(X_{n+1}) = \{y : \phi(Z_1, \ldots, Z_n, (X_{n+1}, y)) = 0\}$ for a test ϕ, which note we can do without a loss of generality since we may always take $\phi(Z) = 1\{Y_{n+1} \notin C_n(X_{n+1})\}$. The assumed coverage property now implies $\mathbb{E}_P[\phi(Z)] \leq \alpha$ for all exchangeable P. By the arguments given above, this implies ϕ must have Neyman structure (21) with respect to $U = \wr Z \int$, and hence $\phi(Z) = 1\{p \leq \alpha\}$ for a permutation p-value p as in (13) and statistic T. Defining the score function by $s(z_{n+1}; z) = T(z)$, we have shown $C_n(X_{n+1})$ is of the form (16), which proves the first claim in the theorem.

For the second claim, if $C_n(X_{n+1})$ is invariant to the ordering of $Z_1, \ldots, Z_n$, then the same must be true of T and s. Thus s meets the symmetry condition (6),[3] and by the arguments given in Section 2.2, the set (16) reduces to (15). This completes the proof. □

4 Efficiency

In what follows, we assume that $Z = (Z_1, \ldots, Z_{n+1}) \sim P$ has i.i.d. entries, i.e., we assume P has the product form $P = P_{X,Y}^{n+1}$ for a joint distribution on the covariates and response $P_{X,Y}$. Letting $P_{X,Y} = P_X \times P_{Y|X}$, note that we can equivalently express $Z_i = (X_i, Y_i) \sim P_{X,Y}$ as $X_i \sim P_X$ and $Y_i | X_i \sim P_{Y|X=X_i}$.

Let $\mathcal{Y}$ denote the response space, and let μ be a measure on $\mathcal{Y}$. For example, for a continuous response we could take $\mathcal{Y} = \mathbb{R}$ and take μ to be Lebesgue measure; for a categorical response $\mathcal{Y}$ would be discrete and we could take μ to be counting measure. Given a prediction set mapping, which takes $X_{n+1}, Z_1, \ldots, Z_n$ to a set $C_n(X_{n+1})$, we define its *efficiency* by

$$\mathbb{E}_P[\mu(C_n(X_{n+1}))],$$

the average measure of $C_n(X_{n+1})$, where the average is over all sources of randomness: $X_{n+1}, Z_1, \ldots, Z_n$. Using the representation (8) for $C_n(X_{n+1})$ as the inversion of a test ϕ, observe that we can write

$$\begin{aligned} \mathbb{E}_P[\mu(C_n(X_{n+1}))] &= \int \mathbb{E}_{P_{X,Y}^n}\big[1\{y \in C_n(x)\}\big] \, \mathsf{d}P_X(x) \, \mathsf{d}\mu(y) \\ &= \int \mathbb{E}_{P_{X,Y}^n}\big[1 - \phi(Z_1, \ldots, Z_n, (x, y))\big] \, \mathsf{d}P_X(x) \, \mathsf{d}\mu(y). \end{aligned} \tag{23}$$

Assuming $\mu(\mathcal{Y}) < \infty$, we can normalize μ to form a probability measure, and define

$$Q = P_{X,Y}^n \times \Big(P_X \times \frac{\mu}{\mu(\mathcal{Y})}\Big). \tag{24}$$

This allows us to express efficiency (23) succinctly as

$$\mathbb{E}_P[\mu(C_n(X_{n+1}))] = \mu(\mathcal{Y}) \cdot \mathbb{E}_Q[1 - \phi(Z)]. \tag{25}$$

[3]Note that the following statements are all equivalent: a score s satisfies (6); it may be written as $s(z_{n+1}; \{z_1, \ldots, z_{n+1}\})$; it may be written as $s(z_{n+1}; \{z_1, \ldots, z_n\})$; and hence $s(z_{n+1}; z)$ is invariant to the order of $z_1, \ldots, z_n$.

In other words, while we have seen that coverage of prediction sets is a statement about type I error under the null hypothesis P (Proposition 1), the formulation in (25) reveals that efficiency is a statement about *type II* error: the right-hand side is precisely a constant (the total mass $\mu(\mathcal{Y})$) times the type II error of ϕ under the alternative hypothesis Q in (24).

In the next two sections, we will leverage the connection in (25) to investigate different questions concerning efficiency. This is a powerful reframing, as it allows us to translate such questions to ones about type I versus type II errors in hypothesis testing.

5 Impossibility

In this section, we use the representation for efficiency developed in the previous section to derive results on the impossibility of conditional coverage, establishing that conformal prediction cannot offer nontrivial guarantees which are conditional—rather than marginal—on the covariate X_{n+1}. In particular, we define the class $\mathcal{P}_{\text{iid}} = \{P : P = P_{X,Y}^{n+1} \text{ for some } P_{X,Y} = P_X \times P_{Y|X}\}$, and we now consider methods which satisfy the conditional property

$$\mathbb{P}_P\{Y_{n+1} \in C_n(x) \,|\, X_{n+1} = x\} \geq 1 - \alpha, \quad \text{for all } x \in \text{supp}(P_X), \text{ and all } P \in \mathcal{P}_{\text{iid}}, \tag{26}$$

where $\text{supp}(P_X)$ denotes the support of P_X.[4] We can recast this equivalently as follows. Let $\mathcal{X}$ denote the covariate space, and δ_x denote the point mass at $x \in \mathcal{X}$ (it is the distribution which places all it mass at x). Define the class

$$\mathcal{P}_{\text{cond}} = \Big\{P : P = P_{X,Y}^n \times (\delta_x \times P_{Y|X=x}) \text{ for some } P_{X,Y} = P_X \times P_{Y|X} \text{ and } x \in \text{supp}(P_X)\Big\},$$

Then the conditional coverage property in (26) is equivalent to

$$\mathbb{P}_P\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha, \quad \text{for all } P \in \mathcal{P}_{\text{cond}}, \tag{27}$$

which is marginal coverage over $\mathcal{P}_{\text{cond}}$.

For a given distribution $P = P_{X,Y}^{n+1}$, and a given non-atom point $x_0 \in \text{supp}(P_X)$, we will study the limits of pointwise efficiency $\mathbb{E}_P[\mu(C_n(x_0))]$ among all methods with conditional coverage over $\mathcal{P}_{\text{iid}}$, as in (26), or equivalently, marginal coverage over $\mathcal{P}_{\text{cond}}$, as in (27). As demonstrated in Section 4, this can be posed as a question about power. If $\mu(\mathcal{Y}) < \infty$, then analogous to (24), we can define

$$Q_0 = P_{X,Y}^n \times \Big(\delta_{x_0} \times \frac{\mu}{\mu(\mathcal{Y})}\Big), \tag{28}$$

and by arguments analogous to those leading to (25), it holds that

$$\mathbb{E}_P[\mu(C_n(x_0))] = \mu(\mathcal{Y}) \cdot \mathbb{E}_{Q_0}[1 - \phi(Z)]. \tag{29}$$

Note that the right-hand side is a constant times the type II error of ϕ against Q_0, or one minus its power. Thus, our question becomes: what is largest possible power among all tests valid over $\mathcal{P}_{\text{cond}}$?

We break our analysis below into two cases: the finite measure case $\mu(\mathcal{Y}) < \infty$ (which is needed in order for Q_0 to be well-defined), and the infinite measure case (which requires a truncation argument). In either case, we will rely on the following core result. This can be found in Theorem 5 of Kraft (1955), however, Kraft attributes the result to Le Cam. Theorem 5 in Kraft (1955) actually establishes a stronger statement than that transcribed below, under conditions on the distributions P, Q. The weaker version we state in Theorem 4 follows directly from their proof without any such conditions.

[4] For a distribution P, its support $\text{supp}(P)$ is the smallest closed set S such that $P(S) = 1$. We work with the conditional coverage property in (26), which holds at each $x \in \text{supp}(P_X)$, because it cleanly transforms to (27). We could alternatively choose to define (26) as holding for P_X-almost every x, which would be more in line with the typical phrasing of conditional coverage properties in the conformal literature; however, this requires slightly more sophisticated arguments which we prefer to avoid for simplicity.

Theorem 4 (Weak version of Theorem 5 in Kraft, 1955)**.** *Let $\mathcal{P}$ and $\mathcal{Q}$ be classes of distributions, and let ϕ be a valid level α test over $\mathcal{P}$, i.e., $\sup_{P\in\mathcal{P}} \mathbb{E}_P[\phi(Z)] \leq \alpha$. Then the power of ϕ over $\mathcal{Q}$ is upper bounded by*

$$\inf_{Q\in\mathcal{Q}} \mathbb{E}_Q[\phi(Z)] \leq \alpha + \mathrm{TV}(\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q})),$$

where $\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q})$ denote the convex hulls of $\mathcal{P}$, $\mathcal{Q}$, respectively, and $\mathrm{TV}(\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q}))$ denotes the total variation distance between the hulls, i.e., the infimum of $\mathrm{TV}(P, Q) = \sup_S |P(S) - Q(S)|$ over all $P \in \mathrm{conv}(\mathcal{P})$ and $Q \in \mathrm{conv}(\mathcal{Q})$.

5.1 Finite measure case

In the finite measure case $\mu(\mathcal{Y}) < \infty$, the distribution Q_0 in (28) is well-defined, and the arguments above already bring us close to a result about the limits of efficiency of prediction set methods which are conditionally valid, as in (26). The key realization is that

$$\mathrm{TV}(\mathrm{conv}(\mathcal{P}_{\mathrm{cond}}), \{Q_0\}) = 0. \tag{30}$$

This claim will be verified shortly. As a consequence, by Theorem 4, any test which is level α over $\mathcal{P}_{\mathrm{cond}}$ must be *powerless*, i.e., have power at most α, against the alternative Q_0 in (28). Rephrasing this in terms of prediction sets, using (29), we arrive at the following hardness result.

Theorem 5 (Conditional hardness, finite measure case)**.** *Assume $\mu(\mathcal{Y}) < \infty$. Consider a mapping which takes X_{n+1} and $Z_1, \ldots, Z_n$ to a prediction set $C_n(X_{n+1})$. If this method satisfies conditional validity over $\mathcal{P}_{\mathrm{iid}}$, as in (26), then for any $P_{X,Y} = P_X \times P_{Y|X}$ and any non-atom point $x_0 \in \mathrm{supp}(P_X)$, it holds that*

$$\mathbb{E}_P[\mu(C_n(x_0))] \geq \mu(\mathcal{Y}) \cdot (1 - \alpha).$$

In other words, the pointwise efficiency $\mathbb{E}_P[\mu(C_n(x_0))]$ of this method is no better than the trivial method which draws $B \sim \mathrm{Bern}(1-\alpha)$ independently of everything else, and outputs $\mathcal{Y}$ if $B = 1$ and $\emptyset$ otherwise.

Proof. As before, we write $C_n(X_{n+1}) = \{y : \phi(Z_1, \ldots, Z_n, (X_{n+1}, y)) = 0\}$ for a test ϕ, which is without a loss of generality as we may always take $\phi(Z) = 1\{Y_{n+1} \notin C_n(X_{n+1})\}$. The assumed conditional validity property (26) is equivalent to the marginal property (27) over $\mathcal{P}_{\mathrm{cond}}$, which means ϕ is level α over $\mathcal{P}_{\mathrm{cond}}$. Combining (29) with Theorem 4, we have

$$\mathbb{E}_P[\mu(C_n(x_0))] \geq \mu(\mathcal{Y}) \cdot (1 - \alpha) - \mu(\mathcal{Y}) \cdot \mathrm{TV}(\mathrm{conv}(\mathcal{P}_{\mathrm{cond}}), \{Q_0\}).$$

Therefore, to prove the theorem, it suffices to verify (30). To this end, define

$$\tilde{P}_{Y|X=x} = \begin{cases} P_{Y|X=x} & \text{if } x \neq x_0, \\ \frac{\mu}{\mu(\mathcal{Y})} & \text{if } x = x_0, \end{cases}$$

and $P_0 = (P_X \times \tilde{P}_{Y|X})^n \times (\delta_{x_0} \times \frac{\mu}{\mu(\mathcal{Y})})$. Then $P_0 \in \mathcal{P}_{\mathrm{cond}}$, and yet $\mathrm{TV}(P_0, Q_0) = 0$, since

$$\begin{aligned}
\mathrm{TV}(P_0, Q_0) &\leq n \cdot \mathrm{TV}(P_X \times \tilde{P}_{Y|X}, P_X \times P_{Y|X}) \\
&= n \cdot \sup_S \left| \int_S \mathsf{d}P_X(x)\, \mathsf{d}\tilde{P}_{Y|X=x}(y) - \int_S \mathsf{d}P_X(x)\, \mathsf{d}P_{Y|X=x}(y) \right| \\
&= n \cdot \sup_S \left| \int_{S\setminus\{x_0\}} \mathsf{d}P_X(x)\, \mathsf{d}\tilde{P}_{Y|X=x}(y) - \int_{S\setminus\{x_0\}} \mathsf{d}P_X(x)\, \mathsf{d}P_{Y|X=x}(y) \right| \\
&= 0,
\end{aligned}$$

where the second-to-last line holds because x_0 is a non-atom point of P_X, and the last line holds because $\tilde{P}_{Y|X=x} = P_{Y|X=x}$ for $x \neq x_0$, by construction. This completes the proof of (30), and the theorem. □

5.2 Infinite measure case

In the infinite measure case $\mu(\mathcal{Y}) = \infty$, we can no longer normalize μ over all of $\mathcal{Y}$ to form the distribution in (28). However, assuming that μ is σ-finite, which is true of Lebesgue measure on $\mathcal{Y} = \mathbb{R}$, we can adapt Theorem 5 using a truncation to reproduce a well-known impossibility result from the conformal literature (cf. Proposition 4 in Vovk 2012, or Lemma 1 in Lei and Wasserman 2014).

Corollary 1 (Conditional hardness, infinite measure case)**.** *Assume $\mu(\mathcal{Y}) = \infty$ and assume μ is σ-finite. Consider a mapping which takes X_{n+1} and $Z_1, \ldots, Z_n$ to a prediction set $C_n(X_{n+1})$. If this method satisfies conditional validity over $\mathcal{P}_{\text{iid}}$, as in (26), then for any $P_{X,Y} = P_X \times P_{Y|X}$ and any non-atom point $x_0 \in \text{supp}(P_X)$, it holds that*

$$\mathbb{E}_P[\mu(C_n(x_0))] = \infty.$$

Proof. Because μ is σ-finite and $\mu(\mathcal{Y}) = \infty$, there exists a sequence of measurable sets $A_k \subseteq \mathcal{Y}$ such that $0 < \mu(A_k) < \infty$ for all k and $\mu(A_k) \to \infty$ as $k \to \infty$. For each k, define the probability measure ν_k on $\mathcal{Y}$ by $\nu_k(B) = \frac{\mu(B \cap A_k)}{\mu(A_k)}$. Then by Theorem 5 applied to ν_k, we have $\mathbb{E}_P[\nu_k(C_n(x_0))] \geq 1 - \alpha$, or equivalently

$$\mathbb{E}_P[\mu(C_n(x_0) \cap A_k)] \geq (1 - \alpha) \cdot \mu(A_k).$$

By monotonicity of μ, we have $\mu(C_n(x_0)) \geq \mu(C_n(x_0) \cap A_k)$, hence

$$\mathbb{E}_P[\mu(C_n(x_0))] \geq (1 - \alpha) \cdot \mu(A_k).$$

Sending $k \to \infty$ completes the proof. □

6 Optimality

We examine another application of the representation for efficiency from Section 4, to study the question of optimality: characterizing the precise form of optimally efficient methods for prediction sets, among all methods with valid coverage over the class $\mathcal{P}_{\text{exch}} = \{P : P \text{ is exchangeable}\}$. As explained in the discussion following (25), this question can be reduced to that of optimal power among all tests which are valid over $\mathcal{P}_{\text{exch}}$. Naturally, Neyman-Pearson theory comes to mind, and this will indeed be our main tool, after we carefully setting up and reformulating the problem, as described below.

For a fixed $P \in \mathcal{P}_{\text{iid}} = \{P : P = P_{X,Y}^{n+1} \text{ for some } P_{X,Y} = P_X \times P_{Y|X}\}$, we consider

$$\begin{aligned} &\text{minimize} && \mathbb{E}_P[\mu(C_n(X_{n+1})] \\ &\text{subject to} && \mathbb{P}_{P'}\{Y_{n+1} \in C_n(X_{n+1})\} \geq 1 - \alpha, \text{ for all } P' \in \mathcal{P}_{\text{exch}}, \end{aligned} \tag{31}$$

where the minimization is understood to be over mappings which take $X_{n+1}, Z_1, \ldots, Z_n$ to a set $C_n(X_{n+1})$. Any such mapping can be written as $C_n(X_{n+1}) = \{y : \phi(Z_1, \ldots, Z_n, (X_{n+1}, y)) = 0\}$ for a test ϕ (we may always take $\phi(Z) = 1\{Y_{n+1} \notin C_n(X_{n+1})\}$). Assuming $\mu(\mathcal{Y}) < \infty$, we now recall the formulation (25) for the expected measure of the prediction set, with the alternative Q defined in (24). The optimization (31) is hence equivalent to

$$\begin{aligned} &\text{maximize} && \mathbb{E}_Q[\phi(Z)] \\ &\text{subject to} && \mathbb{E}_{P'}[\phi(Z)] \leq \alpha, \text{ for all } P' \in \mathcal{P}_{\text{exch}}. \end{aligned} \tag{32}$$

We develop one more reformulation. Writing $\mathbb{E}_Q[\phi(Z)] = \int \mathbb{E}_Q[\phi(Z)|U = u] \, \mathrm{d}Q_U(u)$ for $U = \lfloor Z \rfloor$, consider optimizing the integrand for a fixed $u = \lfloor z \rfloor$,

$$\begin{aligned} &\text{maximize} && \mathbb{E}_{Q_u}[\phi(Z)] \\ &\text{subject to} && \mathbb{E}_{P'}[\phi(Z)] \leq \alpha, \text{ for all } P' \in \mathcal{P}_{\text{exch}}, \end{aligned} \tag{33}$$

where Q_u denotes the distribution of $Z|U = u$ when $Z \sim Q$. If the optimal test in (33) does not depend on u, then it will also be optimal for (32). This will indeed be the case and we will focus on (33) henceforth.

At this point, we are well-positioned to invoke Neyman-Pearson theory. Because (33) involves a composite null hypothesis, we must first establish a least favorable distribution. Once established, the Neyman-Pearson lemma characterizes optimal tests. This approach traces back to the celebrated work of Neyman and Pearson (1933a,b), with extensions to composite hypotheses and least favorable distributions by Wald (1939); Lehmann and Stein (1948); Wald (1950). We state a version from Lehmann and Romano (2022).

Theorem 6 (Theorems 3.2.1 and 3.8.1 in Lehmann and Romano, 2022)**.** *Let $\mathcal{P}$ be a class of null distributions and let Q be a point alternative. For a distribution Λ over the class $\mathcal{P}$, and define the point null P_Λ by $P_\Lambda(S) = \int_{\mathcal{P}} P(S)\, \mathsf{d}\Lambda(P)$. Let p_Λ and q denote the densities (Radon-Nikodym derivatives) of P_Λ and Q, respectively, with respect to a common base measure. Consider a test ϕ_Λ which satisfies*

$$\phi_\Lambda(z) = \begin{cases} 1 & \text{if } q(z) > k p_\Lambda(z), \\ 0 & \text{if } q(z) < k p_\Lambda(z), \end{cases}$$

and suppose k can be chosen such that $\mathbb{E}_{P_\Lambda}[\phi_\Lambda(Z)] = \sup_{P \in \mathcal{P}} \mathbb{E}_P[\phi_\Lambda(Z)] = \alpha$. Then ϕ_Λ achieves the highest power against Q among all tests which are level over $\mathcal{P}$, and the distribution Λ is called least favorable.

6.1 Optimal conformal score

Our next step is to establish a least favorable distribution for testing the composite null $H_0 : P' \in \mathcal{P}_{\text{exch}}$ versus the point alternative $H_1 : P' = Q_u$, where recall $u = \{\!|z|\!\}$ and Q_u denotes the distribution of $Z|U = u$ when $Z \sim Q$. First, we work out Q_u explicitly. Let $p_{X,Y}$ be the density of $P_{X,Y}$ with respect to $\rho \times \mu$, for some measure ρ on $\mathcal{X}$. Factorize $p_{X,Y} = p_X \times p_{Y|X}$. Then Q_u is supported on permutations of z, with

$$\mathbb{P}_{Q_u}\{Z = z_\sigma\} = \frac{\prod_{i=1}^n p_{X,Y}(z_{\sigma(i)}) \cdot p_X(x_{\sigma(n+1)})}{\sum_{\pi \in \mathcal{S}_{n+1}} \prod_{i=1}^n p_{X,Y}(z_{\pi(i)}) \cdot p_X(x_{\pi(n+1)})}.$$

If we multiply and divide in the numerator by $p_{Y|X=x_{\sigma(n+1)}}(y_{\sigma(n+1)})$, and multiply and divide in the denominator by $p_{Y|X=x_{\pi(n+1)}}(y_{\pi(n+1)})$, then observe that the numerator and denominator share a common factor of $\prod_{i=1}^{n+1} p_{X,Y}(x_i, y_i)$, which cancels, we are left with

$$\mathbb{P}_{Q_u}\{Z = z_\sigma\} = \frac{c_u}{p_{Y|X=x_{\sigma(n+1)}}(y_{\sigma(n+1)})}, \tag{34}$$

where $c_u = \big[\sum_{\pi \in \mathcal{S}_{n+1}} \frac{1}{p_{Y|X=x_{\pi(n+1)}}(y_{\pi(n+1)})}\big]^{-1}$ is a normalizing constant.

Consider now the choice $P'_u = \frac{1}{n+1}\sum_{\sigma \in \mathcal{S}_{n+1}} \delta_{z_\sigma} \in \mathcal{P}_{\text{exch}}$. We will show that P'_u is least favorable relative to Q_u. (By this, we mean that the distribution Λ which places all of its mass on P'_u is least favorable.) Note that P'_u and Q_u are absolutely continuous with respect to counting measure on the multiset $u = \{\!|z|\!\}$. We denote their densities by p'_u and q_u, respectively. Then the Neyman-Pearson optimal test for P'_u versus Q_u takes the form

$$\phi(z_\sigma) = \begin{cases} 1 & \text{if } q_u(z_\sigma) > k_u p'_u(z_\sigma), \\ 0 & \text{if } q_u(z_\sigma) < k_u p'_u(z_\sigma), \end{cases}$$

for a threshold k_u. Using $p'_u(z_\sigma) = 1$, and the form of q_u as specified by (34), this can be rewritten as

$$\phi(z_\sigma) = \begin{cases} 1 & \text{if } 1/p_{Y|X=x_{\sigma(n+1)}}(y_{\sigma(n+1)}) > k_u, \\ 0 & \text{if } 1/p_{Y|X=x_{\sigma(n+1)}}(y_{\sigma(n+1)}) < k_u, \end{cases} \tag{35}$$

for a redefined threshold k_u. This threshold is chosen so that

$$\mathbb{E}_{P'_u}[\phi(Z)] = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \phi(z_\sigma) = \alpha. \tag{36}$$

The key observation is these two conditions (35), (36) are fulfilled by a randomized permutation test, as explained below, and proved in Appendix A.4.

Proposition 2. *Let $T(Z) = 1/p_{Y|X=X_{n+1}}(Y_{n+1})$, and define the test $\phi^*(Z) = 1\{p^* \leq \alpha\}$, where*

$$p^* = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \Big(1\{T(Z_\sigma) > T(Z)\} + V \cdot 1\{T(Z_\sigma) = T(Z)\} \Big),$$

for $V \sim \mathrm{Unif}[0,1]$, independently of everything else. Then ϕ^* satisfies (35), (36).

As the Neyman-Pearson optimal test for P'_u versus Q_u takes the form of a randomized permutation test ϕ^*, as defined in Proposition 2, we know that it controls type I error over all of $\mathcal{P}_{\mathrm{exch}}$ (as do all randomized permutation tests; see, e.g., Theorem 17.2.1 in Lehmann and Romano 2022). Thus, we have succeeded in identifying a least favorable distribution, and by Theorem 6, we know that ϕ^* solves (33). Finally, as ϕ^* does not depend on u (note the choice of test statistic $T(Z) = 1/p_{Y|X=X_{n+1}}(Y_{n+1})$ depends only on Q, not Q_u), we conclude that ϕ^* solves (32). Transcribing this back to prediction sets yields the following result.

Theorem 7 (Optimality)**.** *Assume μ is a σ-finite measure on $\mathcal{Y}$. Fix any distribution $P_{X,Y} = P_X \times P_{Y|X}$, denote by $p_{Y|X=x}$ the density of $P_{Y|X=x}$ with respect to μ, and assume that $p_{Y|X=x}(y) > 0$ for all $y \in \mathcal{Y}$, and P_X-almost every x. Let $P = P_{X,Y}^{n+1}$. Among all mappings which take $X_{n+1}, Z_1, \ldots, Z_n$ to a prediction set $C_n(X_{n+1})$, and provide $1-\alpha$ coverage over the class $\mathcal{P}_{\mathrm{exch}}$ of exchangeable distributions, the optimal efficiency $\mathbb{E}_P[\mu(C_n(X_{n+1})]$ is attained by the following randomized conformal predictor:*

$$C_n^*(X_{n+1}) = \left\{ y : \frac{1}{n+1} \sum_{i=1}^{n+1} \Big(1\{s_i^y > s_{n+1}^y\} + V \cdot 1\{s_i^y = s_{n+1}^y\} \Big) > \alpha \right\}, \tag{37}$$

where $V \sim \mathrm{Unif}[0,1]$, independently of everything else, the scores are as in (5), and the score function is

$$s(z_{n+1}; z) = \frac{1}{p_{Y|X=x_{n+1}}(y_{n+1})}. \tag{38}$$

In other words, the construction given in (37), (38) solves problem (31).

Proof. Assume first that $\mu(\mathcal{Y}) < \infty$. Then the distribution Q in (24) is well-defined and the arguments above establish that the test ϕ^* in Proposition 2 solves (32). The corresponding prediction set is

$$C_n^*(X_{n+1}) = \left\{ y : \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \Big(1\{T(Z_\sigma) > T(Z)\} + V \cdot 1\{T(Z_\sigma) = T(Z)\} \Big) > \alpha \right\},$$

which reduces to the claimed form (37) with score $s(z_{n+1}; z) = T(z)$ as in (38), because this score function is invariant to relabeling $z_1, \ldots, z_n$. This completes the proof for the finite measure case.

For the general case, permitting infinite measure $\mu(\mathcal{Y}) = \infty$, we can adapt the above strategy by using a slightly different perspective which does not require defining Q as in (24), deferred to Appendix A.5. □

Several closely related results exist in the literature, as we describe below. Thus, we emphasize the above result not for its novelty, but as a clear demonstration of the power of the hypothesis-theoretic framework for prediction sets. We also remark that the choice of an i.i.d. reference distribution $P = P_{X,Y}^{n+1}$ with which to measure efficiency is made for simplicity. The arguments can be extended to an exchangeable distribution P, which leads to a somewhat more complex choice of optimal score: the inverse conditional density of Y_{n+1} given X_{n+1} and the rest of the samples $Z_1, \ldots, Z_n$.

6.2 Related results

There are numerous related results in the conformal prediction literature which characterize the optimal score at the population-level; these results effectively take $n \to \infty$ (formalized in different ways), and in this large-sample limit, they all arrive at the same answer: the optimal score is the inverse conditional density $1/p_{Y|X=x_{n+1}}(y_{n+1})$. This is covered in Chapter 3.1 of Vovk et al. (2022), and Chapters 5.2 and 5.3 of Angelopoulos et al. (2025); see also Lei et al. (2014, 2018); Sadinle et al. (2019); Izbicki et al. (2020) for

related oracle constructions. The key distinction is that our result in Theorem 7 is finite-sample, i.e., it is about conformal prediction itself, not a large-sample analog.

In terms of finite-sample results, Dandapanthula and Ramdas (2025); Hore and Ramdas (2026) developed what they call the first and second "conformal Neyman-Pearson lemmas" for confidence sets in nonparametric changepoint detection. These results are derived by reducing the question of an optimal score function to one of optimal power in testing, and applying the Neyman-Pearson lemma. While their problem setting is different than ours—changepoint detection over i.i.d. sequences, with tests of a point null versus point alternative—the spirit of the construction is similar.

For prediction sets, versions of the finite-sample result in Theorem 7 appear in recent literature studying optimality: Theorem 4.1 of Hoff (2023) and Theorem 8 of Koning and van Meer (2025). These authors use different approaches, which we discuss separately. Hoff (2023) develops a general framework for prediction sets, attributed originally to Faulkenberry (1973), which at its core is based on conditioning on a complete sufficient statistic. This is closely related to the perspectives on Neyman structure and completeness in Section 3 of this paper. There are some superficial differences in Hoff's Theorem 4.1 and our Theorem 7: their result does not include covariates, and they analyze a Bayesian notion of efficiency, with respect to a mixture distribution P_π defined by a prior π over $\mathcal{P}$. In our view, these are minor and one can adapt their framework to handle covariates, and to analyze frequentist efficiency (by taking π to be a point mass).

A bigger difference lies in the apparent generality of Hoff's conclusion. Theorem 4.1 in Hoff (2023) derives a form for the optimal conformal score based on the posterior predictive density (optimal among conformal methods), but falls short of concluding that the corresponding conformal predictor is optimal among all valid methods for prediction sets. We suspect that this is due to their choice of constraint set: they consider methods which obtain exact coverage over the class $\mathcal{P}_{\text{iid}}$ of i.i.d. distributions, instead of methods which have valid coverage over the class $\mathcal{P}_{\text{exch}}$ of exchangeable distributions, as we do in (31). Hoff's focus on exact coverage stems from their use of the traditional "two-sided" definition of bounded completeness (43), which is tied to exact type I error control. We introduce a "one-sided" notion of bounded completeness (19), as a tool for analyzing valid type I error control. For i.i.d. distributions, there is a gap between these two definitions: the multiset $U = \lbag Z \rbag$ is boundedly complete according to (43) but not (19) (see Appendix A.2). In other words, for i.i.d. distributions, there seems to be a fundamental difference in whether multiset-conditioning is necessary for exact versus valid (but possibly conservative) type I error control.

Koning and van Meer (2025) construct prediction sets by leveraging the duality to testing, as in our paper. Their primary focus is on "fuzzy" prediction sets, which generalize the binary nature (inclusion/exclusion) of a traditional prediction set, and can be seen as rescaled e-values. They observe that efficiency of a traditional prediction set can be represented in terms of type II error/power against a particular alternative, as in Section 4 of this paper; for fuzzy prediction sets, they generalize this by defining a utility function used to measure the amount of evidence presented by the e-value under the alternative. Leveraging tools from the emerging literature on e-values (Koning, 2023, 2024), they first reduce validity over the class $\mathcal{P}_{\text{exch}}$ of exchangeable null distributions to multiset-conditional validity, then derive utility-optimal fuzzy prediction sets for concave, nondecreasing, and upper-semicontinuous utility functions. For the "Neyman-Pearson utility", the consequence in Theorem 8 and Corollary 2 of Koning and van Meer (2025) is that conformal prediction with inverse density score $1/p_{Y|X}$ is optimally efficient among all methods with valid coverage over $\mathcal{P}_{\text{exch}}$. Theorem 7 in our paper reproduces this conclusion with a simpler argument based on classical Neyman-Pearson theory (and technically, extends this result to accommodate a σ-finite measure μ used to measure efficiency, such as Lebesgue measure on $\mathcal{Y} = \mathbb{R}$).

6.3 Interpretations and illustrations

The interpretation of the result in Theorem 7 is that efficient conformal methods should approximate the density $p_{Y|X}$ as closely as possible. This helps us contextualize a choice such as the absolute residual score from a predictor of the conditional mean $f(x) = \mathbb{E}_P[Y|X = x]$. Consider a model in which

$$1/p_{Y|X=x}(y) \text{ is a monotone increasing function of } |y - f(x)|, \tag{39}$$

such as the conditional normal model $Y|X = x \sim N(f(x), \sigma^2)$. Because conformal sets are invariant under monotone transformations of the score, we can view the use of a conditional mean predictor $\hat{f} = A(z)$ (from an algorithm A) together with residual score $s(z_{n+1}; z) = |y_{n+1} - \hat{f}(x_{n+1})|$ as a plug-in strategy for estimating the conditional density under (39).

The model (39) assumes homoskedasticity. If we instead consider the heteroskedastic conditional normal model $Y|X = x \sim N(f(x), \sigma^2(x))$, then

$$1/p_{Y|X=x}(y) \text{ is a monotone increasing function of } \frac{(y - f(x))^2}{2\sigma^2(x)} + \log \sigma(x). \tag{40}$$

It is interesting to note that (40) does *not* lead to studentized score, but something different—it suggests forming conditional mean $\hat{f} = A(z)$ and variance $\hat{\sigma} = B(z)$ predictors (from algorithms A, B), and using

$$s(z_{n+1}; z) = \frac{(y_{n+1} - \hat{f}(x_{n+1}))^2}{2\hat{\sigma}^2(x_{n+1})} + \log \hat{\sigma}(x_{n+1}), \tag{41}$$

as the corresponding plug-in strategy under (40). Studentized score would keep only the first term above, or equivalently (again due to invariance under monotone transformations),

$$s(z_{n+1}; z) = \frac{|y_{n+1} - \hat{f}(x_{n+1})|}{\hat{\sigma}(x_{n+1})}. \tag{42}$$

This score has its own benefits (e.g., when $\hat{f}, \hat{\sigma}$ are accurate it brings us closer to conditional coverage, as studied in Lei et al. 2018), but is inefficient compared to (41). The score (41) shrinks the prediction sets in high-variance regions compared to the studentized score (42), due to the second term $\log \hat{\sigma}(x_{n+1})$, which sacrifices local coverage but improves efficiency. Figure 1 provides an illustration.

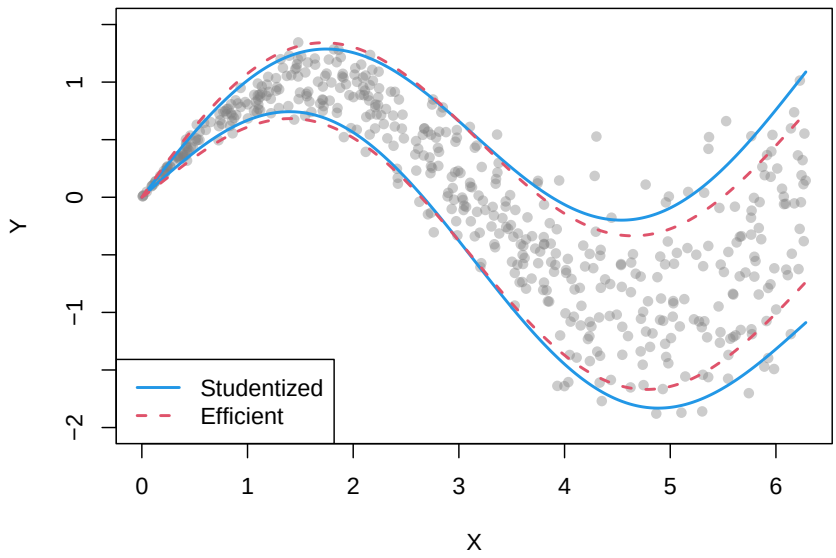

Figure 1: Simple example of conformal prediction intervals using studentized (42) and efficient (41) scores in the Gaussian heteroskedastic model. See Appendix A.6 for further explanation and discussion.

7 Discussion

This paper presents a perspective of conformal prediction methods as inverting tests of exchangeability. This view allows us to reinterpret well-known conformal universality and impossibility results as consequences of hypothesis testing theory from the 1950s. It also facilitates the construction of optimal conformal predictors, as an application of Neyman-Pearson theory.

Broadly, the goal of this work is to motivate and document the direct relevance of foundational statistical inference—predominantly developed in the first half of the 20th century—to modern predictive inference. While we studied standard conformal prediction (for exchangeable samples), the relevance of hypothesis-theoretic tools should extend beyond the standard setting, as many newer variants of conformal can be seen as inverting hypothesis tests given partial information about the data (Barber and Tibshirani, 2026). Much ground remains to be explored in both directions: what hypothesis testing can teach us about conformal prediction, and vice versa.

Acknowledgments

We thank Emmanuel Candès for helpful comments, and Nick Koning for introducing us to his work. R.J.T. was supported by the Office of Naval Research (ONR) grant N00014-20-1-2787, and R.F.B. by ONR grant N00014-24-1-2544.

An AI model was used to brainstorm the key representation in (25) of efficiency in terms of type II error. AI was also used to help identify references, and to implement the example in Figure 1.

References

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. arXiv: 2411.11824, 2025.

Rina Foygel Barber and Ryan J. Tibshirani. Unifying different theories of conformal prediction. *Electronic Journal of Statistics*, 20(1):1428–1474, 2026.

Debabrata Basu. On statistics independent of a complete sufficient statistic. *Sankhya: The Indian Journal of Statistics*, 15(4):377–380, 1955.

David. R. Cox. Prediction intervals and empirical Bayes confidence intervals. *Journal of Applied Probability*, 12(S1):47–55, 1975.

Sanjit Dandapanthula and Aaditya Ramdas. Offline changepoint localization using a matrix of conformal p-values. arXiv: 2505.00292, 2025.

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

Edgar Dobriban and Mengxin Yu. SymmPI: Predictive inference for data with group symmetries. *Journal of the Royal Statistical Society: Series B*, 87(5):1353–1381, 2025.

G. David Faulkenberry. A method of obtaining prediction intervals. *Journal of the American Statistical Association*, 68(342):433–435, 1973.

Ronald A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6(4):391–398, 1935.

Donald A. S. Fraser. Completeness of order statistics. *Canadian Journal of Mathematics*, 6:42–45, 1954.

Manindra N. Ghosh. On the extension of Neyman's structure. *Sankhya: The Indian Journal of Statistics*, 8 (4):329–338, 1948.

Paul G. Hoel. On Neyman's structure. *Annals of Mathematical Statistics*, 19(2):268–271, 1948.

Peter Hoff. Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2):901–928, 2023.

Rohan Hore and Aaditya Ramdas. Conformal changepoint localization. arXiv: 2602.06267, 2026.

Rafael Izbicki, Gilson Shimizu, and Rafael Stern. Flexible distribution-free conditional predictive bands using density estimators. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2020.

Nick W. Koning. Measuring evidence against exchangeability and group invariance with e-values. arXiv: 2310.01153, 2023.

Nick W. Koning. Continuous testing: Unifying tests and e-values. arXiv: 2409.05654, 2024.

Nick W. Koning and Sam van Meer. Fuzzy prediction sets: Conformal prediction with e-values. arXiv: 2509.13130, 2025.

Charles H. Kraft. Some conditions for consistency and uniform consistency of statistical procedures. *University of California Publications in Statistics*, 2(1):125–142, 1955.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, fourth edition, 2022.

Erich L. Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased tests: Part I. *Sankhya: The Indian Journal of Statistics*, 10(4):305–340, 1950.

Erich L. Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased tests: Part II. *Sankhya: The Indian Journal of Statistics*, 15(3):219–236, 1955.

Erich L. Lehmann and Charles Stein. Most powerful tests of composite hypotheses. I. Normal distributions. *Annals of Mathematical Statistics*, 19(4):495–516, 1948.

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, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 501(108):278–287, 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.

Yash Nair and Lucas Janson. Randomization tests for adaptively collected data. *To appear, Journal of the American Statistical Association*, 2026.

Jerzy Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London: Series A*, 236(767):333–380, 1937.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London: Series A*, 231(694–706):289–337, 1933a.

Jerzy Neyman and Egon S. Pearson. The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society*, 29(4):492–510, 1933b.

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.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008.

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, second edition, 2022.

Volodya Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the International Conference on Machine Learning*, 1999.

Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10(4):299–326, 1939.

Abraham Wald. *Statistical Decision Functions*. Wiley, 1950.

Samuel S. Wilks. Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics*, 12(1):91–96, 1941.

Yao Zhang and Qingyuan Zhao. What is a randomization test? *Journal of the American Statistical Association*, 118(544):2928–2942, 2023.

A Appendix

A.1 Proof of Theorem 2

Let ϕ be a test that is valid over $\mathcal{P}$, according to (17). Let $f(u) = \mathbb{E}[\phi(Z)|U = u] - \alpha$, which by sufficiency of U, does not depend on P. Then for any $P_U \in \mathcal{P}_U$, we have $\mathbb{E}_{P_U}[f(U)] = \mathbb{E}_P[\phi(Z)] - \alpha \leq 0$, by assumption. Furthermore, f is bounded (by $1 - \alpha$), so by bounded completeness of U with respect to $\mathcal{P}_U$, it follows that $f(u) \leq 0$ for $\mathcal{P}_U$-almost every u, which is equivalent to (18).

A.2 Bounded completeness

The traditional definition of bounded completeness of a sufficient statistic U for a class $\mathcal{P}$ states that, for any bounded function f,

$$\mathbb{E}_{P_U}[f(U)] = 0 \text{ for all } P \in \mathcal{P} \quad \Longrightarrow \quad f(u) = 0 \text{ for } \mathcal{P}_U\text{-almost every } u. \tag{43}$$

In our version (19), we have replaced equalities with inequalities. Our version may be seen as "one-sided" bounded completeness, and the traditional version as "two-sided" bounded completeness.

The one-sided version (19) is strictly stronger than the traditional two-sided version (43). To see this, note that we can apply (19) to both f and $-f$ to conclude (43). However, the reverse implication is not true.

To demonstrate this, we first recall that it is well-known that $U = \lbag Z \rbag$ is boundedly complete in the two-sided sense (43) with respect to the class of i.i.d. continuous distributions (Fraser, 1954). This extends to some discrete distributions, such as the class of i.i.d. Bernoulli distributions for binary random variables (Lehmann and Scheffé, 1950, 1955). However, consider the following example: draw i.i.d. $Z_1, Z_2 \sim \text{Bern}(p)$ and enumerate the distinct multisets by $u_1 = \lbag 0, 0 \rbag$, $u_2 = \lbag 1, 0 \rbag$, and $u_3 = \lbag 1, 1 \rbag$. Let $f(u_1) = -1$, $f(u_2) = 1$, and $f(u_3) = -1$. Then

$$\mathbb{E}_p[f(U)] = -(1-p)^2 + 2p(1-p) - p^2 = -(2p-1)^2 \leq 0, \quad \text{for all } p \in [0, 1].$$

However, $f(u_2) > 0$, which contradicts one-sided bounded completeness (19).

The above example also reveals an interesting difference in the "richness" of the classes $\mathcal{P}_{\text{iid}}$ and $\mathcal{P}_{\text{exch}}$ of i.i.d. and exchangeable distributions, respectively. Recall, the arguments in Section 3 establish that $\mathcal{P}_{\text{exch}}$ is rich enough to admit $U = \lbag Z \rbag$ as a boundedly complete statistic according to the one-sided definition in (19). But the above example shows that this is not true for $\mathcal{P}_{\text{iid}}$, and $U = \lbag Z \rbag$ is only boundedly complete according to the weaker two-sided definition in (43).

A.3 Proof of claim (22)

We assume $\alpha < 1$, as otherwise the claim is vacuous. Note first that the " $\Longleftarrow$ " direction of the claim is due to the validity of permutation tests over each exchangeable distribution of the form $\frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \delta_{z_\sigma}$.

For the " $\Longrightarrow$ " direction of the claim, let ϕ satisfy (21). Define a test statistic by $T = \phi$ and permutation p-value p by (13). We will show that under this construction $\phi(Z) = 1\{p \leq \alpha\}$, and proceed to check this in two cases.

- If $\phi(Z) = 1$, then

$$p = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} 1\{\phi(Z_\sigma) \geq 1\} = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} \phi(Z_\sigma) \leq \alpha.$$

- If $\phi(Z) = 0$, then

$$p = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} 1\{\phi(Z_\sigma) \geq 0\} = 1 > \alpha.$$

This completes the proof.

A.4 Proof of Proposition 2

Following Chapter 17.2.1 in Lehmann and Romano (2022), let $T_{(1)} \leq \cdots \leq T_{(N)}$ denote the ordered values of $T(z_\sigma)$ over $\sigma \in \mathcal{S}_{n+1}$, where $N = (n+1)!$. Let $k = N - \lfloor N\alpha \rfloor$, where $\lfloor N\alpha \rfloor$ is the largest integer less than or equal to $N\alpha$, and define

$$\phi(z_\sigma) = \begin{cases} 1 & \text{if } T(z_\sigma) > T_{(k)} \\ 0 & \text{if } T(z_\sigma) < T_{(k)}. \end{cases}$$

We see that (35) is met for $k_u = T_{(k)}$. Further, for $N_+ = |\{i : T_{(i)} > T_{(k)}\}|$ and $N_= = |\{i : T_{(i)} = T_{(k)}\}|$, set

$$\phi(z_\sigma) = \frac{N\alpha - N_+}{N_=} \quad \text{if } T(z_\sigma) = T_{(k)}.$$

Then by Theorem 17.2.1 in Lehmann and Romano (2022), the test ϕ has level exactly α, as in (36).

It remains to show that ϕ defined above is equivalent to the test ϕ^* defined in the proposition; precisely, we will show $\phi = \mathbb{E}_V[\phi^*]$, where the expectation is over V. For this, it helps to rewrite ϕ^* in terms of its action over z_σ, $\sigma \in \mathcal{S}_{n+1}$. This is $\phi^*(z_\sigma) = 1\{p^*(z_\sigma) \leq \alpha\}$, where

$$p^*(z_\sigma) = \frac{N_+^\sigma + V \cdot N_=^\sigma}{N},$$

and $N_+^\sigma = |\{i : T_{(i)} > T(z_\sigma)\}|$, $N_=^\sigma = |\{i : T_{(i)} = T(z_\sigma)\}|$. We proceed in cases. Below "almost sure" should be interpreted with respect to V.

- If $N_+^\sigma + N_=^\sigma \leq N\alpha$, then $p^* \leq \alpha$ almost surely and $\mathbb{E}_V[\phi^*(z_\sigma)] = 1$. In this case, we also know at most $\lfloor N\alpha \rfloor$ statistics are $\geq T(z_\sigma)$, or equivalently, more than $N - \lfloor N\alpha \rfloor = k$ statistics are $< T(z_\sigma)$, thus $T(z_\sigma) > T_{(k)}$ and $\phi(z_\sigma) = 1$.
- If $N_+^\sigma > N\alpha$, then $p^* > \alpha$ almost surely and $\mathbb{E}_V[\phi^*(z_\sigma)] = 0$. Also, in this case, at least $\lfloor N\alpha \rfloor + 1 = N - k + 1$ statistics are $> T(z_\sigma)$, thus $T(z_\sigma) < T_{(k)}$ and $\phi(z_\sigma) = 0$.
- If $N_+^\sigma \leq N\alpha < N_+^\sigma + N_=^\sigma$, then

$$\mathbb{E}_V[\phi^*(z_\sigma)] = \mathbb{P}\Big\{N_+^\sigma + V \cdot N_=^\sigma \leq N\alpha\Big\} = \frac{N\alpha - N_+^\sigma}{N_=^\sigma}.$$

  In this case, $T(z_\sigma) = T_{(k)}$, so $N_+ = N_+^\sigma$, $N_= = N_=^\sigma$, and finally $\phi(z_\sigma) = \mathbb{E}_V[\phi^*(z_\sigma)]$.

This completes the proof.

A.5 Proof of Theorem 7, general case

In the general case, where μ is σ-finite but not necessarily finite, we can adapt the arguments presented in finite case, as follows. Recall the representation (23). Since $p_{Y|X=x}(y) > 0$ for all y and P_X-almost every x, we can rewrite this as

$$\mathbb{E}_P[\mu(C_n(X_{n+1}))] = \mathbb{E}_P\left[(1 - \phi(Z)) \cdot \frac{1}{p_{Y|X=X_{n+1}}(Y_{n+1})}\right].$$

Next, by the total law of expectation

$$\mathbb{E}_P[\mu(C_n(X_{n+1}))] = \mathbb{E}_{P_U}\Bigg[\underbrace{\mathbb{E}_P\left[(1 - \phi(Z)) \cdot \frac{1}{p_{Y|X=X_{n+1}}(Y_{n+1})} \,\middle|\, U\right]}_{h(\phi,U)}\Bigg],$$

where by exchangeability the conditional expectation is

$$h(\phi, u) = \frac{1}{(n+1)!} \sum_{\sigma \in \mathcal{S}_{n+1}} (1 - \phi(z_\sigma)) \cdot \frac{1}{p_{Y|X=x_{\sigma(n+1)}}(y_{\sigma(n+1)})}.$$

Fix $u = \lceil z \rfloor$, and define Q_u as in (34). Observe that

$$h(\phi, u) = \frac{1}{c_u(n+1)!} \cdot \mathbb{E}_{Q_u}[1 - \phi(Z)].$$

Hence, even in the general case (where Q itself is not necessarily well-defined), we have reduced the problem to one of minimizing type II error against Q_u. By the same arguments as given earlier, the test ϕ^* in Proposition 2 has optimal type II error among all valid level α tests for $\mathcal{P}_{\text{exch}}$ versus Q_u. That is, for any test ϕ which is level α for $\mathcal{P}_{\text{exch}}$,

$$h(\phi^*, u) \leq h(\phi, u), \quad \text{for all } u.$$

Furthermore, since ϕ^* does not depend on u, we can integrate over U to obtain

$$\mathbb{E}_P[\mu(C_n^*(X_{n+1}))] \leq \mathbb{E}_P[\mu(C_n(X_{n+1}))].$$

This completes the proof.

A.6 Further details for Figure 1

We generated $n = 500$ i.i.d. samples (X_i, Y_i), $i = 1, \ldots, n$, where each $X_i \sim \text{Unif}[0, 2\pi]$, and

$$Y_i | X_i \sim N(f(X_i), \sigma^2(X_i)), \quad \text{where } f(x) = \sin(x) \text{ and } \sigma(x) = \pi x/30.$$

These are the gray dots in Figure 1. For both the efficient score (41) and studentized score (42), we used the true mean and variance functions, i.e., $\hat{f} = f$ and $\hat{\sigma} = \sigma$. We then pursued the conformal construction in (3), (5) at the coverage level $1 - \alpha = 0.9$.

In this setup, due to our use of oracle mean and variance functions (no model training), the score s_i^y does not depend on y for $i \leq n$, for either studentized or efficient score functions. This allows us to compute the conformal sets in closed-form. For studentized score (42), this is

$$C_n(x) = [f(x) - \sigma(x)\hat{q}_n,\ f(x) + \sigma(x)\hat{q}_n],$$

where $\hat{q}_n$ is the $\lceil (n+1)(1-\alpha) \rceil^{\text{th}}$ smallest of $s_1, \ldots, s_n$. For the efficient score (41), this is

$$C_n(x) = \Big[f(x) - \sigma(x)\sqrt{2(\hat{q}_n - \log \sigma(x))},\ f(x) + \sigma(x)\sqrt{2(\hat{q}_n - \log \sigma(x))}\Big],$$

where again $\hat{q}_n$ is the $\lceil (n+1)(1-\alpha) \rceil^{\text{th}}$ smallest of $s_1, \ldots, s_n$.

The form above makes it clear why, for the most part, the efficient intervals (red) are narrower than the studentized intervals (blue) in Figure 1. This is true except in regions where $\sigma(x)$ is very small, in which case $\log \sigma(x) < 0$. In this simulation, the studentized intervals are on an average about 12.5% wider than the efficient intervals.