Full Conformal Prediction under Stochastic Non-Conformity Measure

Thanawat Sornwanee [1]

Abstract

The theory of full conformal prediction uses deterministic non-conformity measure, but modern usage of full conformal prediction often relies on machine learning training, making stochasticity inevitable. A simple sufficient condition of almost sure permutation invariance of the non-conformity measure can be too restrictive, so many have suggested the relaxation to permutation in distribution as a condition for full conformal prediction validity. We, however, show that this commonly known condition is actually insufficient. We then provide a correct sufficient condition: *Conditional Independence & Permutation Invariance in Distribution*, which encompasses several stochastic settings that may be used in machine learning.

1. Full conformal prediction

Given the dataset of

$$\left([z_i = (x_i, y_i)]_{i=1}^{N}, x_{N+1}\right),$$

where x_i is the input and y_i is the output, we are interested in constructing a confidence bound on the unobserved output y_{N+1}.

A standard conformal prediction relies on the routine of retrieving conformal p-value from the rank of the non-conformity score. Afterwards, the conformal prediction returns a confidence set

$$\left\{\hat{y} \in \mathcal{Y} : \text{conformal p-value}\left([z_i]_{i=1}^{N}, (x_{N+1}, \hat{y})\right) \geq \alpha\right\}.$$

In full conformal prediction setting, we are given a non-conformity score function $t : \mathcal{Z}^{N+1} \to \mathcal{S}$, where $\mathcal{S}$ is a totally preordered set endowed with a binary relation[1] $\succcurlyeq$, and is normally chosen to be $(\mathbb{R}, \geq)$. (Vovk et al., 2005; 2016) We the compute the score in leave-one-out manner, and then compute the conformal p-value from the rank of the score as outlined in the algorithm 1.

Algorithm 1 Conformal p-Value via Deterministic t

Require: Full Data $z \in \mathcal{Z}^{N+1}$

1: **for** $i = 1$ to $N + 1$ **do**

2: Compute Non-Conformity Score

$$s_i = t\left(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{N+1}, z_i\right)$$

3: **end for**

4: Compute Conformal p-Value

$$P = \frac{1}{N+1}\left[1 + \sum_{i=1}^{N} \mathbf{1}_{S_{N+1} \succcurlyeq S_i}\right]$$

5: **return** P

The result is that, if $Z = [Z_i = (X_i, Y_i)]_{i=1}^{N+1}$ is exchangeable, and the non-conformity score function t is permutation invariant in the first N arguments, then the conformal p-value of the actual data Z is a valid p-value[2]. This implies that the actual output Y_{N+1} is in the confidence set with respect to $\left([Z_i]_{i=1}^{N}, X_{N+1}\right)$ with probability $\geq 1 - \alpha$, where the randomness is from the stochasticity of Z itself.

In most cases, the function t uses the first N data entries, which is $\{z_j\}_{j \neq i}$, symmetrically to train a regression model[3] $\hat{f}_{-i} : \mathcal{X} \to \mathcal{Y}$, and use the loss $\left\|y_i - \hat{f}_{-i}(x)\right\|$ to be the non-conformity score S_i.

Note that the trained model $\hat{f}_{-i}$ does not have to be good or satisfy any conditions apart from that it training process must be deterministic and permutation invariant in the data input $\{z_j\}_{j \neq i}$. For example, a ridge regression or machine learning with fixed initialization and stochastic descent will satisfy this.

[1] Stanford University. Correspondence to: Thanawat Sornwanee <>.

ICML 2026 Hypothesis Testing Workshop, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

[1] which is reflexive, transitive, and complete.

[2] See appendix A for a formal definition.

[3] If the training returns more estimators, such as variance estimator or quantile estimator, one can modify the score to incorporate them. (Papadopoulos et al., 2002; Romano et al., 2019)

Permutation invariance seems to be natural to most modern machine learning algorithm, but deterministic training requirement is often violated. Modern applications of full conformal prediction often rely on a training of a machine learning model, which may involve stochasticity via stochastic gradient descent as well as random initialization. (Lee & Zhang, 2025; Tailor et al., 2025) Some may also employ different stochastic training model such as random forests. (Linusson et al., 2014; Gauraha & Spjuth, 2018)

Apart from the stochasticity in the training, we will see that the stochasticity in the inference time is also not uncommon. The emergence usage of modern LLM (large language model) as well as other generative model can have stochasticity in the inference such as from the API call itself.[4] (Angelopoulos & Bates, 2022) suggest adding a small white noise to each score can also be used for tie-breaking.

This suggests that a better understanding of full conformal prediction under a stochastic non-conformity measure is crucial.

2. Stochasticity

Before discussing the prior works, we will provide some formulation of full conformal prediction under stochastic measure/score. This is when we replace the routine of computing conformal p-value from using a deterministic function $t : \mathcal{Z}^{N+1} \to \mathcal{S}$ to stochastic function(s) T. Note that we will assume that every function mentioned has to be measurable throughout.

Since we may evaluate different non-conformity score with different randomness[5], we then need to provide flexibility for the stochastic function of each coordinate to be different. Formally, we have the algorithmic random space $(\Omega, \mathcal{F}, \mathbb{P}_{\text{alg}})$, which will throughout be assumed to be independent from that generating data, and a vector of stochastic functions $T = [T_i]_{i=1}^{N+1}$ where $T_i : \mathcal{Z}^{N+1} \times \Omega \to \mathcal{S}$ for each $i \in \{1, 2, \ldots, N+1\}$.

The routine of computing non-conformity score becomes the algorithm 2. Note that, for a given z, the p-value P still inherits randomness from the algorithm.

The question of interest is then

[4]Currently, the usage of LLM as well as other pretrained models in conformal prediction is more limited to a split conformal setting. (Ravfogel et al., 2023; Quach et al., 2024; Su et al., 2024; Epstein et al., 2026) However, one can imagine the use of full conformal prediction by doing a fine-tuning steps, or other post-training alignment methods, with the input data. Though this has not been extensively studied

[5]For example, ω may dictate 3 independently drawn random seed ϵ_1, ϵ_2, and ϵ_3. The training process to get score 1 may only utilize the seed ϵ_1, while that of the score 2 only uses ϵ_2. Therefore, mathematically, we need to put the indicator for T.

Algorithm 2 Conformal p-Value via Stochastic T

Require: Full Data $z \in \mathcal{Z}^{N+1}$, Algorithmic Random Space $(\Omega, \mathcal{F}, \mathbb{P}_{\text{alg}})$

Realize $\omega \sim \mathbb{P}_{\text{alg}}$
for $i = 1$ to $N+1$ **do**
  Compute Stochastic Non-Conformity Score

$$S_i = T_i\left(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{N+1}, z_i; \omega\right)$$

end for
Compute Stochastic Conformal p-Value

$$P = \frac{1}{N+1}\left[1 + \sum_{i=1}^{N} \mathbf{1}_{S_{N+1} \succcurlyeq S_i}\right]$$

return P

What is a sufficient condition on the random functions $T = [T_i]_{i=1}^{N+1}$ for the validity of the conformal prediction?

This is when the actual observation Y_{N+1} is included in the confidence set with probability $\geq 1 - \alpha$, where now the randomness comes from both the data Z itself and the algorithm. This is equivalent to asking for a sufficient condition for the conformal p-value in the algorithm 2 to be a valid p-value whenever the random input Z is an exchangeable random vector. Formally, for any exchangeable Z, we require, for all $\alpha \geq 0$,

$$\mathbb{P}\left(\text{conformal p-value}\left(Z; \omega\right) \leq \alpha\right) \leq \alpha.$$

2.1. Prior Works

As discussed, the stochasticity in full conformal prediction is not uncommon, but the theoretical study is lacking. To the author's knowledge, only 5 papers explicitly mention about condition for stochastic conformal score. The mentions are brief: 3 of them are in footnotes. 4 claims are also incorrect.

Incorrect Claims The paper "Conformal Prediction Beyond Exchangeability"(Barber et al., 2023) has claimed that in a case where the stochastic T is such that

$$\begin{aligned} &T_i(z_1, z_2, \ldots, z_{i-1}, z_{i+1}, z_{i+2}, \ldots, z_{N+1}, z_i; \omega) \\ &= \left(\mathcal{A}(z_1, z_2, \ldots, z_{i-1}, z_{i+1}, z_{i+2}, \ldots, z_{N+1}; \omega)\right)(z_i) \end{aligned}$$

for some random algorithm $\mathcal{A}$ mapping $z_{-i} \in \mathcal{Z}^N$ to a function $\hat{\mu} : \mathcal{X} \to \mathcal{Y}$, which is assumed to be $\mathbb{R}$, then we only require

$$\mathcal{A}\left(z; \cdot\right) \overset{\text{d}}{=} \mathcal{A}\left(\left[z_{\sigma(i)}\right]_{i=1}^{N}; \cdot\right)$$

for all $z \in \mathcal{Z}^{N+1}$ and for all permutation $\sigma \in S_N$.[6] The same permutation symmetry in distribution sense is also claimed to be a sufficient condition in the paper "Training-conditional coverage for distribution-free predictive inference"(Bian & Barber, 2023)[7]and by (Lee et al., 2023)[8]. This condition implies that, for all $i, j \in \{1, 2, \ldots, T+1\}$, $z \in \mathcal{Z}^{N+1}$, $\sigma \in S_N$,

$$T_i(z; \cdot) \overset{\mathrm{d}}{=} T_j\left(\left[z_{\sigma(k)}\right]_{k=1}^{N}, z_{N+1}; \cdot\right),$$

and $T_i = T_j$. Note that the equation can be further decomposed into two parts: 1.) within the same index, equality in distribution across permutation and 2.) equality in distribution across different indices. The latter is implied by $T_i = T_j$, but we keep it as two separate conditions since we will soon relax the equality.

This condition is claimed to be a sufficient condition by (Bai & Jin, 2024).[9]

However, all these 4 papers are not correct, and we will give out a simple counterexample in the section 3.

Correct but Narrow Claim (Lee et al., 2025) seem[10] to claim that a sufficient condition is when, in an almost sure sense,

$$T_i(z; \omega) = T_j\left(\left[z_{\sigma(k)}\right]_{k=1}^{N}, z_{N+1}; \omega\right).$$

Similarly, this condition can be decomposed into two components: almost sure equality across indices, and almost sure equality among permutation in the same index.

This is obviously sufficient but quite strong. For example, it cannot handle stochastic gradient descent in a usual sense.[11]

3. Why equality in distribution is insufficient?

Below we will show a simple example where the equality in distribution under permutation is satisfied for the training algorithm $\mathcal{A}$, so the equality in distribution under permutation of the first N element is satisfied for the stochastic scoring functions T, but the the full conformal prediction is invalid.

Consider the case where $(X_i, Y_i) \overset{\text{iid}}{\sim} \text{Unif}\left([0,1]^2\right)$, and $N = 2$. Thus, the random vector $Z = (Z_1, Z_2, Z_3) \in \left([0,1]^2\right)^3$ is exchangeable.

Let the random space be such that we can define two independent standard Brownian motions $\{B_t(\omega)\}_{t\in[0,1]}$ and $\{B'_t(\omega)\}_{t\in[0,1]}$, and we define a random scoring function for all $i \in \{1, 2, 3\}$ to be

$$T_i\left(z; \omega\right) := \left(\mathcal{A}\left(z_1, z_2; \omega\right)\right)\left(z_3\right) := B_{y_2}(\omega) + B'_{1-y_2}(\omega),$$

for any $z \in \left([0,1]^2\right)^3$, thereby making

$$T\left(z; \cdot\right) \sim N(0,1),$$

and similarly $\mathcal{A}$ is permutation invariant in distribution, since it always return a constant function whose value is distributed according to $N(0,1)$.

The non-conformity scores from the algorithm 2 conditioned on the data Z are

$$\begin{cases} \left(S_1|Z\right)(\cdot) & = T\left(Z_2, Z_3, Z_1; \cdot\right) = B_{Y_3} + B'_{1-Y_3} \\ \left(S_2|Z\right)(\cdot) & = T\left(Z_1, Z_3, Z_2; \cdot\right) = B_{Y_3} + B'_{1-Y_3} \\ \left(S_3|Z\right)(\cdot) & = T\left(Z_1, Z_2, Z_3; \cdot\right) = B_{Y_2} + B'_{1-Y_2} \end{cases},$$

making the conformal p-value

$$\begin{aligned} \left(P|Z\right)(\cdot) &= \left(\frac{1 + \mathbf{1}_{S_3 \geq S_1} + \mathbf{1}_{S_3 \geq S_2}}{3}\middle| Z\right) \\ &= \frac{1}{3} + \frac{2}{3}\left(\mathbf{1}_{B_{Y_2} + B'_{1-Y_2} \geq B_{Y_3} + B'_{1-Y_3}}\middle| Z\right). \end{aligned}$$

Under the case when $Y_2 \neq Y_3$, which happens almost surely, we have that $B_{Y_2} + B'_{1-Y_2} \geq B_{Y_3} + B'_{1-Y_3}$ with

[6]The footnote 2 of (Barber et al., 2023) states that "If $\mathcal{A}$ is a randomized algorithm, then this equality is only required to hold in a distributional sense," where the original equality refers to $\mathcal{A}(z) = \mathcal{A}\left(\left[z_{\sigma(i)}\right]_{i=1}^{N}\right)$, which is a commonly used sufficient condition in deterministic full conformal prediction.

[7]The subsubsection 2.1.3 "A note on randomized algorithm" of (Bian & Barber, 2023) states that "The background given above implicitly the algorithm $\mathcal{A}$ as a deterministic function of the training data—that is, we view $\mathcal{A}$ as a function $((X_1, Y_1), \ldots, (X_n, Y_n)) \mapsto \hat{\mu}$. In many settings, however, it is common to use a randomized regression algorithm-for instance, stochastic gradient descent. In this setting, we can formally view $\mathcal{A}$ as a function $((X_1, Y_1), \ldots, (X_n, Y_n), \xi) \mapsto \hat{\mu}$, where the term ξ introduces stochastic noise (effectively, a random seed). All the results described above hold for both the deterministic and randomized settings. (For results that assume $\mathcal{A}$ is symmetric, the symmetry condition (6) should be understood in the distributional sense-that is, the training data points are treated symmetrically with respect to the randomized training procedure. For example, for stochastic gradient descent, if data points are drawn uniformly at random during the training epochs, then symmetry is satisfied.)"

[8]The footnote 12 of (Lee et al., 2023) mentions that "The framework also allows for a randomized algorithm $\mathcal{A}$, in which case the symmetry condition is required to hold in a distribution sense."

[9]The footnote 1 of (Bai & Jin, 2024) mentions that "In the general case where $\mathcal{V}$ is a randomized algorithm, we require ... Definition 4 to hold in a distributional sense, i.e., $=$ could be replaced by $\overset{\mathrm{d}}{=}$," where the equality refers to the one shown here.

[10]The proof of the theorem 2 of (Lee et al., 2025) states that "The scoring function $V^{(k)}(\cdot)$ was trained invariant to the order of elements inside E_j (this assumes that $V^{(k)}$ does not use external randomness during training; similar results can be attained for random training procedures by conditioning on the random seed)," $V^{(k)}$ is equivalent to T_i in our notation.

[11]We provide some examples where this is satisfied in the subsection 4.2.

probability $\frac{1}{2}$ conditioned on Y_2 and Y_3 (so the randomness is from the algorithmic randomness). Therefore, $(P|Z)(\cdot) \sim \frac{1}{2}\delta_{\frac{1}{3}} + \frac{1}{2}\delta_1$ almost surely in Z (not in the algorithmic random space). Thus, almost surely in Z,

$$\mathbb{P}\left(P \leq \frac{1}{3}\right) = \mathbb{E}\left[\mathbb{P}\left(P \leq \frac{1}{3}\middle| Z\right)\right] = \frac{1}{2} > \frac{1}{3},$$

contradicting with that $\mathbb{P}\left(P \leq \frac{1}{3}\right) \leq \frac{1}{3}$ for P to be a valid p-value. Thus, P cannot be a valid p-value.

Thus, the sufficiency claims in (Barber et al., 2023; Bian & Barber, 2023; Lee et al., 2023; Bai & Jin, 2024) are not correct.[12]

4. Examples of sufficient conditions

Recall that the validity of conformal prediction comes from the validity of permutation test. Thus, we only need that, as long as the data points Z are exchangeable, the non-conformity scores $S := [S_i]_{i=1}^{N+1}$ are also exchangeable (unconditionally).

4.1. Deterministic with Permutation Invariance

For deterministic case, a simple sufficient condition is permutation invariant among first N arguments, meaning that

$$T_i(z;\omega) = T_j(\left[z_{\sigma(k)}\right]_{k=1}^{N}, z_{N+1};\omega') \tag{1}$$

for all $i, j \in \{1, 2, \ldots, N+1\}$, $z \in \mathcal{Z}^{N+1}$, $\sigma \in S_N$, and for almost surely ω, ω'. This is a standard condition, and can be achieved by symmetrization over a deterministic function $f : \mathcal{Z}^{N+1} \to \mathbb{R}$ such as choosing $T_i(z;\omega) = \frac{1}{N!}\sum_{\sigma\in S_N} f\left(z_{\sigma(1)}, z_{\sigma(2)}, \ldots, z_{\sigma(N)}, z_{N+1}\right)$, but this requires $N!$ computation.

4.2. Stochastic with Almost Sure Permutation Invariance

As mentioned, this is when we weaken the degeneracy requirement (by dropping ω') from the condition 1 into

$$T_i(z;\omega) = T_j(\left[z_{\sigma(k)}\right]_{k=1}^{N}, z_{N+1};\omega) \tag{2}$$

almost surely. This is still hard to achieve, as it is translated into that the permutation invariance holds for each random seed ω, and we have to reuse the same seed for each score evaluation process. For example, this can be satisfied when the randomness only comes in through random initialization of neural network parameters and full gradient descent is implemented. Another example is when symmetrization is used with the same fixed seed.

[12]It is easier to construct a counter example when we allow the random function for each score computation to use different part of the randomness. For example, we can consider the case when $\omega \sim \text{Unif}([0,1]^2)$, and $S_1 = S_2 = \omega_1$, while $S_3 = \omega_2$.

For most machine learning training, this rarely holds. Stochastic gradient descent with mini-batch[13] may sample a random index $I(\omega)$ of the data to be used in the gradient computation. By applying the permutation before running the algorithm, the $I(\omega)^{\text{th}}$ datapoint in the permuted dataset will correspond to the $\sigma\left(I(\omega)\right)^{\text{th}}$ datapoint of the original dataset.

Moreover, this condition is difficult to satisfy if we have randomness in the inference.

4.3. Stochastic with Independence & Permutation Invariance in Distribution

We weaken the almost sure equality 2 to equality in distribution

$$T_i\left(z;\cdot\right) \overset{\text{d}}{=} T_j\left(\left[z_{\sigma(k)}\right]_{k=1}^{N}, z_{N+1};\cdot\right), \tag{3}$$

but with additional independence condition that

$$\left\{T_i\left(z_1, z_2, \ldots, z_{i-1}, z_{i+1}, z_{i+2}, \ldots, z_{N+1}, z_i;\cdot\right)\right\}_{i=1}^{N+1}$$

is a collection of independent variables.

This is perhaps the most commonly used method in full conformal prediction. Each time we train a machine learning model, the stochasticity is assumed to be independent from those in the past. If there exists randomness in the inference, as long as such randomness are independent from each other, then the validity of full conformal prediction will still held. However, we can see that this can be violated in some case, such as when each training share the same random seed, which may be used for reproducibility.

5. General Sufficient Condition: Conditional Independence & Permutation Invariance in Distribution

Note that the deterministic case 4.1 is a special case for both stochastic cases 4.2 and 4.3, but the two stochastic cases are not a special case of each other. We generalize the combination of the two stochastic cases to show that a more general condition of conditional independence and permutation invariance in distribution is sufficient for full conformal prediction validity.

Definition 1 (Conditional Independence & Permutation Invariance in Distribution)**.** *The stochastic non-conformity score functions $T = [T_i]_{i=1}^{N+1}$, where $T_i : \mathcal{Z}^{N+1} \times \Omega \to \mathcal{S}$ for each $i \in \{1, 2, \ldots, N+1\}$ are conditionally independent and permutation invariant in distribution if there exists some random variable C defined on the algorithmic random*

[13]Unless we enforce the algorithm to re-sort the datapoints first before using any stochasticity.

space such that, for all $z \in \mathcal{Z}^{N+1}$ and for almost surely C,

$$(T_i(z;\cdot)|C) \overset{d}{=} \left(T_j\left(\left[z_{\sigma(k)}\right]_{k=1}^{N}, z_{N+1};\cdot\right)\middle|C\right) \quad (4)$$

for all $\sigma \in S_N$, for all $i, j \in \{1, 2, \ldots, N+1\}$, and the collection of random variables

$$\{T_i(z_1, z_2, \ldots, z_{i-1}, z_{i+1}, z_{i+2}, \ldots, z_{N+1}, z_i;\cdot)\}_{i=1}^{N+1}$$

is independent conditioned on C.

By taking C to be non-informative, this reverts back to the independence and permutation invariance in distribution setting 4.3, and, by taking C to be ω, this reverts back to the almost sure permutation invariance 4.2.

Theorem 1. *The conformal p-value whose stochastic non-conformity score functions T are conditionally independent and permutation invariant in distribution is valid.*

The proof[14], first, shows that the joint score computation, with $z_{-i} := (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{N+1})$,

$$A(z;\omega) := [T_i(z_{-i}, z_i;\omega)]_{i=1}^{N+1}$$

is permutation equivariance in joint distribution. This means that

$$\left[(A(z;\cdot))_{\pi(i)}\right]_{i=1}^{N+1} \overset{\mathrm{d}}{=} A(\left[z_{\pi(i)}\right]_{i=1}^{N+1};\cdot)$$

for all $z \in \mathcal{Z}^{N+1}$ and $\pi \in S_{N+1}$.

Afterwards, we use exchangeability of Z to conclude that the scores

$$S = A(Z;\omega)$$

are also exchangeable.

This condition is general enough that could be used to satisfied most valid use case of full conformal prediction. For example, one can imagine when the random seed for the random initialization of the neural network parameters are fixed, while the random seeds for the order of stochastic gradient descent is redrawn each time. In such case, the condition C will represent the initialization of the neural network parameters.

It is also easy to see that this is not a necessary condition. For example, we can replace with the condition like conditional exchangeability. This condition will imply the condition "Conditional Independence & Permutation Invariance in Distribution" we introduced but at the cost of harder verification.

Another sufficient condition will be to revert to only use deterministic score function/measure but augmenting the inputs z with white noise. (Vovk et al., 2005) It can be seen that this is not implying nor implied by our condition. This is because the conditional independence requirement is only "on the path", and does not require every possible permutations in a two-dimensional table to satisfy the conditional independence.

[14] which is in the appendix C

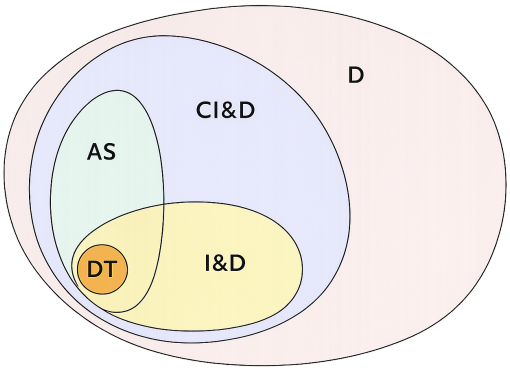

Figure 1. *DT* stands for deterministic permutation invariance, which is a standard assumption in full conformal prediction as outlined in the subsection 4.1. (Vovk et al., 2005) *AS* stands for almost sure permutation invariance as outlined in the subsection 4.2 *I&D* stands for independence and permutation invariance in distribution as outlined in the subsection 4.3. *CI&D* stands for conditional independence and permutation invariance in distribution as defined in the definition 1. We have established through the main theorem 1 that *CI&D* is sufficient to ensure the validity of full conformal prediction when the non-conformity measure is stochastic. *D* stands for permutation invariance in distribution, which is the incorrect sufficient condition in the literature (Barber et al., 2023; Bian & Barber, 2023; Lee et al., 2025; Bai & Jin, 2024). In this figure, we show that DT $\subseteq$ AS $\cap$ I&D, while AS $\cup$ I&D $\subseteq$ CI&D $\subseteq$ D. Note that these conditions are defined with respect to an algorithmic random space $(\Omega, \mathcal{F}, \mathbb{P}_{\text{alg}})$ as well as $\mathcal{S}$ and $\mathcal{Z}$. It is easy to show that, for each $\subseteq$ relation, we can find a setting such that every $\subseteq$ relation is indeed $\subsetneq$.

6. Conclusion

Our paper shows that the long incorrectly believed assumption for the validity of full conformal prediction is however incorrect. We show it by a simple construction with a generous interpretation of the claims made in the literature. We then provide a sufficient condition.

Acknowledgement

The author thanks Lihua Lei and Vladimir Vovk for useful discussions.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty

quantification, 2022. URL https://arxiv.org/abs/2107.07511.

Bai, T. and Jin, Y. Optimized conformal selection: Powerful selective inference after conformity score optimization. *arXiv preprint arXiv:2411.17983*, 2024.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Bian, M. and Barber, R. F. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023.

Epstein, E. L., Winnicki, J., Sornwanee, T., and Dwaraknath, R. V. LLMs are overconfident: Evaluating confidence interval calibration with fermieval. In *AAAI 2026 Workshop on Assessing and Improving Reliability of Foundation Models in the Real World*, 2026. URL https://openreview.net/forum?id=yUyFITL0wv.

Gauraha, N. and Spjuth, O. conformalclassification: A conformal prediction r package for classification, 2018. URL https://arxiv.org/abs/1804.05494.

Lee, J., Popov, I., and Ren, Z. Full-conformal novelty detection: A powerful and non-random approach. *arXiv preprint arXiv:2501.02703*, 2025.

Lee, K. and Zhang, Y. Leave-one-out stable conformal prediction. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Bt1vnCnAVS.

Lee, Y., Barber, R. F., and Willett, R. Distribution-free inference with hierarchical data. *arXiv preprint arXiv:2306.06342*, 2023.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer, 2005.

Linusson, H., Johansson, U., Boström, H., and Löfström, T. Efficiency comparison of unstable transductive and inductive conformal classifiers. In *IFIP International conference on artificial intelligence applications and innovations*, pp. 261–270. Springer, 2014.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzUhfQ74c5.

Ravfogel, S., Goldberg, Y., and Goldberger, J. Conformal nucleus sampling. In *Findings of the association for computational linguistics: ACL 2023*, pp. 27–34, 2023.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Su, J., Luo, J., Wang, H., and Cheng, L. Api is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 979–995, 2024.

Tailor, D., Correia, A., Nalisnick, E., and Louizos, C. Approximating full conformal prediction for neural network regression with gauss-newton influence. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vcX0k4rGTt.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer, 2005.

Vovk, V., Fedorova, V., Nouretdinov, I., and Gammerman, A. Criteria of efficiency for conformal prediction. In *Symposium on conformal and probabilistic prediction with applications*, pp. 23–39. Springer, 2016.

A. p-Value & permutation test

We define a valid p-value, and provide background on permutation test, which is the backbone conformal prediction.

Definition 2 (Valid p-Value)**.** *A random variable P is a valid p-value if*

$$\mathbb{P}\left(P \leq \alpha\right) \leq \alpha$$

for all $\alpha \geq 0$.

Lemma 1. *Let $S := [S_i]_{i=1}^n$ be an exchangeable random variable where $S_i \in \mathcal{S}$ for each $i \in \{1, 2, \ldots, n\}$ and $(\mathcal{S}, \succcurlyeq)$ is a total preorder. Define*

$$P_1 := \frac{\sum_{i=1}^n \mathbf{1}_{S_1 \succcurlyeq S_i}}{n}.$$

Then its distribution satisfies

$$\mathcal{L}(P_1) \succcurlyeq_{\text{FOSD}} \text{Unif}\left(\left\{\frac{1}{n}, \ldots, 1\right\}\right) \succcurlyeq_{\text{FOSD}} \text{Unif}([0, 1]).$$

Proof. Define $P_i := \frac{\sum_{j=1}^n \mathbf{1}_{S_i \succcurlyeq S_j}}{n}$. Exchangeability of S implies exchangeability of P. Note that, for any integer $k \in \{1, 2, \ldots, n\}$, there can be at most k values of $i \in$

$\{1, 2, \ldots, n\}$ where $P_i \leq \frac{k}{n}$. Hence,

$$\mathbb{P}(P_1 \leq k/n) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{P_i \leq k/n}\right] \leq \frac{k}{n},$$

implying the first order stochastic dominations. □

Corollary 1. *P_1 in the lemma 1 is a valid p-value if S is exchangeable.*

Proof. From the lemma 1, we have that $\mathbb{P}(P_1 \leq \alpha) \leq (\text{Unif}([0,1]))([0,\alpha]) = \alpha$, so P_1 is a valid p-value. □

In permutation testing (Lehmann & Romano, 2005), we test a hypothesis of whether random variables $Z := [Z_i]_{i=1}^n$ are exchangeable, by randomly permuting the variables and compute test statistic each time. Under such hypothesis, the test statistics should be exchangeable, so $P_1 \leq \alpha$ with probability $\leq \alpha$, meaning that it is a valid p-value.

B. Permutation Equivariance in Distribution algorithm

To prove the main theorem 1, we recall that Z is an exchangeable random variable whose stochasticity is independent from that of the stochastic algorithm and $(\mathcal{S}, \succcurlyeq)$ is already assumed to be a total preorder. Thus, we only need to ensure that a stochastic algorithm $A : \mathcal{Z}^{N+1} \times \Omega \to \mathcal{S}^{N+1}$ that jointly return the score

$$S := [S_i]_{i=1}^{N+1} := A(Z; \omega)$$

will be able to transfer the exchangeability in Z into the exchangeability in S, both of which in a distribution sense. We can then construct a p-value

$$P := \frac{1}{N+1}\left[1 + \sum_{i=1}^{N} \mathbf{1}_{S_{N+1} \succcurlyeq S_i}\right],$$

and get its validity by the corollary 1.

In this section, we first show that permutation equivariant in joint distribution is a sufficient condition (in the definition 3) for the stochastic algorithm A to preserve exchangeability. Next, we will show a condition on the non-conformity score functions T to achieve permutation equivariant in joint distribution.

B.1. Definition & property

A convenient sufficient condition for exchangeability to be inherited through A is that the stochastic algorithm A is permutation equivariant in joint distribution.

Definition 3 (Permutation Equivariance in Joint Distribution)**.** *A random function $A : \mathcal{Z}^{N+1} \times \Omega \to \mathcal{S}^{N+1}$ is permutation equivariant in joint distribution if, for any permutation $\pi \in S_{N+1}$, for any $z \in \mathcal{Z}^{N+1}$,*

$$A\left(\left[z_{\pi(i)}\right]_{i=1}^{N+1}; \cdot\right) \overset{d}{=} \left[(A(z;\cdot))_{\pi(i)}\right]_{i=1}^{N+1}.$$

Recall that exchangeable is the same as permutation invariance in distribution. It is then unsurprising that $S = A(Z; \omega)$ can inherit the exchangeability[15] in distribution from Z.

Lemma 2 (Exchangeability & Permutation Equivariance imply Exchangeability)**.** *If $A : \mathcal{Z}^{N+1} \times \Omega \to \mathcal{S}^{N+1}$ is permutation equivariant in joint distribution, then $A(Z; \omega)$ is exchangeable.*

Proof. Consider a permutation $\pi \in S_{N+1}$. From exchangeability of Z, and the permutation equivariance of A, we have that

$$A(Z) \overset{\text{d}}{=} A\left(\left[Z_{\pi(i)}\right]_{i=1}^{N+1}\right) \overset{\text{d}}{=} \left[(A(Z))_{\pi(i)}\right]_{i=1}^{N+1},$$

meaning that it is exchangeable. □

Therefore, we get the following corollary.

Corollary 2. *If $A : \mathcal{Z}^{N+1} \times \Omega \to \mathcal{S}^{N+1}$ is permutation equivariant in joint distribution, and $S = A(Z; \omega)$, then, by defining*

$$P := \frac{1}{N+1}\left[1 + \sum_{i=1}^{N} \mathbf{1}_{S_{N+1} \succcurlyeq S_i}\right],$$

we will have that P is a valid p-value.

Proof. From the lemma 2, we have that S is exchangeable. Note that the construction of P is the same (but use the last index instead of first index) as P_1 in the lemma 1, so by the corollary 1, P is a valid p-value. □

B.2. Construction & sufficient condition

In the last subsection, we have established that permutation equivariance will be sufficient for the validity of the rank p-value. Note that this is more general than a conformal p-value, since A does not have to create non-conformity scores like in the algorithm 2. In this subsection, we will restrict the attention when the algorithm A generates non-conformity scores under full conformal prediction framework. For any vector $u = (u_1, \ldots, u_{N+1})$, write $u_{-i} := (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_{N+1})$. This is when

$$A(z; \omega) := [T_i(z_{-i}, z_i; \omega)]_{i=1}^{N+1}$$

[15] Note that the distribution of Z is fixed to be an exchangeable one, so we leave it out of the lemma statement.

for any $z \in \mathcal{Z}^{N+1}$.

Consider a permutation $\pi \in S_{N+1}$ and write $z^\pi := \left[z_{\pi(i)}\right]_{i=1}^{N+1}$. For each coordinate i, we have

$$\begin{aligned}(A(z^\pi;\omega))_i &= T_i\left((z^\pi)_{-i}, z_{\pi(i)};\omega\right),\\ (A(z;\omega))_{\pi(i)} &= T_{\pi(i)}\left(z_{-\pi(i)}, z_{\pi(i)};\omega\right).\end{aligned}$$

Thus, for each coordinate, the two random vectors share the same last element, but with different ordering and index of the score function. If we want permutation equivariance in joint distribution, we then need the marginal distribution of each coordinate to be the same. This can be easily satisfied by enforcing permutation invariance in distribution 3. However, as we have shown in the section 3, this is not sufficient, since this will only give permutation equivariance in marginal distributions but not joint distribution. Thus, a stronger condition is required.

For example, if we have almost sure permutation invariance 2, we can rearrange the first N terms and change the index from T_i to $T_{\pi(i)}$ for each coordinate. This gives $A\left(\left[z_{\pi(i)}\right]_{i=1}^{N+1};\omega\right) = \left[(A(z;\omega))_{\pi(i)}\right]_{i=1}^{N+1}$ almost surely, which is stronger than what required in permutation equivariance in joint distribution.

Next, we will show that conditional independence and permutation invariance in distribution condition of the stochastic scoring functions T ensures that the stochastic algorithm A is permutation equivariant in joint distribution.

Lemma 3 (Conditional Independence & Permutation Invariance in Distribution implies Permutation Equivariance)**.** *If the stochastic non-conformity score functions T are conditionally independent and symmetric, then the stochastic algorithm $A(z;\omega) := [T_i(z_{-i}, z_i;\omega)]_{i=1}^{N+1}$ almost surely for all $z \in \mathcal{Z}^{N+1}$ is permutation equivariant in joint distribution.*

Proof. From Dynkin's $\pi-\lambda$ theorem, it is sufficient to show the equality in probability over rectangular set. Consider an arbitrary $B := \bigotimes_{i=1}^{N+1} B_i$, where each B_i is a measurable subset of $\mathcal{S}$. Write $z^\pi := \left[z_{\pi(i)}\right]_{i=1}^{N+1}$. From the definition 1 and the definition of A, we have that the collection

$$\{(A(z;\cdot))_i\}_{i=1}^{N+1}$$

is independent conditioned on C, so

$$\begin{aligned}\mathbb{P}(A(z^\pi) \in B|C) &= \mathbb{P}((A(z^\pi))_i \in B_i; \forall i|C)\\ &= \prod_{i=1}^{N+1} \mathbb{P}((A(z^\pi))_i \in B_i|C).\end{aligned}$$

Note that, from the conditional symmetry in 4,

$$\begin{aligned}((A(z^\pi))_i|C) &= \left(T_i\left((z^\pi)_{-i}, z_{\pi(i)}\right)\middle|C\right)\\ &\overset{\mathrm{d}}{=} \left(T_{\pi(i)}\left(z_{-\pi(i)}, z_{\pi(i)}\right)\middle|C\right)\\ &= \left((A(z))_{\pi(i)}\middle|C\right),\end{aligned}$$

so

$$\begin{aligned}\mathbb{P}(A(z^\pi) \in B|C) &= \prod_{i=1}^{N+1} \mathbb{P}\left((A(z))_{\pi(i)} \in B_i\middle|C\right)\\ &= \mathbb{P}\left((A(z))_{\pi(i)} \in B_i; \forall i\middle|C\right)\\ &= \mathbb{P}\left(\left[(A(z))_{\pi(i)}\right]_{i=1}^{N+1} \in B\middle|C\right).\end{aligned}$$

Note that all equalities here are held almost surely C. We then take expectation, so

$$\mathbb{P}(A(z^\pi) \in B) = \mathbb{P}\left(\left[(A(z))_{\pi(i)}\right]_{i=1}^{N+1} \in B\right).$$

Thus, A is permutation equivariant in joint distribution. □

C. Proof of the main theorem

Now that we have proved other claims, we can put them together to prove the main theorem.

Proof. From the lemma 3, we have that the joint score function is permutation invariant in joint distribution. From the corollary 2, we then have that P is a valid p-value. □

D. Strict relationships

As mentioned in the figure 1, the relationships

$$\text{DT} \subseteq \text{AS} \cap \text{I\&D},$$

and

$$\text{AS} \cup \text{I\&D} \subseteq \text{CI\&D} \subseteq \text{D}$$

can be strengthened from $\subseteq$ to $\subsetneq$ for some appropriate setting $((\Omega, \mathcal{F}, \mathbb{P}_{\text{alg}}), \mathcal{Z}, \mathcal{S})$.

Consider the case when $N = 3$, $\Omega = [0,1]^2$, $\mathcal{F}$ is the Borel σ-algebra, $\mathbb{P}_{\text{alg}}$ is the Lebesgue measure, $\mathcal{Z}$ is $\{0,1\} \times [0,1]$, and $\mathcal{S} = \mathbb{R}$.

D.1. CI&D $\subsetneq$ D

Since the random space is sufficiently rich to define 2 independent Brownian motion, the same construction in the section 3[16] will provide the example of T that is permutation

[16]The construction in the section 3 assumes $\mathcal{X} = [0,1]$, but $\mathcal{X}$ is not used in the definition, so the result will be the same.

invariance in distribution but does not yield a valid conformal p-value. From the theorem 1, we can then conclude that T is not conditionally independence and permutation invariance in distribution. Therefore,

$$\text{CI\&D} \subsetneq \text{D}.$$

D.2. DT $\subsetneq$ AS $\cap$ I&D

We define

$$T_i(z;\omega) := \begin{cases} \omega_1 & \text{if } x_1 = x_2 = 0 \text{ and } x_3 = 1 \\ \omega_2 & \text{if } x_1 = x_2 = 1 \text{ and } x_3 = 0 \\ 0 & \text{otherwise} \end{cases}$$

for all $i \in \{1, 2, 3\}$, $z \in \mathcal{Z}^3$, and $\omega \in \Omega$.

Thus, T is not deterministic, but it is almost surely permutation invariant.[17]

Next, we will show that this is independent and permutation invariant in distribution. As we have discussed, permutation invariant in distribution is weaker than the almost sure condition, so it is automatically satisfied. Since, among $T_1(z_2, z_3, z_1; \omega)$, $T_2(z_1, z_3, z_2; \omega)$, and $T_3(z_1, z_3, z_2; \omega)$, only at most one can be non-zero, we then get independence. Therefore,

$$\text{DT} \subsetneq \text{AS} \cap \text{I\&D}.$$

[17] Recall that we only require the permutation invariance over the first $N = 2$ arguments.