Conformal Validity Guarantees Exist for Any Data Distribution (and How to Find Them)

Drew Prinster, Samuel Stanton, Anqi Liu, Suchi Saria

Duke B&B

October 31, 2025

Presented by Yuqi Li

Uncertainty & Conformal Prediction

What is Conformal Prediction (CP)?

- CP constructs a **prediction set** $\widehat{\mathcal{C}}_n(x)$ that contains the true label Y with probability at least $1-\alpha$, *without assumptions on the model form*.
- Validity guarantee (finite-sample):

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

- Intuitively, CP ranks how "unusual" a test point (x, y) is compared to calibration samples, and includes y if it is not too extreme.

Conformal Prediction (CP) provides **distribution-free coverage guarantees** without assuming model correctness.

Uncertainty & Conformal Prediction

Motivation & Limitation of Existing Methods:

- As AI/ML systems increasingly make autonomous decisions, it becomes critical to **quantify uncertainty** and **control risk** of their predictions.
- Standard uncertainty quantification (UQ) often fails when the data distribution shifts — e.g., in *active learning* or *reinforcement learning*, where models influence future data.
- Prior CP variants rely on *exchangeability* or its relaxed form *weighted exchangeability*. These assumptions break under sequential or feedback-loop shifts common in AI agents.

Core contribution of this paper:

- Establishes that conformal prediction can, in theory, extend to **any joint data distribution** with a valid density f—exchangeability is not a required assumption.
- Introduces a **general procedure** for deriving **weighted CP algorithms** tailored to specific non-exchangeable distributions.

Simple case: univariate prediction

Problem. Suppose $Y_1, Y_2, \ldots$ are exchangeable samples from the same distribution. How can we estimate Y_{n+1} after seeing $Y_1, \ldots, Y_n$?

Goal. A prediction interval C_n such that

$$\Pr\{Y_{n+1} \in C_n\} \geq 1 - \alpha.$$

Intuition.

- All Y_i's are identically distributed: the distribution of Y_{n+1} is the same as the distribution of $Y_1, \ldots, Y_n$.
- $Y_1, \ldots, Y_n$ are exchangeable: quantiles of the distribution of Y_{n+1} can be estimated by the empirical quantiles of $Y_1, \ldots, Y_n$.

Solution. Construct $(1 - \alpha)$ prediction interval using quantiles of the observed samples

$$\widehat{C}_n = \left[\widehat{Q}_{\alpha/2}(Y_{1:n}),\ \widehat{Q}_{1-\alpha/2}(Y_{1:n}) \right].$$

Prediction with additional information

Problem. Suppose $(X_1, Y_1), (X_2, Y_2), \ldots$ are exchangeable samples from the same distribution. How can we estimate Y_{n+1} after seeing $(X_1, Y_1), \ldots, (X_n, Y_n)$, for a newly drawn X_{n+1} from the same distribution?

Goal. A prediction interval C_n such that

$$\Pr\{Y_{n+1} \in C_n\} \geq 1 - \alpha.$$

Note that our goal is not $\Pr\{Y_{n+1} \in C_n \mid X_{n+1}\} \geq 1 - \alpha$*, which requires knowledge on relation between* X *and* Y*, thus not distribution-free.*

Naive approach. Can we simply ignore X?
Yes, because (X_i, Y_i)'s are exchangeable and identically distributed implies that Y_i's are exchangeable and identically distributed.

Prediction with additional information

- For any function $f(x, y)$, (X_i, Y_i)'s are exchangeable and identically distributed would imply that $S_i := f(X_i, Y_i)$'s are exchangeable and identically distributed.
- Finding the prediction interval for Y_{n+1} reduces to finding the prediction interval for S_{n+1}!
- Prediction interval for S_{n+1}:

$$\widehat{C}_{n,S} = [\widehat{Q}_{\alpha/2}(S_{1:n}), \widehat{Q}_{1-\alpha/2}(S_{1:n})].$$

- Prediction interval for Y_{n+1}:

$$\widehat{C}_{n,Y} = \{y : f(X_{n+1}, y) \in \widehat{C}_{n,S}\}.$$

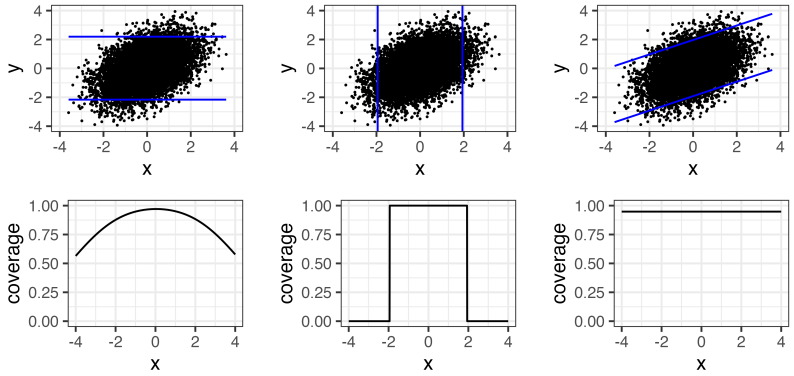
Why does this work?

$$\Pr\{Y_{n+1} \in \widehat{C}_{n,Y}\} = \Pr\{S_{n+1} = f(X_{n+1}, Y_{n+1}) \in \widehat{C}_{n,S}\} = 1 - \alpha.$$

Prediction with additional information

Example. (X_i, Y_i) are i.i.d. from a bivariate normal distribution: $X_i \sim \mathcal{N}(0,1)$, $Y_i = 0.5X_i + \mathcal{N}(0,1)$.

Choices of $f(x,y)$: • $f(x,y) = y$; • $f(x,y) = x$; • $f(x,y) = y - 0.5x$.



Prediction with additional information

- Any choice of $f(x, y)$ guarantees that $\Pr\{Y_{n+1} \in \widehat{C}_{n,Y}\} \geq 1 - \alpha$: ensures coverage across the whole test set
- However, not all choices of $f(x, y)$ guarantee that $\Pr\{Y_{n+1} \in \widehat{C}_{n,Y} \mid X_{n+1}\} \geq 1 - \alpha$: does not ensure coverage for a particular test case

- When $f(X_i, Y_i)$ is independent of X_i, conditional coverage holds: $\Pr\{Y_{n+1} \in \widehat{C}_{n,Y} \mid X_{n+1}\} \geq 1 - \alpha$
- This suggests choosing $f(X_i, Y_i) = Y_i - \widehat{\mu}(X_i)$, where $\widehat{\mu}(\cdot)$ is a prediction model for Y given X.

The framework is **distribution-free**: marginal coverage holds for any choice of f, but a prediction model may improve conditional coverage.

Standard Conformal Prediction

Algorithmic form:

- Let $\widehat{S}(x, y) = \mathcal{S}((x, y), \bar{Z})$ denote the fitted score function, where $Z_i = (X_i, Y_i)$.
- Compute calibration scores $V_i = \widehat{S}(Z_i)$ for $i = 1, \ldots, n$.
- Let $Q_{1-\alpha}$ be the empirical $(1 - \alpha)$ quantile of $\{V_i\}$.
- Construct the prediction set:

$$\widehat{C}_n(x) = \{y \in \mathcal{Y} : \widehat{S}(x, y) \leq Q_{1-\alpha}(V_{1:n} \cup \{\infty\})\}.$$

Limitation: the assumption of exchangeability excludes many realistic cases, such as sequential or feedback-dependent data—precisely the problem addressed in this paper.

Another perspective on conformal prediction

Key idea: "Nonconformity" or "strangeness" of a sample.

- Define a real-valued **score function** $\mathcal{S}((x, y), \bar{Z})$ measuring how atypical (x, y) is relative to a bag of examples $\bar{Z}$, where $Z_i = (X_i, Y_i)$.
- Example: residual score $|\widehat{\mu}_{\bar{Z}}(x) - y|$ where $\widehat{\mu}_{\bar{Z}}$ is a regression predictor fitted on $\bar{Z}$.
- Compute scores for all points, compare the new sample's score to the calibration distribution of past scores.

Intuition: If the new sample's score is not among the largest α-fraction of calibration scores, its label y is "plausible" and included in $\widehat{C}_n(x)$.

Conformal Prediction with Distributional Shift

- In some applications, the data from the training set and the test set may be different in distribution.
- e.g. in-hospital patients used to train a model may be very different from office-visit patients whose outcomes are to be predicted

- Assume the training set and the test set still shares the same conditional distribution $P_{Y|X}(Y \mid X)$, but have different marginal distribution $P_X^{(0)}(X)$ and $P_X^{(1)}(X)$, respectively.
- Can we provide a distribution-free CI for Y_{n+1} given $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_X^{(0)} \times P_{Y|X}$, and $X_{n+1} \sim P_X^{(1)}$?

- **Goal:** achieve marginal coverage of $1 - \alpha$:

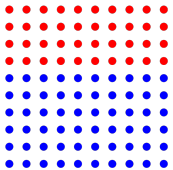
$$\Pr\{Y_{n+1} \in C_n\} \geq 1 - \alpha, \quad (X_{n+1}, Y_{n+1}) \sim P_X^{(1)} \times P_{Y|X}.$$

Conformal Prediction with Distributional Shift

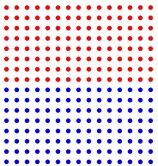
- Why does the previous approach fail? The distribution of X in the test set is different from that in the training set.
- **Intuition of solution:** Manipulate the training set so that it resembles the test set.

- Toy example:
  In training set, boys : girls $= 3 : 2$; in test set, boys : girls $= 1 : 1$.
- We can manipulate the training set by doubling the boys and tripling the girls, so that it looks like the test set.

60 boys, 40 girls



120 boys, 120 girls



Conformal Prediction with Distribution Shift

Generalizing the idea to non-binary predictors, we statistically "manipulate" the training set through **weighting**:

- The probability of drawing x in the training set is $P_X^{(0)}(x)$, but probability of drawing x in the test set is $P_X^{(1)}(x)$.
- Each time we draw x in the training set, we pretend that we have drawn $w(x) := P_X^{(1)}(x)/P_X^{(0)}(x)$ copies of x.
- Hence, the training set $(X_1, Y_1), \ldots, (X_n, Y_n)$ should be weighted by $w(X_1), \ldots, w(X_n)$, respectively.

Intuitively, one may think of weighting as each value (X_i, Y_i) in the training set has $w(X_i)$ copies.

Conformal Prediction with Distributional Shift

- Define $S_i = Y_i - \widehat{\mu}(X_i)$. Construct prediction interval for S_{n+1} using weighted quantile:

$$\widehat{C}_{n,S} = [\widehat{Q}_{\alpha/2}(S_{1:n}, w_{1:n}), \widehat{Q}_{1-\alpha/2}(S_{1:n}, w_{1:n})],$$

  where $\widehat{Q}_p(S, w)$ is the weighted empirical p-quantile of S with weights w.
- Construct prediction interval for Y_{n+1}: $\widehat{C}_{n,Y} = \widehat{\mu}(X_{n+1}) + \widehat{C}_{n,S}$.

General View of CP

- Challenge of active learning: training data $(X_1, Y_1), \ldots, (X_n, Y_n)$ are not exchangeable (and usually not identically distributed).
- Therefore, the distribution of (X_i, Y_i) cannot be simply denoted by $P_X^{(0)} \times P_{Y|X}$.
- Like previously did, we define $S_i = Y_i - \hat{\mu}(X_i)$, where $S_1, \ldots, S_n$ are from training set and S_{n+1} is from the test set.
  Subtlety here: Y_{n+1} is not observable, so S_{n+1} should actually be understood as a function of $Y_{n+1} = y$: $S_{n+1}(y) = y - \widehat{\mu}(X_{n+1})$, instead of a fixed observed quantity.

- **Goal:** derive a confidence interval for S_{n+1}, from which we can solve out the confidence interval for Y_{n+1}.

General View of CP

- **Question:** Given $S_1, \ldots, S_n$, can we construct a $(1-\alpha)$ CI for S_{n+1}, when these observations are not exchangeable?
- Generally not when distribution-free. We need to assume a distributional model $f(s_1, \ldots, s_{n+1})$ for the joint distribution of $(S_1, \ldots, S_{n+1})$.
- Similar to the case with distributional shift, we need to assign with each $S_i = s_i$ in the training set a weight w_i:

$$w_i = \frac{\Pr(s_i \text{ comes from the test set})}{\Pr(s_i \text{ comes from the training set})} = \frac{\Pr(S_{n+1} = s_i)}{\Pr(S_j = s_i, j \neq n+1)} \approx \Pr(S_{n+1} = s_i).$$

- How to calculate w_i?

$$w_i = \Pr(S_{n+1} = s_i) = \iiint \cdots \int f(u_1, u_2, \cdots, u_n, s_i)\, \mathrm{d}u_1 \mathrm{d}u_2 \cdots \mathrm{d}u_n.$$

Computationally hard!

General View of CP - computational trick

- Recall that for each $S_i = s_i$ in the training set,

$$w_i \propto \Pr(S_{n+1} = s_i).$$

- Therefore,

$$\begin{aligned}(w_1, \cdots, w_{n+1}) &\propto (\Pr(S_{n+1} = s_1), \cdots, \Pr(S_{n+1} = s_{n+1})) \\ &\propto (\Pr(S_{n+1} = s_1 \mid E), \cdots, \Pr(S_{n+1} = s_{n+1} \mid E)),\end{aligned}$$

  where E is the event that $(S_1, \ldots, S_{n+1})$ is a permutation of $(s_1, \ldots, s_{n+1})$.
- Calculation of conditional probability is much easier:

$$\Pr(S_{n+1} = s_i \mid E) = \frac{\sum_{\sigma:\sigma(n+1)=i} f(s_{\sigma(1)}, s_{\sigma(2)}, \cdots, s_{\sigma(n+1)})}{\sum_{\sigma} f(s_{\sigma(1)}, s_{\sigma(2)}, \cdots, s_{\sigma(n+1)})}.$$

reduced from integral to finite summation!

General View of CP

- With the calculated weights $w_1, \ldots, w_n$, we can now **define** $(1-\alpha)$ CI for S_{n+1}:

$$\widehat{C}_{n,S} = [\widehat{Q}_{\alpha/2}(S_{1:n}, w_{1:n}), \widehat{Q}_{1-\alpha/2}(S_{1:n}, w_{1:n})].$$

- Then we can construct $(1-\alpha)$ CI for Y_{n+1}:

$$\widehat{C}_{n,Y} = \{y : S_{n+1}(y) \in \widehat{C}_{n,S}\}.$$

- The reason that I used **define** instead of **construct** here is that $\widehat{C}_{n,S}$ depends on the unobserved value $Y_{n+1} = y$ as well.
- Thus, the CI $\widehat{C}_{n,Y}$ is not as straightforward to solve as it seems.

Comments

Core contributions of this paper:

- Establishes that conformal prediction can, in theory, extend to **any joint data distribution** with a valid density f—exchangeability is not a required assumption.
- Introduces a **general procedure** for deriving **weighted CP algorithms** tailored to specific non-exchangeable or dependent-data settings.

Limitations and challenges:

- The general formulation is **computationally intractable**, requiring $\mathcal{O}((n+1)!)$ evaluations of f.
- The true joint density f is typically **unknown in practice**, limiting the direct applicability of the theoretical framework.

Comment: This paper provides a unifying theoretical perspective on CP and its validity under any distribution, but for practical understanding of conformal prediction, it is recommended to first study **Tibshirani et al. (2019)** (*Conformal Prediction under Covariate Shift*) for intuition and implementation insights.

References

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer, Volume 29.

Tibshirani, R. J., Foygel Barber, R., Candès, E., and Ramdas, A. (2019). *Conformal Prediction under Covariate Shift*. *Advances in Neural Information Processing Systems*, 32.

Vovk, V., et al. (2024). *Conformal Validity Guarantees Exist for Any Data Distribution (and How to Find Them)*. arXiv preprint `arXiv:2402.01810`.

Appendix 1: Weighted Conformal Prediction

Weighted nonconformity scores.

- Compute calibration scores $V_i = \widehat{\mathcal{S}}(X_i, Y_i)$ and test score $V_{n+1} = \widehat{\mathcal{S}}(X_{n+1}, y)$.
- Assign weights proportional to the covariate density ratio:

$$w_i = \frac{p_X^{(1)}(X_i)}{p_X^{(0)}(X_i)}, \qquad w_{n+1} = \frac{p_X^{(1)}(X_{n+1})}{p_X^{(0)}(X_{n+1})}.$$

- Define normalized weights

$$p_i^w(x) = \frac{w_i}{\sum_{j=1}^n w_j + w_{n+1}(x)}, \quad p_{n+1}^w(x) = \frac{w_{n+1}(x)}{\sum_{j=1}^n w_j + w_{n+1}(x)}, \quad i = 1, \cdots, n$$

Appendix 1: Weighted Conformal Prediction

Prediction set definition (Tibshirani et al., 2019).

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

Where:

- $V_i^{(x,y)} = \widehat{\mathcal{S}}(X_i, Y_i)$ are calibration scores, and $V_{n+1}^{(x,y)} = \widehat{\mathcal{S}}(x, y)$ is the test score;
- $\delta_{V_i^{(x,y)}}$ denotes a Dirac measure (point mass) at score value $V_i^{(x,y)}$.

Coverage guarantee.

$$\Pr\left\{ Y_{n+1} \in \widehat{\mathcal{C}}_n(X_{n+1}) \right\} \geq 1-\alpha, \quad \text{if data are } \textit{weighted exchangeable}.$$

Intuitively, calibration scores are reweighted according to their importance under the test covariate distribution; this restores exchangeability in the weighted sense.