IMPERIAL

Conformal Prediction as Bayesian Quadrature

ICML 2025

Xinzhe Luo
July 31, 2025

Motivation

- Distribution-free uncertainty quantification: to flexibly and reliably quantify the suitability of a model for deployment without making too many assumptions about how the model was trained or in which setting it will be used.
- Conformal prediction is distribution-free, but are based on frequentist statistics, making it difficult to incorporate prior knowledge that might be available about specific models.

Distribution-Free Uncertainty Quantification
Split conformal prediction

- Assume a black-box model mapping inputs X to outputs Y, a calibration set $\{X_i, Y_i\}_{i=1}^{n}$, and a score function $s(x, y)$ which measures the disagreement between a predictor's output and the ground truth.
- Assuming exchangeability of $\{(X_i, Y_i)\}_{i=1}^{n+1}$, the α-acceptable conformal prediction set $\mathcal{C}(X_{n+1})$ of for a test point (X_{n+1}, Y_{n+1}) is

$$P(Y_{n+1} \notin \mathcal{C}(X_{n+1})) \leq \alpha, \quad \mathcal{C}(X_{n+1}) = \{y : s(X_{n+1}, y) \leq \widehat{q}\}, \tag{1.1}$$

where $\widehat{q}$ is the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$-th quantile of the scores on the calibration set.

Proof.

Note that $P(Y_{n+1} \notin \mathcal{C}(X_{n+1})) = P(s(X_{n+1}, Y_{n+1}) > \widehat{q})$, where $\widehat{q}$ is the $\lceil (n+1)(1-\alpha) \rceil$-th value of the sorted scores. Thus, due to the exchangeability of the samples, we have

$$P(s(X_{n+1}, Y_{n+1}) > \widehat{q}) = \frac{(n+1) - \lceil (n+1)(1-\alpha) \rceil}{n+1} \leq \frac{(n+1) - (n+1)(1-\alpha)}{n+1} = \alpha. \tag{1.2}$$

$\square$

- Conformal risk control considers prediction sets that minimise the expected risk of miscoverage,

$$\mathbb{E}\left[\ell\left(\mathcal{C}_\lambda(X_{n+1}), Y_{n+1}\right)\right] \leq \alpha, \tag{1.3}$$

for any bounded loss function $\ell \leq B < \infty$ that shrinks as $\mathcal{C}_\lambda(X_{n+1})$ grows, and larger λ leads to larger (more conservative) prediction sets.

Theorem (Conformal Risk Control)

Denote $L_i(\lambda) \triangleq \ell(\mathcal{C}_\lambda(X_i), Y_i)$ for $i = 1, \ldots, n+1$. Assume that $L_i(\lambda)$ is non-increasing in λ, right continuous, and $L_i(\lambda_{\max}) \leq \alpha$, $\sup_\lambda L_i(\lambda) \leq B < \infty$ a.s. Let $\widehat{R}_n(\lambda) = \frac{1}{n}\sum_{i=1}^n L_i(\lambda)$, then

$$\mathbb{E}\left[L_{n+1}(\widehat{\lambda})\right] \leq \alpha, \quad \widehat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\}. \tag{1.4}$$

Bayesian Quadrature

- Bayesian quadrature estimates the value of an integral $J[f] := \int_a^b f(x)\mathrm{d}x$ by
  1. place a prior $p(f)$ on functions;
  2. evaluate f at a finite set of points $\{x_i\}_{i=1}^n$;
  3. compute the posterior distribution $p(f \mid x_{1:n}, y_{1:n}) \propto p(f) \prod_{i=1}^n \delta(y_i - f(x_i))$;
  4. estimate $\int_a^b f(x)\mathrm{d}x \approx \int_a^b f_n(x)\mathrm{d}x$, where $f_n(x) = \mathbb{E}[f(x) \mid x_{1:n}, y_{1:n}]$.

Proof.

By Fubini's theorem, we have

$$\int_a^b f(x)\mathrm{d}x \approx \mathbb{E}\left[\int_a^b f(x)\mathrm{d}x \mid x_{1:n}, y_{1:n}\right] = \int_a^b \mathbb{E}[f(x) \mid x_{1:n}, y_{1:n}]\mathrm{d}x = \int_a^b f_n(x)\mathrm{d}x. \tag{1.5}$$

$\square$

Decision-Theoretic Formulation

- Let $z = (z_1, \ldots, z_n)$ be a set of calibration data where each $z_i = (x_i, y_i)$ is a pair of input and ground truth.
- Let θ denote the true state of nature that defines a shared density $f(z_i \mid \theta)$ for the data. A new test point z_{new} is assumed to have the same distribution.
- Let $\lambda(z)$ be a control parameter that must be chosen based on the calibration data. We assume the presence of a loss function $L(\theta, \lambda)$ which quantifies the loss incurred by selecting λ when the true state of nature is θ.
- The decision-theoretic goal is to choose a decision rule $\lambda(z)$ that controls the *risk*, defined as the expected loss:

$$R(\theta, \lambda) = \mathbb{E}_{z \sim f(z|\theta)}[L(\theta, \lambda(z))]. \tag{1.6}$$

- The *maximum risk* is defined as $\overline{R}(\lambda) = \sup_\theta R(\theta, \lambda)$. We want to find *$\alpha$-acceptable decision rules* whose risk is upper bounded by a constant α:

$$\overline{R}(\lambda) \leq \alpha, \tag{1.7}$$

and use another criterion to select among these.

Decision-Theoretic Formulation

Recovering Split Conformal Prediction

Proposition (3.1)

Let $L_{scp}(\theta, \lambda) \triangleq P(s(z_{new}) > \lambda) = 1 - \int \mathbf{1}(s(z_{new}) \leq \lambda) f(z_{new} \mid \theta) dz_{new}$ *be the miscoverage loss, where* s *is an arbitrary nonconformity function. Define* $s_i \triangleq s(z_i)$ *for* $i = 1, \ldots, n$ *and let* $s_{(1)} \leq s_{(2)} \leq \cdots \leq s_{(n)}$ *be the corresponding order statistics. Let* λ_{scp} *be the following decision rule:*

$$\lambda_{scp} = \widehat{q}_{1-\alpha} := \begin{cases} s_{(\lceil (n+1)(1-\alpha) \rceil)}, & \textit{if } \lceil (n+1)(1-\alpha) \rceil \leq n, \\ \infty, & \textit{otherwise.} \end{cases} \tag{1.8}$$

Then λ_{scp} *is an* α*-acceptable decision rule for the miscoverage loss* L_{scp}*.*

Proof.

By exchangeability, $P(s_{\text{new}} \leq \widehat{q}_{1-\alpha}) \geq 1 - \alpha$. Therefore, for $\lambda = \widehat{q}_{1-\alpha}$, $R(\theta, \lambda) = P(s_{\text{new}} > \lambda) \leq \alpha$. This statement holds for any θ, so we have $\overline{R}(\lambda_{\text{scp}}) \leq \alpha$ and $\lambda_{\text{scp}} = \widehat{q}_{1-\alpha}$. □

Recovering Conformal Risk Control

Proposition (3.2)

Let $L_{crc}(\theta, \lambda) \triangleq \int \ell(z_{new}, \lambda) f(z_{new} \mid \theta) \mathrm{d}z_{new}$ *where* $\ell(z_{new}, \lambda)$ *is an individual loss function that is monotonically non-increasing in* λ*. Let* λ_{crc} *be the following decision rule:*

$$\lambda_{crc} = \inf\left\{\lambda : \frac{1}{n+1}\left(\sum_{i=1}^{n} \ell(z_i, \lambda) + B\right) \leq \alpha\right\}. \tag{1.9}$$

Then λ_{crc} *is an* α*-acceptable decision rule for the loss* L_{crc}*.*

Proof.

By definition, we identify $L_i(\lambda) = \ell(z_i, \lambda)$ for $i = 1, \ldots, n$ and $L_{n+1}(\lambda) = \ell(z_{\mathrm{new}}, \lambda)$, $\lambda_{\mathrm{crc}} = \widehat{\lambda}$, and $R(\theta, \lambda_{\mathrm{crc}}) = \mathbb{E}[L_{n+1}(\widehat{\lambda})] \leq \alpha$ for any θ. Thus, we have $\overline{R}(\lambda_{\mathrm{crc}}) \leq \alpha$. $\square$

Bayes Risk

- Since the true state of nature θ is uncertain, we want a decision rule that protects against high loss for a range of possible θ. The idea is expressed as the *integrated risk*:

$$r(\pi, \lambda) = \int R(\theta, \lambda)\pi(\theta)\mathrm{d}\theta, \tag{2.1}$$

  where $\pi(\theta)$ is a prior distribution over the true state of nature θ.

- The *Bayes decision rule* is defined as the minimiser of the posterior risk:

$$\lambda^{\pi} \triangleq \arg\min_{\lambda} r(\lambda \mid z), \tag{2.2}$$

  where $r(\lambda \mid z)$ is the *posterior risk*

$$r(\lambda \mid z) = \int L(\theta, \lambda)\pi(\theta \mid z)\mathrm{d}\theta. \tag{2.3}$$

Reformulation as Bayesian Quadrature

- Consider the posterior risk $r(\lambda \mid z) = \int L(\theta, \lambda)\pi(\theta \mid z)\mathrm{d}\theta$ where $L(\theta, \lambda) = \int \ell(z_{\text{new}}, \lambda) f(z_{\text{new}} \mid \theta)\mathrm{d}z_{\text{new}}$.
- Define the distribution function of individual losses induced by λ for a particular value of θ: $F(\ell) \triangleq P(\ell(z_{\text{new}}, \lambda) \leq \ell \mid \theta)$. The corresponding quantile function is defined as $K(t) := F^{-1}(t) \triangleq \inf\{\ell : F(\ell) \geq t\}$. Since the expectation of an r.v. is equal to the integral of its quantile function, we have $L(\theta, \lambda) = \int_0^1 K(t)\mathrm{d}t =: J[K]$.
- The posterior risk given the observed individual losses $\ell_i \triangleq \ell(z_i, \lambda)$ for $i = 1, \ldots, n$ can be expressed as

$$r(\lambda \mid \ell_{1:n}) = \mathbb{E}_\theta \left[L(\theta, \lambda) \mid \ell_{1:n}\right] = \mathbb{E}_K \left[J[K] \mid \ell_{1:n}\right] = \int J[K] p(K \mid \ell_{1:n})\mathrm{d}K. \tag{2.4}$$

- The posterior over quantile functions can be expressed as

$$p(K \mid \ell_{1:n}) = \int p(K \mid t_{1:n}, \ell_{1:n}) p(t_{1:n} \mid \ell_{1:n})\mathrm{d}t_{1:n}, \quad p(K \mid t_{1:n}, \ell_{1:n}) \propto \pi(K) \prod_{i=1}^{n} \delta\left(\ell_i - K(t_i)\right). \tag{2.5}$$

Elimination of the Prior Distribution

Theorem (4.1)

Let $t_{(0)} = 0$*,* $t_{(n+1)} = 1$*, and* $\ell_{(n+1)} = B$*. Then, given the evaluation sites* $t_{1:n}$*,*

$$\sup_{\pi} \mathbb{E}[L \mid t_{1:n}, \ell_{1:n}] \leq \sum_{i=1}^{n+1} u_i \ell_{(i)}, \tag{2.6}$$

where $u_i = t_{(i)} - t_{(i-1)}$*.*

Proof.

By Lemma B.2, we have

$$\mathbb{E}[L \mid t_{1:n}, \ell_{1:n}] = \int J[K] p(K \mid t_{1:n}, \ell_{1:n}) \mathrm{d}K \leq \sup_{K \in \mathcal{K}_n} J[K] \leq \sum_{i=1}^{n+1} (t_{(i)} - t_{(i-1)}) \ell_{(i)}. \tag{2.7}$$

□

Elimination of the Prior Distribution

Lemma (B.1)

Consider the variational maximisation problem $I[f] = \int_a^b f(x)\mathrm{d}x$ *subject to* $f(a) = f_a$, $f(b) = f_b$, *and* $f_a \leq f(x) \leq f_b$ *for all* $x \in [a, b]$ *where* $f_a \leq f_b$. *Then, the solution is given by*

$$f^*(x) = \begin{cases} f_a & \text{if } x = a, \\ f_b & \text{otherwise,} \end{cases} \tag{2.8}$$

and $I[f^*] = (b - a)f_b$.

Lemma (B.2)

Let $\mathcal{K}_n$ *be the set of quantile functions for which* $K(t_i) = \ell_i$ *for* $i = 1, \dots, n$. *Then,*

$$\sup_{K \in \mathcal{K}_n} J[K] = \sum_{i=1}^{n+1} (t_{(i)} - t_{(i-1)})\ell_{(i)}, \tag{2.9}$$

where $t_{(0)} = 0$, $t_{(n+1)} = 1$, $\ell_{(n+1)} = B$, *and* $J[K] = \int_0^1 K(t)\mathrm{d}t$.

Bound on Maximum Posterior Risk

Theorem (4.3)

Define $\ell_{(i)}$ *to be the order statistics of* $\ell_{1:n}$, *and* $\ell_{(n+1)} \triangleq B$. *Let* L^+ *be the r.v. defined as follows:*

$$L^+ = \sum_{i=1}^{n+1} U_i \ell_{(i)}, \quad U_1, \ldots, U_{n+1} \sim Dir(1, \ldots, 1). \tag{2.10}$$

Then, for any $b \in (-\infty, B]$,

$$\inf_\pi P(L \leq b \mid \ell_{1:n}) \geq P(L^+ \leq b). \tag{2.11}$$

Corollary (4.4)

For any desired confidence level $\beta \in (0, 1)$, *define*

$$b^*_\beta = \inf_b \{b : P(L^+ \leq b \mid \ell_{1:n}) \geq \beta\}. \tag{2.12}$$

Then, $\inf_\pi P(L \leq b \mid \ell_{1:n}) \geq \beta$ *for any* $b \geq b^*_\beta$.

Recovering Conformal Methods

- Taking the expected value of L^+, we have,

$$\mathbb{E}[L^+] = \sum_{i=1}^{n+1} \mathbb{E}[U_i]\ell_{(i)} = \frac{1}{n+1}\left[\sum_{i=1}^{n+1} \ell_i + B\right]. \tag{2.13}$$

  The *Conformal Risk Control* decision rule is the infimum over λ for which $\mathbb{E}[L^+] \leq \alpha$.

- For *Split Conformal Prediction*, the individual loss is defined as $\ell_i = 1 - \mathbf{1}(s_i \leq \lambda)$. Let $\lambda = s_{(k)}$. The expected value of L^+ becomes

$$\mathbb{E}[L^+] = \frac{1}{n+1}\left[n + 1 - \sum_{i=1}^{n} \mathbf{1}(s_i \leq s_{(k)})\right] = 1 - \frac{k}{n+1}. \tag{2.14}$$

  Therefore, $\mathbb{E}[L^+] \leq \alpha$ is satisfied when $k \geq (n+1)(1-\alpha)$, and in particular $k^* = \lceil (n+1)(1-\alpha) \rceil$.

Conformal Prediction as Bayesian Quadrature

We use the Bayesian quadrature-based method to compute the decision rule based on the one-sided highest posterior density (HPD) interval

$$\lambda_{\mathrm{hpd}}^{\beta} \triangleq \inf_{\lambda} \{ \lambda : P(L^{+} \leq \alpha \mid \ell_{1:n}) \geq \beta \} \tag{2.15}$$

by finding the corresponding critical values b_{β}^{*} via Monte Carlo simulation of Dirichlet random variables with 1000 samples.

Related Work: Risk-Controlling Prediction Sets

Definition (Risk-Controlling Prediction Set, Bates et al., J. ACM 2021)

Let $\mathcal{T}$ be a random function taking values in the space of functions $\mathcal{X} \to \mathcal{Y}'$, where $\mathcal{Y}'$ is some space of sets. We say that $\mathcal{T}$ is a *(α, δ)-risk-controlling prediction set (RCPS)* if, with probability at least $1 - \delta$, we have $R(\mathcal{T}) \leq \alpha$, where $R(\mathcal{T}) = \mathbb{E}[L(Y, \mathcal{T}(X))]$ for some loss function $L : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}$.

- Assume that $\mathcal{T}$ is indexed by a parameter $\lambda \in [-\infty, \infty]$, and we have access to a pointwise *upper confidence bound (UCB)* $\widehat{R}^+$ for the risk R for each λ:

$$P\left(R(\lambda) \leq \widehat{R}^+(\lambda)\right) \geq 1 - \delta, \tag{2.16}$$

where $\widehat{R}^+(\lambda)$ may depend on the calibration set $\{(X_i, Y_i)\}_{i=1}^n$.

Related Work: Risk-Controlling Prediction Sets

Theorem (Validity of UCB Calibration)

Suppose $R(\lambda)$ is a continuous monotone nonincreasing function such that $R(\lambda) \leq \alpha$ for some $\lambda \in \Lambda$. Then, for $\widehat{\lambda} \triangleq \inf\{\lambda \in \Lambda : \widehat{R}^{+}(\lambda') < \alpha, \ \forall \lambda' \geq \lambda\}$,

$$P\left(R(\widehat{\lambda}) \leq \alpha\right) \geq 1 - \delta. \tag{2.17}$$

That is, $\mathcal{T}_{\widehat{\lambda}}$ is a (α, δ)-RCPS.

Proof.

Define $\lambda^* \triangleq \inf\{\lambda \in \Lambda : R(\lambda) \leq \alpha\}$. By continuity, $R(\lambda^*) = \alpha$. Then,

$$R(\widehat{\lambda}) > \alpha \Rightarrow \widehat{\lambda} < \lambda^* \Rightarrow \widehat{R}^{+}(\lambda^*) < \alpha = R(\lambda^*). \tag{2.18}$$

Thus, $P\left(R(\widehat{\lambda}) > \alpha\right) \leq P\left(\widehat{R}^{+}(\lambda^*) < R(\lambda^*)\right) \leq \delta$. □

Related Work: Risk-Controlling Prediction Sets

Example (Smplified Hoeffding Bound)

It is natural to construct a UCB for $R(\lambda)$ based on the empirical risk $\widehat{R}(\lambda) \triangleq \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \mathcal{T}_\lambda(X_i))$. Suppose the loss is bounded above by 1, then

$$P\left(\widehat{R}(\lambda) - R(\lambda) \leq -x\right) \leq \exp\left(-2nx^2\right). \tag{2.19}$$

This implies a UCB of the form

$$\widehat{R}^{+}_{\mathrm{sHoef}}(\lambda) = \widehat{R}(\lambda) + \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}, \tag{2.20}$$

and an RCPS $\mathcal{T}_{\widehat{\lambda}_{\mathrm{sHoef}}}$ with

$$\widehat{\lambda}_{\mathrm{sHoef}} = \inf\left\{\lambda \in \Lambda : \widehat{R}^{+}_{\mathrm{sHoef}}(\lambda') < \alpha,\ \forall\, \lambda' \geq \lambda\right\} = \inf\left\{\lambda \in \Lambda : \widehat{R}(\lambda) < \alpha - \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}\right\} \tag{2.21}$$

Experiment: Synthetic Binomial Data

- Assume the acceptance rate is $\alpha = 0.4$ and the loss function is

$$\ell(z_i, \lambda) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}(V_{ik} > \lambda), \tag{3.1}$$

where $V_{ik} \sim \text{Unif}(0,1)$ for $i = 1, \ldots, n$.

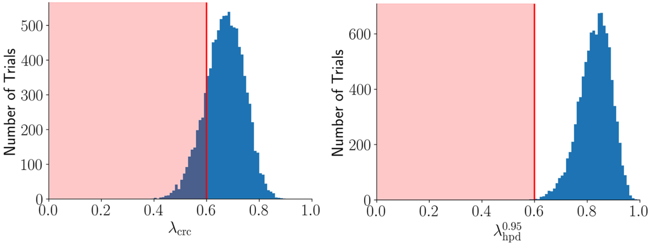

Figure 3. Comparison of risk incurred by each procedure across multiple trials. Left: Histogram of the decision rule λ_{crc} chosen by Conformal Risk Control across M = 10,000 randomly sampled calibration sets. The region where per-trial risk exceeds α is highlighted in red. Right: Histogram of the $\lambda_{\text{hpd}}^{0.95}$ chosen according to our 95% Bayesian posterior interval.

Experiment: Synthetic Heteroskedastic Data

- Let $X \sim U[0,4]$ and $Y \sim N(0, X^2)$.
- The prediction intervals are defined as $[-\lambda, \lambda]$. The loss is the miscoverage loss and the target loss is set to $\alpha = 0.1$.
- The maximum acceptable failure rate is set to $1 - \beta = 0.05$.

Table 2. Relative frequency of trials (out of 10,000) for which the resulting decision rule λ exceeded the target risk threshold α in the synthetic heteroskedastic experiment.

Decision Rule	Relative Freq.	95% CI	Mean Prediction Interval Length
Split Conformal Prediction / CRC	46.19%	[45.21%, 47.17%]	7.99
RCPS	0.0%	[0.0%, 0.04%]	14.29
Ours ($\beta = 0.95$)	3.42%	[3.07%, 3.80%]	9.50

Note: Error bars are computed as 95% Clopper-Pearson confidence intervals for binomial proportions.

IMPERIAL

Thank you!
Q&A