Conformal Prediction as Bayesian Quadrature

by Snell & Griffiths, ICML 2025 Outstanding Paper

Presenter: Yichen Ji

November 11, 2025

Overview

1. Background Check
   - Conformal Prediction
     - Split & Full Conformal Prediction
     - Statistical Decision Theory
     - Conformal Risk Control
   - Bayesian Quadrature

2. Paper's Approach
   - Motivation: Conditional Coverage, More than Marginal Coverage
   - Reformulate CP as BQ
   - Remove Prior Specification by an Upper Bound
   - Dirichlet Quantile Spacings & Bound Maximum Risk
   - Recover Conformal Methods by Posterior Mean

Conformal Prediction 101

- Goal: Distribution-free uncertainty quantification in future observations of ANY black-box prediction models/algorithms
- How? Construct prediction sets that contain the ground-truth output with high probability
  i.e. finite-sample coverage guarantees in terms of coverage level $1 - \alpha$
- How? Build a wrapper on top of black-box algos by converting prediction values into prediction sets.
- Setup: Assume having access to a calibration set $(X_i, Y_i), i = 1, \ldots, n$ the model hasn't seen, then given a new test point (X_{n+1}, Y_{n+1}), construct a prediction set $\hat{\mathcal{C}}_n$ such that

$$\mathbb{P}\left(Y_{n+1} \in \hat{\mathcal{C}}_n\left(X_{n+1}\right)\right) \geq 1 - \alpha \Leftrightarrow \mathbb{P}\left(Y_{n+1} \notin \hat{\mathcal{C}}_n\left(X_{n+1}\right)\right) \leq \alpha$$

Key idea: use rank to form adjusted quantiles

- Denote conformity score $s(x, y)$ as a (dis)agreement metric, e.g., absolute residual in regression, then the prediction set is formed by

$$\hat{\mathcal{C}}_n(X_{n+1}) = \{y : s(X_{n+1}, y) \leq \hat{q}_n\}$$

- $\hat{q}_n$ is the $\frac{\lceil (1-\alpha)(n+1) \rceil}{n}$ quantile of $s(X_1, Y_1), \ldots, s(X_n, Y_n)$, that is, $\lceil (1-\alpha)(n+1) \rceil$ smallest of $s(X_1, Y_1), \ldots, s(X_n, Y_n))$
- Key assumption: exchangeability of scores $s(X_i, Y_i), i = 1, \ldots, n+1$
- Then, the rank of s_{n+1} is uniformly distributed over $\{1, \ldots, n+1\}$

$$\underbrace{\overset{\bullet}{s_{(1)}} \quad \overset{\bullet}{s_{(2)}} \quad \overset{\bullet}{s_{(3)}} \quad \overset{\bullet}{s_{(4)}} \quad \cdots \quad \overset{\bullet}{s_{\lceil (1-\alpha)(n+1) \rceil}}}_{\geq 1-\alpha \text{ fraction}} \quad \cdots \quad \overset{\bullet}{s_{(n+1)}}$$

- $\Longrightarrow \mathbb{P}(s(X_{n+1}, Y_{n+1}) \leq \hat{q}_n) \in \left[1-\alpha, 1-\alpha+\frac{1}{n+1}\right)$

Regression Example

A Taste of Conformal Prediction by Emmanuel Candès starting at 10:12

Split & Full Conformal Prediction

- Statistical & Computational tradeoff: how to exploit accessible data $\{X_i, Y_i\}_{i=1}^n$ at hand?
- Split CP:
  - split data into training set D_1 & calibration set D_2
    $D_1 \cup D_2 = \{1, \dots, n\}, D_1 \cap D_2 = \emptyset$
  - fit prediction model $\hat{f}_{D_1}$ using D_1, compute scores using D_2
  - train model once (fast), but lose statistical efficiency due to the sample splitting (only use half the data points to fit the model)
- Full CP:
  - use all data points for training and calibration via leave-one-out fitting
  - fit model $\hat{f}_{-i}$ using all data except (X_i, Y_i), and compute score s_i for each $i = 1, \dots, n$
  - train model n times (expensive), but get exact coverage with tighter prediction sets
- Exchangeability needed for both! (Counterexamples: time series, covariate shift, heteroscedastic data $\Rightarrow$ modified CP methods)

Concrete Algorithm (Full CP, split as a special case)

1. Propose a test query value y
2. Pick any conformity score $s(x, y)$ (residual score $s(x, y) = |y - \hat{\mu}(x)|$)
3. Fit model to all $n+1$ data via a symmetric algorithm to $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{n+1}, y) \sim S(X_i, Y_i)$
4. Compute quantile $\hat{q}_n \triangleq Q_{\frac{\lceil (1-\alpha)(n+1) \rceil}{n}} (S(X_i, Y_i))$
   note: $Q_{1-\alpha} \{s_i : i \in \mathcal{I}_{\text{cal}}\}$ for split CP
5. If $S(X_{n+1}, y) \leq \hat{q}_n$, include y in prediction set $\hat{\mathcal{C}}_n$

Review of Statistical Decision Theory

Let $z = (z_1, \ldots, z_n)$ be a set of calibration data $\{(x_i, y_i)\}_{i=1}^n$. Denote θ as the likelihood parameter and $\lambda(z)$ as a control parameter chosen based on z. Loss $L(\theta, \lambda(z))$ is incurred by selecting λ when the true nature is θ.

- risk

$$R(\theta, \lambda) = \int L(\theta, \lambda(z)) f(z \mid \theta) dz$$

- maximum risk

$$\bar{R}(\lambda) = \sup_{\theta} R(\theta, \lambda)$$

- average risk

$$r(\pi, \lambda) = \int R(\theta, \lambda) \pi(\theta) d\theta$$

- posterior risk

$$r(\lambda \mid z) = E\left(L_\lambda \mid z\right) = \int L(\theta, \lambda(z)) \pi(\theta \mid z) d\theta$$

- Find λ to control $\bar{r}(\lambda) \triangleq \sup_\pi r(\pi, \lambda) = \sup_\theta R(\theta, \lambda) = \bar{R}(\lambda) \leq \alpha$

Conformal Risk Control (Angelopoulos et al, 2024)

- Generalize to per-sample loss functions $\ell(z, \lambda) \equiv \ell\left(\mathcal{C}_\lambda\left(x\right), y\right)$ that are assumed to be monotonic of a single parameter λ. The risk is

$$R(\theta, \lambda) = \int \ell\left(z_{\text{new}}, \lambda\right) f\left(z_{\text{new}} \mid \theta\right) dz_{\text{new}} = \mathbb{E}_{f(z|\theta)}\left[\ell\left(\mathcal{C}_\lambda\left(X_{n+1}\right), Y_{n+1}\right)\right]$$

- Goal: control expected loss (risk) under minimal assumptions (exchangeability), e.g., false positive rate, F1 score, conditional error

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

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

where B is assumed to be the maximum possible loss value.
$\ell(z, \lambda)$ measures how well the conformal set $\mathcal{C}(x_i)$ covers true label y_i,
$\frac{1}{n+1}\sum_{i=1}^{n} \ell(z_i, \lambda)$ empirical risk, $\frac{B}{n+1}$ correction term by exchangeability.

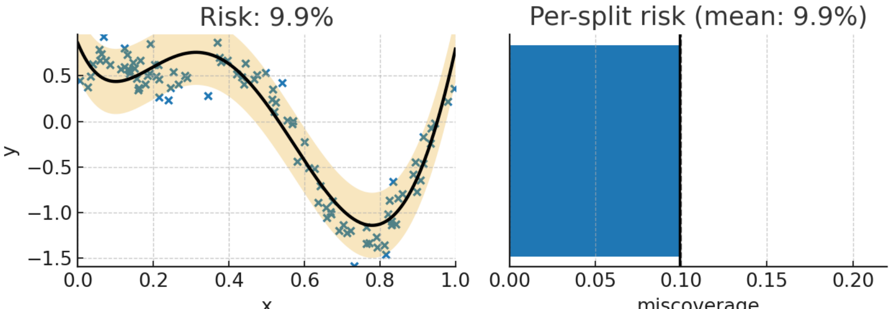
Bayesian Quadrature

A class of probabilistic numerical methods viewing numerical integration of $\int_a^b f(x)dx$ as Bayesian inference in the following steps:

1. Place a prior $p(f)$ on functions ,e.g., Gaussian process;
2. Evaluate f at $x_1, \ldots, x_n$ $y_i = f(x_i)$;
3. Compute posterior given the observed values of $p(f \mid x_{1:n}, y_{1:n}) \propto p(f) \prod_{i=1}^{n} \delta(y_i - f(x_i))$;
4. Estimate

$$\int_a^b f(x)dx \approx \int_a^b f_n(x)dx, f_n(t) = \mathbb{E}\left[f(t) \mid x_{1:n}, y_{1:n}\right]$$

- Marginal coverage:
  $\mathbb{P}\left(Y_{n+1} \in \hat{C}_n\left(X_{n+1}\right) \mid \left(X_i, Y_i\right), i \in D_1\right) \in \left[1-\alpha, 1-\alpha+\frac{1}{n_2+1}\right)$
  The probability or expectation $\mathbb{P}, \mathbb{E}$ is over both calibration and test data $i = 1, \ldots, n_2+1 \Rightarrow$ coverage guaranteed in aggregate over multiple calibration sets (model trained, test data fixed), but the guarantee might not hold for each set in particular
- Conditional coverage on each calibration set?
  $\mathbb{P}\left(Y_{n+1} \in \hat{C}_n(x) \mid \left(X_i, Y_i\right), i \in D_1 \cup D_2, X_{n+1} = x\right) \geq 1-\alpha$



Quantile as Random Variable

- By change of variable $u = F(x)$, the expectation of a random variable X is the integral of its quantile function over its domain

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 Q(u) du$$

- Consider the conditional CDF of loss functions given θ

$$F(\ell \mid \theta) \triangleq \mathbb{P}\{\ell(z_{\text{new}}, \lambda) \leq \ell \mid \theta\}$$

- The corresponding quantile function is

$$K_\theta(t) \equiv F_\theta^{-1}(t) = \inf\{\ell : F_\theta(\ell) \geq t\}$$

- Loss ℓ as a random variable, then the expected loss is

$$\mathbb{E}[\ell(z, \lambda)] \triangleq J[K] = \int_0^1 K(t) dt$$

as a functional of the latent quantile function $K(t)$

Reformulate CP as BQ

- Idea: Instead of working directly on $p(\theta \mid z_{1:n})$, reparametrize the model using Bayesian Quadrature for quantile function
- Recall the expected loss (risk) over future data is

$$L(\theta, \lambda) = \int \ell\left(z_{\text{new}}, \lambda\right) f\left(z_{\text{new}} \mid \theta\right) dz_{\text{new}} = \mathbb{E}_{z \mid \theta}\left[\ell(z_{\text{new}}, \lambda)\right] \triangleq J[K]$$

- The posterior risk given the observed individual losses $\ell_i = \ell(z_i, \lambda)$ is

$$\begin{aligned} r(\lambda \mid z_{1:n}) = \mathbb{E}[L \mid z_{1:n}] &= \int L(\theta, \lambda) p\left(\theta \mid z_{1:n}\right) d\theta \\ = \mathbb{E}\left[L \mid \ell_{1:n}\right] &= \int J[K] p(K \mid \ell_{1:n}) dK \end{aligned}$$

- The posterior over quantile functions is

$$p\left(K \mid \ell_{1:n}\right) = \int p\left(K \mid t_{1:n}, \ell_{1:n}\right) p\left(t_{1:n} \mid \ell_{1:n}\right) dt_{1:n}$$

$$p\left(K \mid t_{1:n}, \ell_{1:n}\right) \propto \pi(K) \prod_{i=1}^{n} \delta\left(\ell_i - K\left(t_i\right)\right)$$

Remove Prior Specification by an Upper Bound

- But to be assumption-free, we need to avoid specifying prior $p(K)$!
- An upper bound on the posterior risk by the right rectangular rule

Theorem 4.1

Let $t_{(0)} = 0, t_{(n+1)} = 1$, and $\ell_{(n+1)} = B$. Then

$$\sup_{\pi} E\left(L \mid t_{1:n}, \ell_{1:n}\right) \leq \sum_{i=1}^{n+1} u_i \ell_{(i)}$$

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

- LHS $=$ worst posterior risk
- RHS $=$ piecewise-constant numerical quadrature approximation as a weighted sum of the observed losses by the spacing between consecutive quantiles

Illustration

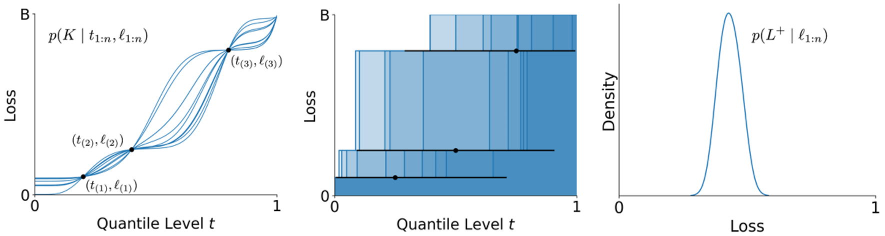


Figure: Left: Bayesian quadrature places a prior over the quantile function of the loss distribution. In practice, quantile levels are not observed.
Middle: quantile spacings with a right rectangular integration rule to construct an upper bound on the posterior distribution of the expected loss. Randomly sampled spacings and corresponding quantile functions are shown in blue along with a 95% credible interval for each quantile level in black.
Right: The posterior distribution for a random variable L^+ that upper bounds the expected loss is constructed by integrating over the unknown quantile levels $t_{1:n}$.

Illustration of Theorem 4.1

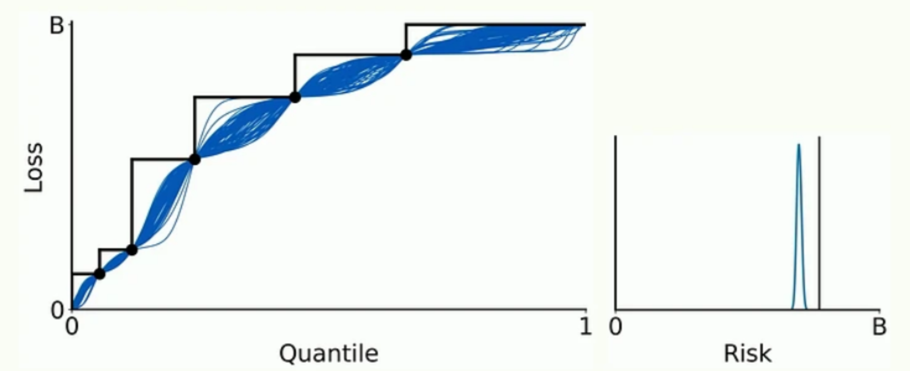


Figure: Regardless of priors, the stepwise function gives an upper bound on the risk (in black)

Dirichlet Quantile Spacings & Bound Maximum Risk

- Note that we only observe the loss values $\ell_{1:n}$, not the quantiles $t_{1:n}$
- It turns out the distribution of quantile spacings follows a Uniform Dirichlet $Dir(1, \ldots, 1)$, independent of the loss distribution.

Theorem 4.3

Define $\ell_{(i)}$ to be the order statistics of $\ell_1, \ldots, \ell_n$ for $i = 1, \ldots, n$ and $\ell_{(n+1)} \triangleq B$. Let L^+ be the random variable defined as follows:

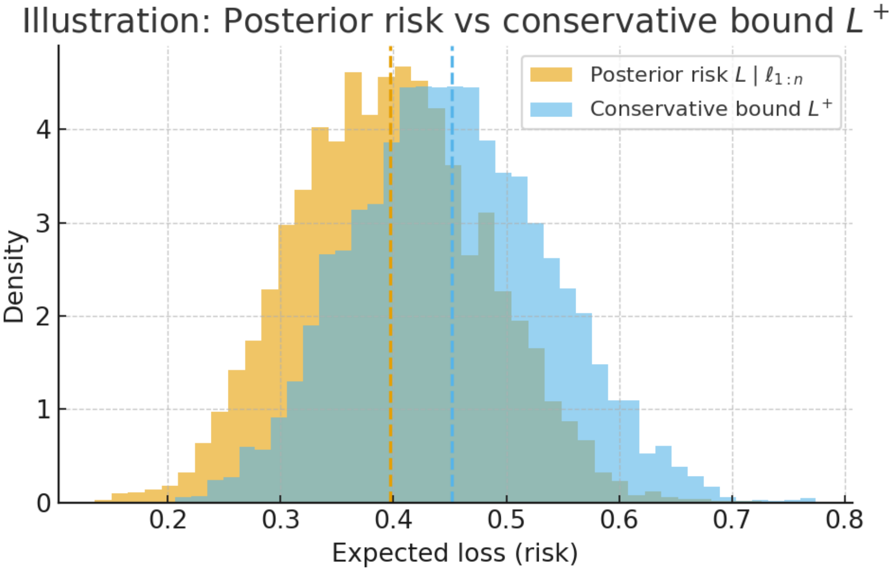
$$U_1, \ldots, U_{n+1} \sim \mathrm{Dir}(1, \ldots, 1), L^+ = \sum_{i=1}^{n+1} U_i \ell_{(i)}$$

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

$$\inf_{\pi} \Pr\left(L \leq b \mid \ell_{1:n}\right) \geq \Pr\left(L^+ \leq b\right)$$

L^+ stochastically dominates posterior risk (density of L^+ more at the right)

Illustration of Theorem 4.3



Recover Conformal Methods by Posterior Mean

Then we can directly construct upper confidence bounds as follows:

Corollary 4.4

For any desired coverage level $1-\alpha \in (0,1)$, define

$$b^*_{1-\alpha} = \inf_b \left\{ b : \Pr\left(L^+ \le b \mid \ell_{1:n} \right) \ge 1-\alpha \right\}.$$

Then $\inf_\pi \Pr\left(L \le b \mid \ell_{1:n}\right) \ge 1-\alpha$ for any $b \ge b^*_{1-\alpha}$.

The expected value of L^+ recovers conformal methods:

- Split CP:

$$E\left(L^+\right) = \frac{1}{n+1}\left(n+1 - \sum_{i=1}^{n} \mathbb{1}\left\{ s_i \le s_{(k)} \right\} \right) = 1 - \frac{k}{n+1}$$

- Conformal Risk Control:

$$E\left(L^+\right) = \sum_{i=1}^{n+1} E\left(U_i\right) \ell_{(i)} = \frac{1}{n+1}\left(\sum_{i=1}^{n} \ell_i + B \right)$$