Distribution-free uncertainty quantification for classification under label shift

Aleksandr Podkopaev, Aaditya Ramdas

Carnegie Mellon University

Carnegie Mellon University

Setup

Distribution-free (DF) uncertainty quantification

Conformal predictive inference | Calibrating probabilistic output

Conformal classification Construct $C : \mathcal{X} \to 2^{\mathcal{Y}}$:

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geqslant 1 - \alpha.$$

Calibration A predictor $f : \mathcal{X} \to \Delta_K$ is calibrated if

$$\mathbb{P}\left(Y = y \mid f(X)\right) = f_y(X), \quad y \in \mathcal{Y} = \{1, \dots, K\}.$$

Let P, Q stand for the source (generating training data) and target (generating test data) distributions defined on $\mathcal{X} \times \mathcal{Y}$.

Label shift assumption $q(x \mid y) = p(x \mid y)$, $q(y) \neq p(y)$.

Exchangeable (split-)conformal

The form of the *oracle* prediction sets when $\pi_y(x) = \mathbb{P}[Y = y \mid X = x]$ is known suggests to conformalize the following sequence of nested sets ($u \sim \mathrm{Unif}([0,1])$):

$$\mathcal{F}_\tau(x, u; \widehat{\pi}) = \{y : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leqslant \tau\}, \quad \tau \in [0,1],$$
$$\rho_y(x; \widehat{\pi}) = \sum_{y'} \widehat{\pi}_{y'}(x) \mathbb{1}\{\widehat{\pi}_{y'}(x) > \widehat{\pi}_y(x)\}.$$

For any triple (X, Y, U), its non-conformity score:

$$r(X, Y, U) = \inf\{\tau \in \mathcal{T} : \rho_Y(X; \widehat{\pi}) + U \cdot \widehat{\pi}_Y(X) \leqslant \tau\} = \rho_Y(X; \widehat{\pi}) + U \cdot \widehat{\pi}_Y(X).$$

Choose $\tau^\star = Q_{1-\alpha}\left(\{r_i\}_{i \in \mathcal{I}_{\mathrm{cal}}} \cup \{1\}\right)$. Then:

$$\mathbb{P}\left(Y_{n+1} \in \mathcal{F}_{\tau^\star}(X_{n+1}, U_{n+1}; \widehat{\pi}) \mid \{(X_i, Y_i)\}_{i \in \mathcal{I}_{\mathrm{tr}}}\right) \geqslant 1 - \alpha.$$

Calibration for i.i.d. data

Binning is necessary for obtaining DF guarantees: $\Delta_K = \bigcup_{m=1}^M B_m$, $B_i \cap B_j = \varnothing$, $i \neq j$. In the binary setting uniform-mass, or equal frequency, binning guarantees a sufficient number of calibration data points in each bin. To achieve approximate calibration, use empirical frequencies of class labels in each bin:

$$\widehat{\pi}^P_{y,m} = \frac{1}{N_m} \sum_{i=1}^n \mathbb{1}\{Y_i = y, \ f(X_i) \in B_m\},$$
$$N_m = |\{i \in \mathcal{I}_{\mathrm{cal}} : f(X_i) \in B_m\}|.$$

Let $h : \mathcal{X} \to \Delta_K$ denote the 'recalibrated' predictor: $h(x) = \widehat{\pi}_{g(x)}$ where $g : \mathcal{X} \to \mathcal{M}$ is the bin-mapping function: $g(x) = m \Leftrightarrow f(x) \in B_m$. For any given $\alpha \in (0,1)$, we show that with probability $\geqslant 1 - \alpha$, $\left\|\widehat{\pi}^P_m - \pi^P_m\right\|_1$, where

$$\varepsilon_m := \frac{2}{\sqrt{N_m}} \sqrt{\frac{1}{2} \ln\left(\frac{M 2^K}{\alpha}\right)}.$$

Consequently, it implies approximate calibration of the resulting predictor.

Label-shifted conformal

Let $w(y) = q(y)/p(y)$ (*importance weights*). Then:

$$\mathcal{F}^{(w)}(x, u; \widehat{\pi}) = \{y : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leqslant \tau^\star_w(y)\},$$
$$\tau^\star_w(y) = Q_{1-\alpha}\left(\sum_{i=1}^n \widetilde{p}^w_i(y) \delta_{r_i} + \widetilde{p}^w_{n+1}(y) \delta_1\right),$$
$$\widetilde{p}^w_i(y) = \frac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(y)},$$
$$\widetilde{p}^w_{n+1}(y) = \frac{w(y)}{\sum_{j=1}^n w(Y_j) + w(y)},$$

are provably valid (the proof relies on the concept of *weighted exchangeability*).

Exchangeability arguments yield a guarantee for known importance weights, in practice only an estimator is available. If a consistent estimator $\widehat{w}_k$ is used, then under mild assumptions:

$$\lim_{k \to \infty} \mathbb{P}(Y_{n+1} \in \mathcal{F}^{(\widehat{w}_k)}(X_{n+1}, U_{n+1}; \widehat{\pi}) \mid \{(X_i, Y_i)\}_{i \in \mathcal{I}_{\mathrm{tr}}}) \geqslant 1 - \alpha,$$

where $k = |\mathcal{D}_{\mathrm{est}}|$ is the size of sets used for constructing $\widehat{w}_k$.

- **Simulated data** Class proportions: $p = (0.1, 0.6, 0.3)$ and $q = (0.3, 0.2, 0.5)$. Covariates: $X \mid Y = y \sim \mathcal{N}(\mu_y, \Sigma)$ where $\mu_1 = (-2; 0)^\top$, $\mu_2 = (2; 0)^\top$, $\mu_3 = (0; 2\sqrt{3})^\top$, $\Sigma = \mathrm{diag}(4, 4)$.
- **Real data** `Wine quality dataset` with $p = (0.1, 0.4, 0.5)$, $q = (0.4, 0.5, 0.1)$.

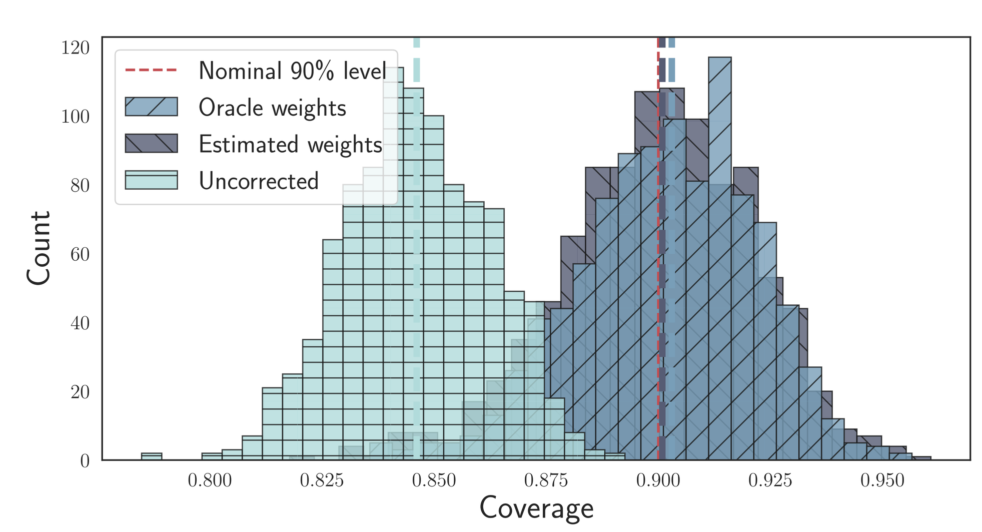


Fig. 2: Coverage on the simulated dataset.

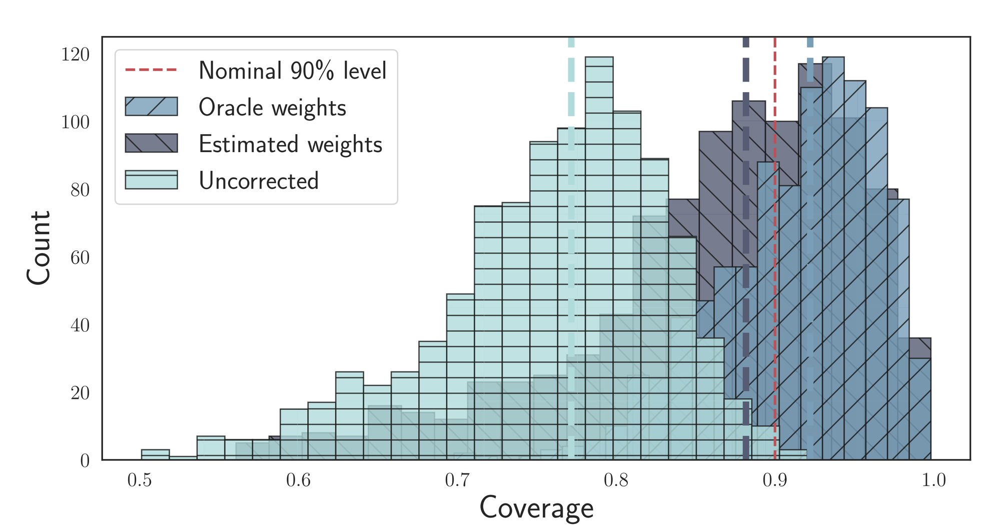


Fig. 3: Coverage on the `wine quality` dataset.

Label-conditional conformal (LCC)

Choose a set of significance levels for each class $\{\alpha_y\}_{y \in \mathcal{Y}}$ (e.g., $\alpha_y = \alpha$). Split the calibration set $\mathcal{I}_{\mathrm{cal}}$ into $|\mathcal{Y}| = K$ groups: $\mathcal{I}^y_{\mathrm{cal}} := \{i \in \mathcal{I}_{\mathrm{cal}} : Y_i = y\}$. Consider:

$$\mathcal{F}^c(x, u; \widehat{\pi}) = \{y : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leqslant \tau^\star_c(y)\},$$
$$\tau^\star_c(y) = Q_{1-\alpha_y}\left(\{r_i\}_{i \in \mathcal{I}^y_{\mathrm{cal}}} \cup \{1\}\right).$$

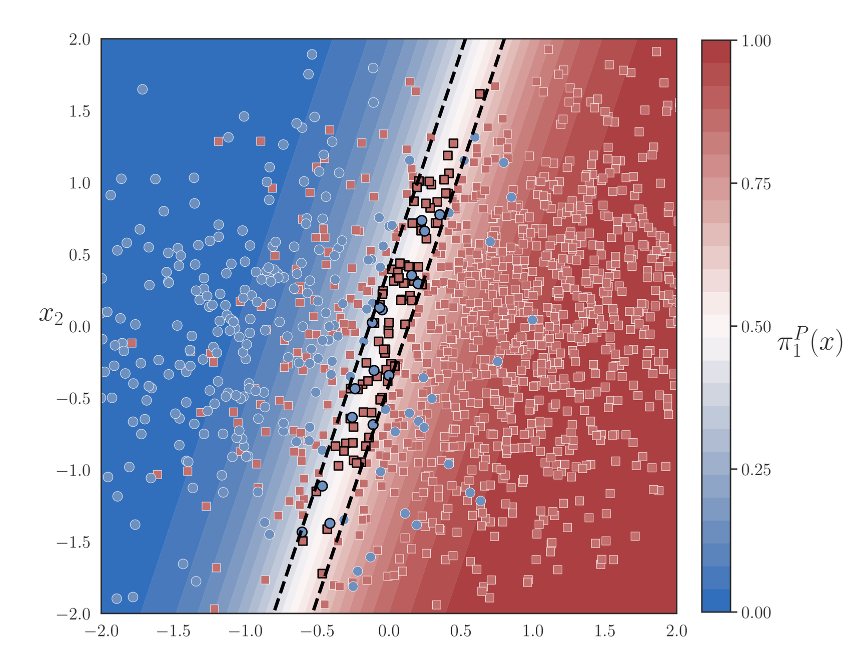
Then for any $y \in \mathcal{Y}$:

$$\mathbb{P}\left(Y_{n+1} \notin \mathcal{F}^c(X_{n+1}, U_{n+1}; \widehat{\pi}) \mid Y_{n+1} = y\right) \leqslant \alpha_y.$$

- LCC yields a stronger guarantee which makes it automatically robust to changes in class proportions. The price to pay is given by larger prediction sets.
- LCC does not require importance weights estimation and has exact finite-sample guaranee.
- LCC requires splitting available calibration data into K parts that could result in large losses of statistical efficiency when the number of classes K is large.

Label shift hurts calibration

- Data are sampled from a mixture of two Gaussians: $p(0) = p(1) = 1/2$ and $q(0) = 0.2$, $q(1) = 0.8$.
- The Bayes-optimal rule $\pi^P_1(x)$, which is calibrated, is plotted using the background coloring.
- The area $S = \left\{x \in \mathbb{R}^2 : \pi^P_1(x) \in [0.4; 0.6]\right\}$ has boundary given by the black dashed lines.



Label-shifted calibration

Bayes rule suggests an appropriate correction for achieving approximate calibration on the target:

$$\widehat{\pi}^{(\widehat{w})}_{y,m} = \frac{\widehat{w}(y) \cdot \widehat{\pi}^P_{y,m}}{\sum_{k=1}^K \widehat{w}(k) \cdot \widehat{\pi}^P_{k,m}}, \quad y \in \mathcal{Y}, \quad m \in \{1, \dots, M\}.$$

Performance depends on the *condition number*:

$$\kappa := \sup_k w(k) / \inf_{k : w(k) \neq 0} w(k),$$

with $\kappa = 1$ corresponding to label shift not being present.

Theorem 1. *For any bin $m \in \mathcal{M}$, it holds that:*

$$\left\|\widehat{\pi}^{(\widehat{w})}_m - \pi^Q_m\right\|_1 \leqslant \underbrace{2\kappa \cdot \left\|\widehat{\pi}^P_m - \pi^P_m\right\|_1}_{(a)} + \underbrace{\frac{2\left\|\widehat{w} - w\right\|_\infty}{\inf_{l : w(l) \neq 0} w(l)}}_{(b)}.$$

(a) is controlled by the calibration error on the source and (b) is controlled by the importance weights estimation error.

Label-shifted calibration yields an approximately calibrated predictor on the target while uncorrected fails.

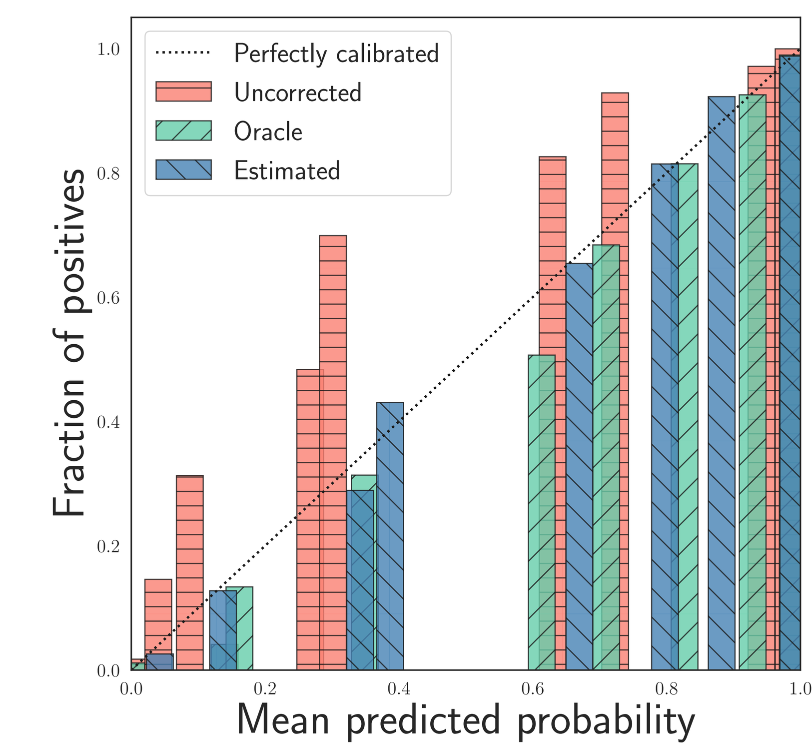


Fig. 5: Example of a reliability curve on the simulated dataset.

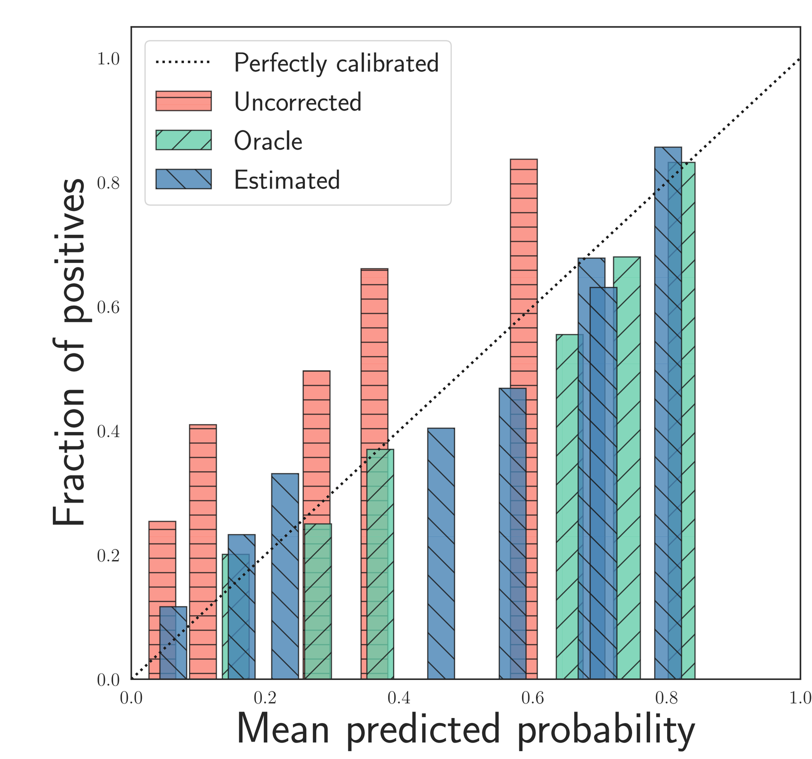


Fig. 6: Example of a reliability curve on the `wine quality` dataset.