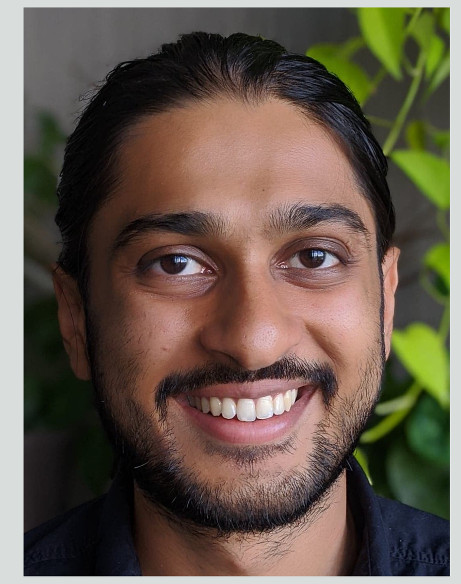
# Distribution-free binary classification: prediction sets, confidence intervals and calibration

Chirag Gupta, Aleksandr Podkopaev, Aaditya Ramdas

Carnegie Mellon University



NEURAL INFORMATION PROCESSING SYSTEMS

## Notions of uncertainty quantification for classification

**Setup.** Let $\mathcal{X}$ and $\mathcal{Y} = \{0, 1\}$ denote the feature and label spaces for binary classification. Given predictor $f : \mathcal{X} \to \mathcal{Z}$ (e.g. $\mathcal{Z} = [0, 1]$ for logistic regression, $\mathcal{Z} = \mathbb{R}$ for SVM) trained on some labeled data and an independent sample $\{(X_i, Y_i)\}_{i\in[n]} \sim P^n$, we consider a question of providing a measure of uncertainty for the produced prediction in distribution-free setting, i.e. without making assumptions on $P$.

**Confidence Intervals (CI) and Prediction Sets (PS).** Let $\mathcal{I}$ denote the set of all subintervals of $[0, 1]$ and denote $\mathcal{L} \equiv \{\{0\}, \{1\}, \{0, 1\}, \varnothing\}$.

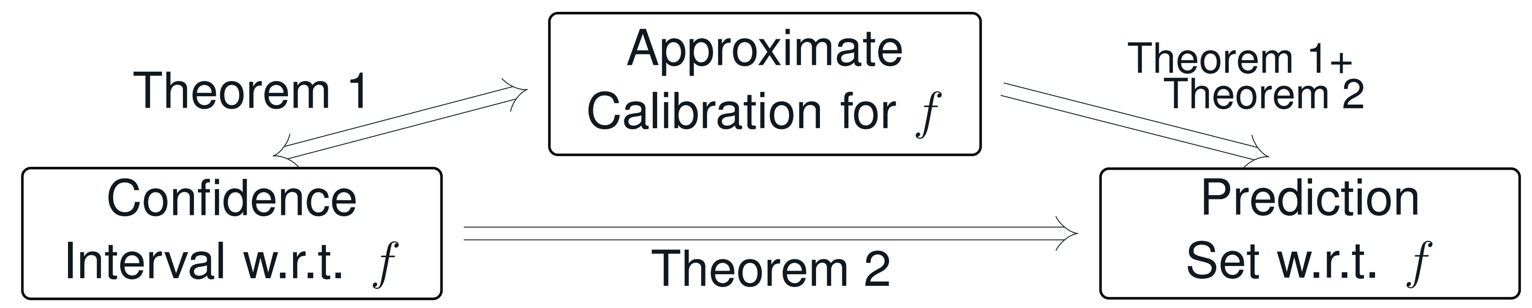
- A function $C : \mathcal{Z} \to \mathcal{I}$ is a $(1 - \alpha)$-CI with respect to $f : \mathcal{X} \to \mathcal{Z}$ if
$$\mathbb{P}(\mathbb{E}[Y \mid f(X)] \in C(f(X))) \geqslant 1 - \alpha.$$
- A function $S : \mathcal{Z} \to \mathcal{L}$ is a $(1 - \alpha)$-PS with respect to $f : \mathcal{X} \to \mathcal{Z}$ if
$$\mathbb{P}(Y \in S(f(X))) \geqslant 1 - \alpha.$$

**Perfect Calibration.** A predictor $f : \mathcal{X} \to [0, 1]$ is (perfectly) calibrated if
$$\mathbb{E}[Y \mid f(X) = a] = a \quad \text{a.s. for all } a \text{ in the range of } f.$$

**Approximate Calibration.** A predictor $f : \mathcal{X} \to [0, 1]$ is $(\varepsilon, \alpha)$-approximately calibrated for some $\alpha \in (0, 1)$ and a function $\varepsilon : [0, 1] \to [0, 1]$ if with probability at least $1 - \alpha$, we have
$$|\mathbb{E}[Y|f(X)] - f(X)| \leqslant \varepsilon(f(X)).$$

**Asymptotic Calibration.** A sequence of predictors $\{f_n\}_{n\in\mathbb{N}}$ from $\mathcal{X} \to [0, 1]$ is asymptotically calibrated at level $\alpha \in (0, 1)$ if there exists a sequence of functions $\{\varepsilon_n\}_{n\in\mathbb{N}}$ such that $f_n$ is $(\varepsilon_n, \alpha)$-approximately calibrated for every $n$, and $\varepsilon_n(f_n(X_{n+1})) = o_P(1)$.

## Relationship between notions




**Theorem 1.** *Let $f : \mathcal{X} \to [0, 1]$ be a predictor that is $(\varepsilon, \alpha)$-approximately calibrated for some function $\varepsilon$. Then the function $C$:*
$$C(f(x)) = [f(x) - \varepsilon(f(x)), f(x) + \varepsilon(f(x))], \tag{1}$$
*is a $(1 - \alpha)$-CI with respect to $f$.*

**Corollary 1.** *If a sequence of predictors $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated at level $\alpha$, then* (1) *yields a sequence $\{C_n\}_{n\in\mathbb{N}}$ such that each $C_n$ is a $(1 - \alpha)$-CI with respect to $f_n$ and $|C_n(f_n(X_{n+1}))| = o_P(1)$.*

**Theorem 2.** *Fix $f : \mathcal{X} \to \mathcal{Z}$. If $\widehat{C}_n$ is a $(1 - \alpha)$-CI with respect to $f$ for all distributions $P$, then disc$(\widehat{C}_n) = \widehat{C}_n \cap \{0, 1\} \subseteq \mathcal{L}$ is a $(1 - \alpha)$-PS with respect to $f$ for all distributions $P$ for which $P_{f(X)}$ is nonatomic.*

**Corollary 2.** *Fix $f : \mathcal{X} \to \mathcal{Z}$. If $\widehat{C}_n$ is a $(1 - \alpha)$-CI with respect to $f$ for all $P$, and there exists a $P$ such that $P_{f(X)}$ is nonatomic, then we can construct a distribution $Q$ such that $\mathbb{E}_{Q^{n+1}}|\widehat{C}_n(f(X_{n+1}))| \geqslant 0.5 - \alpha$.*

## Necessary condition for asymptotic calibration in distribution-free setting

**Partition view-point.** Actual values taken by $f$ are only as informative as the *partition* of $\mathcal{X}$ provided by its level sets. Denote this partition as $\{\mathcal{X}_z\}_{z\in\mathcal{Z}}$, where $\mathcal{X}_z = \{x \in \mathcal{X} : f(x) = z\}$.

**Theorem 3** (informal). *If a sequence $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated at level $\alpha$ for all $P$, then the cardinality of the partition induced by $f_n$ must be at most countable for large enough $n$.*

**Implications.** Popular continuous scoring functions such as logistic regression, deep neural-nets with softmax output and SVMs cannot be asymptotically calibrated without distributional assumptions.
This impossibility result can be extended to many parametric calibration schemes that 're-calibrate' an existing $f$ through a wrapper $h_n : \mathcal{Z} \to [0, 1]$ learnt on the calibration data (Platt/temperature scaling, beta calibration).

## Achieving approximate calibration in distribution-free setting via binning

**Notation.** Sample-space $\mathcal{X}$ is partitioned into $B$ regions $\{\mathcal{X}_b\}_{b\in[B]}$ with $\pi_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$ being the expected label probability in $\mathcal{X}_b$. Denote the partition-identity function as $\mathcal{B} : \mathcal{X} \to [B]$ where $\mathcal{B}(x) = b$ if and only if $x \in \mathcal{X}_b$. Let $\widehat{s}_b := |\{i \in [n] : \mathcal{B}(X_i) = b\}|$ be the number of points from the calibration set that belong to region $\mathcal{X}_b$. Define
$$\widehat{\pi}_b := \frac{1}{\widehat{s}_b} \sum_{i:\mathcal{B}(X_i)=b} Y_i \quad \text{and} \quad \widehat{V}_b := \frac{1}{\widehat{s}_b} \sum_{i:\mathcal{B}(X_i)=b} (Y_i - \widehat{\pi}_b)^2$$
as the empirical average and variance of the $Y$ values in a partition.

**Theorem 4.** *For any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$,*
$$|\pi_b - \widehat{\pi}_b| \leqslant \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{\widehat{s}_b}} + \frac{3\ln(3B/\alpha)}{\widehat{s}_b},$$
*simultaneously for all $b \in [B]$.*

Let $b^\star = \arg\min_{b\in[B]} \widehat{s}_b$ denote the index of the region with the minimum number of calibration examples.

**Corollary 3.** *For $\alpha \in (0, 1)$, $f_n(x) := \widehat{\pi}_{\mathcal{B}(x)}$ is $(\varepsilon, \alpha)$-approximately calibrated with*
$$\varepsilon(\cdot) = \sqrt{\frac{\widehat{V}_{b^\star} \ln(3B/\alpha)}{2\widehat{s}_{b^\star}}} + \frac{3\ln(3B/\alpha)}{2\widehat{s}_{b^\star}}. \tag{2}$$
*Thus, $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated at level $\alpha$.*

Results are also generalized to online setting when extra calibration data can be queried until a desired confidence level and covariate shift setting when the test data distribution changes, but unlabeled data from a 'target' domain is available.

**Data-dependent sample-space partition.** Guarantee 2 can be unsatisfactory if the sample-space partition is constructed poorly. Uniform-mass binning is a partitioning scheme based on the sample splitting idea that provably guarantees that $\widehat{s}_{b^\star}$ scales as $\Omega(n/B)$ with high probability.