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

Chirag Gupta[*1], **Aleksandr Podkopaev**[*1,2], **Aaditya Ramdas**[1,2]
Machine Learning Department[1]
Department of Statistics and Data Science[2]
Carnegie Mellon University
{chiragg,podkopaev,aramdas}@cmu.edu

Abstract

We study three notions of uncertainty quantification—calibration, confidence intervals and prediction sets—for binary classification in the distribution-free setting, that is without making any distributional assumptions on the data. With a focus towards calibration, we establish a 'tripod' of theorems that connect these three notions for score-based classifiers. A direct implication is that distribution-free calibration is only possible, even asymptotically, using a scoring function whose level sets partition the feature space into at most countably many sets. Parametric calibration schemes such as variants of Platt scaling do not satisfy this requirement, while nonparametric schemes based on binning do. To close the loop, we derive distribution-free confidence intervals for binned probabilities for both fixed-width and uniform-mass binning. As a consequence of our 'tripod' theorems, these confidence intervals for binned probabilities lead to distribution-free calibration. We also derive extensions to settings with streaming data and covariate shift.

1 Introduction

Let $\mathcal{X}$ and $\mathcal{Y} = \{0, 1\}$ denote the feature and label spaces for binary classification. Consider a predictor $f : \mathcal{X} \to \mathcal{Z}$ that produces a prediction in some space $\mathcal{Z}$. If $\mathcal{Z} = \{0, 1\}$, f corresponds to a point prediction for the class label, but often class predictions are based on a 'scoring function'. Examples are, $\mathcal{Z} = \mathbb{R}$ for SVMs, and $\mathcal{Z} = [0, 1]$ for logistic regression, random forests with class probabilities, or deep models with a softmax top layer. In such cases, a higher value of $f(X)$ is often interpreted as higher belief that $Y = 1$. In particular, if $\mathcal{Z} = [0, 1]$, it is tempting to interpret $f(X)$ as a probability, and hope that

$$f(X) \approx \mathbb{P}(Y = 1 \mid X). \tag{1}$$

However, such hope is unfounded, and in general (1) will be far from true without strong distributional assumptions, which may not hold in practice. Valid uncertainty estimates that are related to (1) can be provided, but ML models do not satisfy these out of the box. This paper discusses three notions of uncertainty quantification: calibration, prediction sets (PS) and confidence intervals (CI), defined next. A function $f : \mathcal{X} \to [0, 1]$ is said to be (perfectly) calibrated if

$$\mathbb{E}\left[Y \mid f(X) = a\right] = a \qquad \text{a.s. for all } a \text{ in the range of } f. \tag{2}$$

Define $\mathcal{L} \equiv \{\{0\}, \{1\}, \{0, 1\}, \varnothing\}$ and fix $\alpha \in (0, 1)$. A function $S : \mathcal{X} \to \mathcal{L}$ is a $(1 - \alpha)$-PS if

$$\mathbb{P}(Y \in S(X)) \geqslant 1 - \alpha. \tag{3}$$

Finally, let $\mathcal{I}$ denote the set of all subintervals of $[0, 1]$. A function $C : \mathcal{X} \to \mathcal{I}$ is a $(1 - \alpha)$-CI if

$$\mathbb{P}(\mathbb{E}\left[Y \mid X\right] \in C(X)) \geqslant 1 - \alpha. \tag{4}$$

[*]equal contribution

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

All three notions are 'natural' in their own sense, but also different at first sight. We show that they are in fact tightly connected (see Figure 1), and focus on the implications of this result for calibration. Our analysis is in the distribution-free setting, that is, we are concerned with understanding what kinds of valid uncertainty quantification is possible without distributional assumptions on the data.

Our work primarily extends the ideas of Vovk et al. [47, Section 5] and Barber [3]. We also discuss Platt scaling [36], binning [51] and the recent work of Vaicenavicius et al. [44]. Other related work is cited as needed, and further discussed in Section 5. All proofs appear ordered in the Appendix.

Notation: Let P denote any distribution over $\mathcal{X} \times \mathcal{Y}$. In practice, the available labeled data is often split randomly into the *training set* and the *calibration set*. Typically, we use n to denote the number of calibration data points, so $\{(X_i, Y_i)\}_{i\in[n]}$ is the calibration data, where we use the shorthand $[a] := \{1, 2, \dots a\}$. A prototypical test point is denoted (X_{n+1}, Y_{n+1}). All data are drawn i.i.d. from P, denoted succinctly as $\{(X_i, Y_i)\}_{i\in[n+1]} \sim P^{n+1}$. As above, random variables are denoted in upper case. The learner observes realized values of all random variables (X_i, Y_i), except Y_{n+1}. (All sets and functions are implicitly assumed to be measurable.)

2 Calibration, confidence intervals and prediction sets

Calibration captures the intuition of (1) but is a weaker requirement, and was first studied in the meteorological literature for assessing probabilistic rain forecasts [5, 7, 31, 39]. Murphy and Epstein [31] described the ideal notion of calibration, called *perfect calibration* (2), which has also been referred to as *calibration in the small* [45], or sometimes simply as *calibration* [7, 12, 44]. The types of functions that can achieve perfect calibration can be succinctly captured as follows.

Proposition 1. *A function $f : \mathcal{X} \to [0, 1]$ is perfectly calibrated if and only if there exists a space $\mathcal{Z}$ and a function $g : \mathcal{X} \to \mathcal{Z}$, such that*

$$f(x) = \mathbb{E}\left[Y \mid g(X) = g(x)\right] \quad \textit{almost surely } P_X. \tag{5}$$

Vaicenavicius et al. [44] stated and gave a short proof for the 'only if' direction. While the other direction is also straightforward, together they lead to an appealingly simple and complete characterization. The proof of Proposition 1 is in Appendix A.

It is helpful to consider two extreme cases of Proposition 1. First, setting g to be the identity function yields that the Bayes classifier $\mathbb{E}\left[Y|X\right]$ is perfectly calibrated. Second, setting $g(\cdot)$ to any constant implies that $\mathbb{E}\left[Y\right]$ is also a perfect calibrator. Naturally, we cannot hope to estimate the Bayes classifier without assumptions, but even the simplest calibrator $\mathbb{E}\left[Y\right]$ can only be approximated in finite samples. Since Proposition 1 states that calibration is possible iff the RHS of (5) is known exactly for some g, perfect calibration is impossible in practice. Thus we resort to satisfying the requirement (2) approximately, which is implicitly the goal of many empirical calibration techniques.

Definition 1 (Approximate calibration)**.** A predictor $f : \mathcal{X} \to [0, 1]$ is (ε, α)-approximately calibrated for some $\alpha \in (0, 1)$ and a function $\varepsilon : [0, 1] \to [0, 1]$ if with probability at least $1 - \alpha$,

$$|\mathbb{E}\left[Y|f(X)\right] - f(X)| \leqslant \varepsilon(f(X)). \tag{6}$$

Note that when the definition is applied to a test point (X_{n+1}, Y_{n+1}), there are two sources of randomness: the randomness in (X_{n+1}, Y_{n+1}), and the randomness in f, ε. The randomness in f, ε can be statistical (through random training data) or algorithmic (if the learning procedure is non-deterministic). All probabilistic statements in this paper should be viewed through this lens. Often in this paper, ε is a constant function. In this case, we represent its value simply as ε. With increasing data, a good calibration procedure should have ε vanishing to 0. We formalize this next.

Definition 2 (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 (ε_n, α)-approximately calibrated for every n, and $\varepsilon_n(f_n(X_{n+1})) = o_P(1)$.

For readers familiar with the ℓ_1-ECE metric for calibration [32], we note that the ℓ_1-ECE $\mathbb{E}_X\left|\mathbb{E}\left[Y|f(X)\right] - f(X)\right|$ is closely related to (6). For instance, when ε is a constant function, (6) implies that with probability at least $1 - \alpha$ (over ε, f), $\mathbb{E}_X\left|\mathbb{E}\left[Y|f(X)\right] - f(X)\right| \leqslant (1 - \alpha)\varepsilon + \alpha \leqslant \varepsilon + \alpha$. (Some readers may be more familiar with the definition of ℓ_1-ECE for vector-valued predictions in multiclass problems. The ECE of the vectorized prediction $[1 - f, f]$ is also related to (6) since $\mathbb{E}_X\left|\mathbb{E}\left[Y|f(X)\right] - f(X)\right| = 0.5 \cdot \ell_1\text{-ECE}([1 - f, f])$.)

We show that the notions of approximate and asymptotic calibration are related to prediction sets (3) and confidence intervals (4). PSs and CIs are only 'informative' if the sets or intervals produced by

them are small: confidence intervals are measured by their length (denoted as $|C(\cdot)|$), and prediction sets are measured by their diameter ($\text{diam}(S(\cdot)) := |\text{convex hull}(S(\cdot))|$). Observe that for binary classification, the diameter of a PS is either 0 or 1.

For a given distribution, one might expect prediction sets to have a larger diameter than the length of the confidence intervals, since we want to cover the actual value of Y_{n+1} and not its (conditional) expectation. As an example, if $\mathbb{E}[Y|X = x] = 0.5$ for every x, then the shortest possible confidence interval is $(0.5, 0.5]$ whose diameter is 0. However, a valid $(1-\alpha)$-PS has no choice but to output $\{0, 1\}$ for at least $(1-2\alpha)$ fraction of the points (and a random guess for the other 2α fraction), and thus must have expected diameter $\geqslant 1-2\alpha$ even in the limit of infinite data.

Recently, Barber [3] built on an earlier result of Vovk et al. [47] to show that if an algorithm provides an interval C which is a $(1-\alpha)$-CI for all product distributions P^{n+1} (of the training data and test-point), then $S := C \cap \{0, 1\}$ is also a $(1-\alpha)$-PS whenever P_X is a nonatomic distribution. Since this implication holds for all nonatomic distributions P_X, including the ones with $\mathbb{E}[Y|X] \equiv 0.5$ discussed above, it implies that distribution-free CIs must necessarily be wide. Specifically, their length cannot shrink to 0 as $n \to \infty$. This can be treated as an impossibility result for the existence of (distribution-free) informative CIs.

One way to circumvent these impossibilities is to consider CIs for functions with 'lower resolution' than $\mathbb{E}[Y|X]$. To this end, we introduce a notion of a CI or PS 'with respect to f' (w.r.t.f). As we discuss in Section 3 (and Section 3.1 in particular), these notions are connected to calibration.

Definition 3 (CI or PS w.r.t. f)**.** A function $C : \mathcal{Z} \to \mathcal{I}$ is a $(1-\alpha)$-CI with respect to $f : \mathcal{X} \to \mathcal{Z}$ if

$$\mathbb{P}(\mathbb{E}[Y \mid f(X)] \in C(f(X))) \geqslant 1-\alpha. \tag{7}$$

Analogously, a function $S : \mathcal{Z} \to \mathcal{L}$ is a $(1-\alpha)$-PS with respect to $f : \mathcal{X} \to \mathcal{Z}$ if

$$\mathbb{P}(Y \in S(f(X))) \geqslant 1-\alpha. \tag{8}$$

When instantiated for a test point (X_{n+1}, Y_{n+1}), the probability in definitions (7) and (8) is not only over the test point, but also over the randomness in the pair (f, C) or (f, S), which are usually learned on labeled data. In order to produce PSs and CIs, one typically fixes a function f learned on an independent split of the labeled data, and considers learning a C or S that provides guarantees (7) and (8). For example, S can be produced using inductive conformal techniques [26, 34, 37]. In this case, C or S would be random as well; to make this explicit, we often denote C or S as $\widehat{C}_n$ or $\widehat{S}_n$.

3 Relating notions of distribution-free uncertainty quantification

As preluded to above, we consider a standard setting for valid distribution-free uncertainty quantification where the 'training' data is used to learn a scoring function $f : \mathcal{X} \to \mathcal{Z}$ and then held-out 'calibration' data is used to estimate uncertainty. We establish that in this setting, the notions of calibration, PSs and CIs are closely related. Figure 1 summarizes this section's takeaway message. Here, and in the rest of the section, if P is the distribution of data, then we denote the distribution of the random variable $Z = f(X)$ as $P_{f(X)}$.

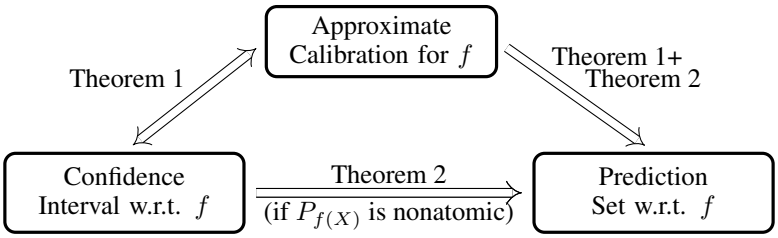


Figure 1: Relationship between notions of distribution-free uncertainty quantification.

In Section 3.1, we show that if an algorithm provides a CI, it can be used to provide a calibration guarantee and vice-versa (Theorem 1). In Section 3.2, we show that for all distributions P such that $P_{f(X)}$ is nonatomic, if an algorithm constructs a distribution-free CI with respect to f, then it can be used to construct a distribution-free PS with respect to f (Theorem 2). However, one typically expects the length of CIs to shrink to 0 in the limit of infinite data, whereas PSs have a fixed distribution-dependent lower bound on their diameter. In Section 3.3, we use this intuition to prove a formal impossibility result for asymptotic calibration (Theorem 3). This result shows that for a

large class of standard scoring functions f (such as logistic regression, deep networks with a final softmax layer, SVMs), it is impossible to achieve distribution-free asymptotic calibration without a 'discretization' step. Parametric schemes such as Platt scaling [36] do not perform such discretization and thus cannot lead to distribution-free calibration. To complement this lower bound, we provide calibration guarantees for one possible discretization step (histogram binning) in Section 4.

3.1 Relating calibration and confidence intervals

Given a predictor f that is (ε, α)-approximately calibrated, there is a trivial way to construct a function C that is a $(1-\alpha)$-CI: for $x \in \mathcal{X}$,

$$\underbrace{|\mathbb{E}\left[Y \mid f(x)\right] - f(x)| \leqslant \varepsilon(f(x))}_{\text{calibration}} \implies \underbrace{\mathbb{E}\left[Y \mid f(x)\right] \in C(f(x))}_{\text{CI w.r.t. } f} := [f(x) - \varepsilon(f(x)), f(x) + \varepsilon(f(x))]. \quad (9)$$

On the other hand, given C that is a $(1-\alpha)$-CI with respect to f, define for $z \in \text{Range}(f)$ the left-endpoint, right-endpoint and midpoint functions respectively:

$$u_C(z) := \sup\{g : g \in C(z)\},\ l_C(z) := \inf\{g : g \in C(z)\},\ m_C(z) := (u_C(z) + l_C(z))/2. \quad (10)$$

Consider the midpoint $m_C(f(x))$ as a 'corrected' prediction for $x \in \mathcal{X}$:

$$\widetilde{f}(x) := m_C(f(x)),\ x \in \mathcal{X}, \quad (11)$$

and let $\varepsilon = \sup_{z \in \text{Range}(f)} \{|C(z)|/2\}$ be a constant function returning the largest interval radius. Then $\widetilde{f}$ is (ε, α)-approximately calibrated for a non-trivial ε. These claims are formalized next.

Theorem 1. *Fix any $\alpha \in (0,1)$. Let $f : \mathcal{X} \to [0,1]$ be a predictor that is (ε, α)-approximately calibrated for some function ε. Then the function C in* (9) *is a $(1-\alpha)$-CI with respect to f.*

Conversely, fix a scoring function $f : \mathcal{X} \to \mathcal{Z}$. If C is a $(1-\alpha)$-CI with respect to f, then the predictor $\widetilde{f}$ in (11) *is (ε, α)-approximately calibrated for $\varepsilon = \sup_{z \in \text{Range}(f)} \{|C(z)|/2\}$.*

The proof is in Appendix B. An important implication of Theorem 1 is that having a sequence of predictors that is asymptotically calibrated yields a sequence of confidence intervals with vanishing length as $n \to \infty$. This is formalized in the following corollary, also proved in Appendix B.

Corollary 1. *Fix any $\alpha \in (0,1)$. If a sequence of predictors $\{f_n\}_{n \in \mathbb{N}}$ is asymptotically calibrated at level α, then construction* (9) *yields a sequence of functions $\{C_n\}_{n \in \mathbb{N}}$ such that each C_n is a $(1-\alpha)$-CI with respect to f_n and $|C_n(f_n(X_{n+1}))| = o_P(1)$.*

Next, we show that for a large class of scoring functions, CIs and PSs are also related in the distribution-free setting. This connection along with Corollary 2 (below) leads to an impossibility result for distribution-free asymptotic calibration for certain functions f (Theorem 3 in Section 3.3).

3.2 Relating distribution-free confidence intervals and prediction sets

Suppose a function satisfies a CI guarantee with respect to f no matter what the data-generating distribution P is. We show that such a function would also provide a PS guarantee for all P such that $P_{f(X)}$ is nonatomic. To write our theorem, we define the 'discretize' function to transform a confidence interval C to a prediction set: $\text{disc}(C) := C \cap \{0,1\} \subseteq \mathcal{L}$. In the following theorem, the CI and PS guarantees provided (per equations (7) and (8)) are to be understood as marginal over both the calibration and test-data. To make this explicit, we denote the CI function as $\widehat{C}_n$.

Theorem 2. *Fix $\alpha \in (0,1)$ and a function f with domain $\mathcal{X}$. If $\widehat{C}_n$ is a $(1-\alpha)$-CI with respect to f for all distributions P, then disc$(\widehat{C}_n)$ is a $(1-\alpha)$-PS with respect to f for all distributions P for which $P_{f(X)}$ is nonatomic.*

The proof is in Appendix B. It adapts the proof of Barber [3, Theorem 1]. Their result connects the notions of CI and PS, but not with respect to f (like in equations (3), (4)). By adapting the result for CIs and PSs with respect to f, and using Theorem 1, we are able to relate CIs and PSs to calibration and use this to prove an impossibility result for asymptotic calibration. This is done in the proof of Theorem 3 in the Section 3.3. A corollary of Theorem 2 that is used in Theorem 3 (but is also important on its own) is stated next.

Corollary 2. *Fix $f : \mathcal{X} \to \mathcal{Z}$ and $\alpha \in (0,1)$. 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*

$$\mathbb{E}_{Q^{n+1}} |\widehat{C}_n(f(X_{n+1}))| \geqslant 0.5 - \alpha.$$

The proof is in Appendix B. For a given f, the bound in the corollary needs existence of P such that $P_{f(X)}$ is nonatomic. These f are characterized in the discussion after Corollary 3 (Section 3.3), and formally in the proof of Theorem 3. One expects the length of a confidence interval to vanish as $n \to \infty$. Corollary 2 shows that this is impossible in a distribution-free manner for certain f.

3.3 Necessary condition for distribution-free asymptotic calibration

Proposition 1 shows that a function f is calibrated if and only if it takes the form (5) for some function g. Observe that in (5), the actual values taken by g are only as informative as the *partition* of $\mathcal{X}$ induced by its level sets. Denote this partition as $\{\mathcal{X}_z\}_{z\in\mathcal{Z}}$, where $\mathcal{X}_z = \{x \in \mathcal{X} : g(x) = z\}$. Then we may equivalently define f in (5) through a set of values $\{f_z = P(Y_{n+1} = 1 \mid X_{n+1} \in \mathcal{X}_z)\}_{z\in\mathcal{Z}}$, setting $f(\cdot) = f_{g(\cdot)}$. Thus we can re-interpret calibration as follows.

Corollary 3 (to Proposition 1)**.** *Any calibrated classifier f is characterized by a partition of $\mathcal{X}$ into subsets $\{\mathcal{X}_z\}_{z\in\mathcal{Z}}$ and corresponding conditional probabilities $\{f_z\}_{z\in\mathcal{Z}}$ for some index set $\mathcal{Z}$.*

We now establish a necessary condition on the cardinality of the partition corresponding to an asymptotically calibrated scoring function. In the following statement, $\mathcal{X}^{(f)}$ denotes the level sets of f on $\mathcal{X}$, and $|\mathcal{X}^{(f)}|$ denotes its cardinality (which may be infinite). Also $\aleph_0$ denotes the largest cardinality of a countable set, which is the cardinality of $\mathbb{N}$.

Theorem 3. *Let $\alpha \in (0, 0.5)$ be a fixed threshold. If a sequence of scoring functions $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated at level α for every distribution P then*

$$\limsup_{n\to\infty} |\mathcal{X}^{(f_n)}| \leqslant \aleph_0.$$

In words, the cardinality of the partition induced by f_n must be at most countable for large enough n. The following phrasing is convenient: f is said to lead to a *fine partition* of $\mathcal{X}$ if $|\mathcal{X}^{(f)}| > \aleph_0$. Then, for the purposes of distribution-free asymptotic calibration, Theorem 3 necessitates us to consider f that do not lead to fine partitions. Popular scoring functions such as logistic regression, deep neural-nets with softmax output and SVMs lead to continuous f that induce fine partitions of $\mathcal{X}$ and thus cannot be asymptotically calibrated without distributional assumptions.

The proof of Theorem 3 is in Appendix B, but we briefly sketch its intuition below. Corollary 1 shows that asymptotic calibration allows construction of CIs whose lengths vanish asymptotically. Corollary 2 shows however that asymptotically vanishing CIs are impossible (without distributional assumptions) for f if there exists a distribution P such that $P_{f(X)}$ is nonatomic. Consequently asymptotic calibration is also impossible for such f. If $\mathcal{Z} = \text{Range}(f)$ is countable, then by the axioms of probability, $\sum_{z\in\mathcal{Z}} \mathbb{P}(X \in \mathcal{X}_z) = \mathbb{P}(X \in \mathcal{X}) = 1$, and so $\mathbb{P}(X \in \mathcal{X}_z) \neq 0$ for at least some z. Thus $P_{f(X)}$ cannot be nonatomic for any P. On the other hand, if $\mathcal{Z}$ is uncountable we can show that there always exists a P such that $P_{f(X)}$ is nonatomic. Hence distribution-free asymptotic calibration is impossible for such f, which leads to the result of the theorem.

Theorem 3 also applies to many parametric calibration schemes that 'recalibrate' an existing f through a wrapper $h_n : \mathcal{Z} \to [0, 1]$ learnt on the calibration data, with the goal that $h_n(f(\cdot))$ is nearly calibrated: $\mathbb{E}[Y \mid h_n(f(X))] \approx h_n(f(X))$. For instance, consider methods like Platt scaling [36], temperature scaling [12] and beta calibration [20]. Each of these methods learns a continuous and monotonic[2] (hence bijective) wrapper h_n, and thus $\mathbb{E}[Y \mid h_n(f(X))] = \mathbb{E}[Y \mid f(X)]$. If h_n is a good calibrator, we would have $\mathbb{E}[Y \mid f(X)] \approx h_n(f(X))$. One way to formalize this is to consider whether an interval around $h_n(f(X))$ is a CI for $\mathbb{E}[Y \mid f(X)]$. In other words — does there exist a function $\varepsilon_n : [0, 1] \to [0, 1]$ such that for every distribution P,

$$\widetilde{C}_n(f(X)) := [h_n(f(X)) - \varepsilon_n(h_n(f(X))), h_n(f(X)) + \varepsilon_n(h_n(f(X)))]$$

is a $(1 - \alpha)$-CI with respect to f and $\varepsilon_n(h_n(f(X))) = o_P(1)$? Theorem 3 shows that this is impossible if f leads to a fine partition of $\mathcal{X}$, irrespective of the properties of h_n. Thus the aforementioned parametric calibration methods cannot lead to asymptotic calibration in general (that is, without further distributional assumptions). It is likely that the implications of our result also applies to other continuous parametric methods that are not necessarily monotonic, as well as calibration schemes that aim to learn a calibrated predictor without sample-splitting.

[2]This assumes that the parameters satisfy natural constraints as discussed in the original papers: $a, b \geqslant 0$ for beta scaling with at least one of them nonzero, $A < 0$ for Platt scaling and $T > 0$ for temperature scaling.

On the other hand, the calibration methods of isotonic regression [52] and histogram binning [51] do provide a countable partition of $\mathcal{X}$, and thus may satisfy distribution-free approximate calibration guarantees. In Section 4, we analyze histogram binning and show that any scoring function can be 'binned' to achieve distribution-free calibration. We explicitly quantify the finite-sample approximate calibration guarantees that automatically also lead to asymptotic calibration. We also discuss calibration in the online setting and calibration under covariate shift.

4 Achieving distribution-free approximate calibration

In Section 4.1, we prove a distribution-free approximate calibration guarantee given a fixed partitioning of the feature space into finitely many sets. This calibration guarantee also leads to asymptotic calibration. In Section 4.2, we discuss a natural method for obtaining such a partition using sample-splitting, called histogram binning. Histogram binning inherits the bound in Section 4.1. This shows that binning schemes lead to distribution-free approximate calibration. In Section 4.3 and 4.4 we discuss extensions of this scheme to adaptive sampling and covariate shift respectively.

4.1 Distribution-free calibration given a fixed sample-space partition

Suppose we have a fixed partition of $\mathcal{X}$ into B regions $\{\mathcal{X}_b\}_{b\in[B]}$, and let $\pi_b = \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]$ be the expected label probability in region $\mathcal{X}_b$. Denote the partition-identity function as $\mathcal{B} : \mathcal{X} \to [B]$ where $\mathcal{B}(x) = b$ if and only if $x \in \mathcal{X}_b$. Given a calibration set $\{(X_i, Y_i)\}_{i\in[n]}$, let $N_b := |\{i \in [n] : \mathcal{B}(X_i) = b\}|$ be the number of points from the calibration set that belong to region $\mathcal{X}_b$. In this subsection, we assume that $N_b \geqslant 1$ (in Section 4.2 we show that the partition can be constructed to ensure that N_b is $\Omega(n/B)$ with high probability). Define

$$\widehat{\pi}_b := \frac{1}{N_b} \sum_{i:\mathcal{B}(X_i)=b} Y_i \qquad \text{and} \qquad \widehat{V}_b := \frac{1}{N_b} \sum_{i:\mathcal{B}(X_i)=b} (Y_i - \widehat{\pi}_b)^2 \tag{12}$$

as the empirical average and variance of the Y values in a partition. We now deploy an empirical Bernstein bound [2] to produce a confidence interval for π_b.

Theorem 4. *For any $\alpha \in (0,1)$, with probability at least $1-\alpha$,*

$$|\pi_b - \widehat{\pi}_b| \leqslant \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b}, \quad \textit{simultaneously for all } b \in [B].$$

The theorem is proved in Appendix C. Using the crude deterministic bound $\widehat{V}_b \leqslant 1$ we get that the length of the confidence interval for partition b is $O(1/\sqrt{N_b})$. However, if for some b, $\mathcal{X}_b$ is highly informative or homogeneous in the sense that π_b is close to 0 or 1, we expect $\widehat{V}_b \ll 1$. In this case, Theorem 4 *adapts* and provides an $O(1/N_b)$ length interval for π_b. Let $b^\star = \arg\min_{b\in[B]} N_b$ denote the index of the region with the minimum number of calibration examples.

Corollary 4. *For $\alpha \in (0,1)$, the function $f_n(\cdot) := \widehat{\pi}_{\mathcal{B}(\cdot)}$ is (ε, α)-approximately calibrated with*

$$\varepsilon = \sqrt{\frac{2\widehat{V}_{b^\star} \ln(3B/\alpha)}{N_{b^\star}}} + \frac{3\ln(3B/\alpha)}{N_{b^\star}}.$$

Thus, $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated at level α.

The proof is in Appendix C. Thus, any finite partition of $\mathcal{X}$ leads to asymptotic calibration. However, the finite sample guarantee of Corollary 4 can be unsatisfactory if the sample-space partition is chosen poorly, since it might lead to small $N_{b^\star}$. In Section 4.2, we present a data-dependent partitioning scheme that provably guarantees that $N_{b^\star}$ scales as $\Omega(n/B)$ with high probability. The calibration guarantee of Corollary 4 can also be stated conditional on a given test point $X_{n+1} = x$:

$$|\mathbb{E}\left[Y \mid f(X) = f(x)\right] - f(x)| \leqslant \varepsilon, \text{ almost surely } P_X\ . \tag{13}$$

This holds since Theorem 4 provides simultaneously valid CIs for all regions $\mathcal{X}_b$.

4.2 Identifying a data-dependent partition using sample splitting

Here, we describe ways of constructing the partition $\{\mathcal{X}_b\}_{b\in[B]}$ through histogram binning [51], or simply, binning. Binning uses a sample splitting strategy to learn the partition of $\mathcal{X}$ as described in Section 4.1. A split of the data is used to learn the partition and an independent split is used

to estimate $\{\widehat{\pi}_b\}_{b\in[B]}$. Formally, the labeled data is split at random into a training set $\mathcal{D}_{\text{tr}}$ and a calibration set $\mathcal{D}_{\text{cal}}$. Then $\mathcal{D}_{\text{tr}}$ is used to train a scoring function $g : \mathcal{X} \to [0, 1]$ (in general the range of g could be any interval of $\mathbb{R}$ but for simplicity we describe it for $[0, 1]$). The scoring function g usually does not satisfy a calibration guarantee out-of-the-box but can be calibrated using binning.

A *binning scheme* $\mathcal{B}$ is any partition of $[0, 1]$ into B non-overlapping intervals $I_1, \ldots, I_B$, such that $\bigcup_{b\in[B]} I_b = [0, 1]$ and $I_b \cap I_{b'} = \varnothing$ for $b \neq b'$. $\mathcal{B}$ and g induce a partition of $\mathcal{X}$ as follows:

$$\mathcal{X}_b = \{x \in \mathcal{X} : \ g(x) \in I_b\}\,, \ b \in [B]. \tag{14}$$

The simplest binning scheme corresponds to *fixed-width binning*. In this case, bins have the form

$$I_i = \left[\frac{i-1}{B}, \frac{i}{B}\right), i = 1, \ldots, B-1 \text{ and } I_B = \left[\frac{B-1}{B}, 1\right].$$

However, fixed-width binning suffers from the drawback that there may exist bins with very few calibration points (low N_b), while other bins may get many calibration points. For bins with low N_b, the $\widehat{\pi}_b$ estimates cannot be guaranteed to be well calibrated, since the bound of Theorem 4 could be large. To remedy this, we consider *uniform-mass binning*, which aims to guarantee that each region $\mathcal{X}_b$ contains approximately equal number of calibration points. This is done by estimating the empirical quantiles of $g(X)$. First, the calibration set $\mathcal{D}_{\text{cal}}$ is randomly split into two parts, $\mathcal{D}^1_{\text{cal}}$ and $\mathcal{D}^2_{\text{cal}}$. For $j \in [B-1]$, the (j/B)-th quantile of $g(X)$ is estimated from $\{g(X_i), i \in \mathcal{D}^1_{\text{cal}}\}$. Let us denote the empirical quantile estimates as $\widehat{q}_j$. Then, the bins are defined as:

$$I_1 = [0, \widehat{q}_1)\,, I_i = [\widehat{q}_{i-1}, \widehat{q}_i]\,, i = 2, \ldots, B-1 \text{ and } I_B = (\widehat{q}_{B-1}, 1]\,.$$

This induces a partition of $\mathcal{X}$ as per (14). Now, only $\mathcal{D}^2_{\text{cal}}$ is used for calibrating the underlying classifier, as per the calibration scheme defined in Section 4.1. Kumar et al. [21] showed that uniform-mass binning provably controls the number of calibration samples that fall into each bin (see Appendix F.2). Building on their result and Corollary 4, we show the following guarantee.

Theorem 5. *Fix $g : \mathcal{X} \to [0, 1]$ and $\alpha \in (0, 1)$. There exists a universal constant c such that if $\left|\mathcal{D}^1_{cal}\right| \geqslant cB \ln(2B/\alpha)$, then with probability at least $1 - \alpha$,*

$$N_{b^\star} \geqslant \left|\mathcal{D}^2_{cal}\right|/2B - \sqrt{|\mathcal{D}^2_{cal}| \ln(2B/\alpha)/2}.$$

Thus even if $|\mathcal{D}^1_{cal}|$ does not grow with n, as long as $|\mathcal{D}^2_{cal}| = \Omega(n)$, uniform-mass binning is $(\widetilde{O}(\sqrt{B \ln(1/\alpha)/n}), \alpha)$-approximately calibrated, and hence asymptotically calibrated for any α.

The proof is in Appendix C. In words, if we use a small number of points (independent of n) for uniform-mass binning, and the rest to estimate bin probabilities, we achieve approximate/asymptotic distribution-free calibration. Note that the probability is conditional on a fixed predictor g, and hence also conditional on the training data $\mathcal{D}_{\text{tr}}$. Since Theorem 5 uses Corollary 4, the calibration guarantee can also be stated conditionally on a fixed test point, akin to equation (13).

4.3 Distribution-free calibration in the online setting

So far, we have considered the batch setting with a fixed calibration set of size n. However, often a practitioner might want to query additional calibration data until a desired confidence level is achieved. This is called the *online* or *adaptive* setting. In this case, the results of Section 4 are no longer valid since the number of calibration samples is unknown a priori and may even be dependent on the data. In order to quantify uncertainty in the online setting, we use *time-uniform* concentration bounds [14, 15]; these hold simultaneously for all possible values of the calibration set size $n \in \mathbb{N}$.

Fix a partition of $\mathcal{X}$, $\{\mathcal{X}_b\}_{b\in[B]}$. For some value of n, let the calibration data be given as $\mathcal{D}^{(n)}_{\text{cal}}$. We use the superscript notation to emphasize the dependence on the current size of the calibration set. Let $\{(X^b_i, Y^b_i)\}_{i\in[N^{(n)}_b]}$ be examples from the calibration set that fall into the partition $\mathcal{X}_b$, where $N^{(n)}_b := |\{i \in [n] : \mathcal{B}(X_i) = b\}|$ is the total number of points that are mapped to $\mathcal{X}_b$. Let the empirical label average and cumulative (unnormalized) empirical variance be denoted as

$$\widehat{V}^+_b = 1 \vee \sum_{i=1}^{N^{(n)}_b} \left(Y^b_i - \overline{Y}^b_{i-1}\right)^2, \text{ where } \overline{Y}^b_i := \frac{1}{i}\sum_{j=1}^{i} Y^b_j \text{ for } i \in [N^{(n)}_b]. \tag{15}$$

Note the normalization difference between $\widehat{V}^+_b$ and $\widehat{V}^b$ used in the batch setting. The following theorem constructs confidence intervals for $\{\pi_b\}_{b\in[B]}$ that are valid uniformly for any value of n.

Theorem 6. *For any $\alpha \in (0,1)$, with probability at least $1-\alpha$,*

$$|\pi_b - \widehat{\pi}_b| \leqslant \frac{7\sqrt{\widehat{V}_b^+ \ln\left(1+\ln \widehat{V}_b^+\right)} + 5.3\ln\left(\frac{6.3B}{\alpha}\right)}{N_b^{(n)}}, \quad \textit{simultaneously for all } b \in [B] \textit{ and all } n \in \mathbb{N}. \tag{16}$$

Thus $\widehat{\pi}_b$ is asymptotically calibrated at any level $\alpha \in (0,1)$.

The proof is in Appendix C. Due to the crude bound: $\widehat{V}_b^+ \leqslant N_b^{(n)}$, we can see that the width of confidence intervals roughly scales as $O(\sqrt{\ln(1+\ln N_b^{(n)})/N_b^{(n)}})$. In comparison to the batch setting, only a small price is paid for not knowing beforehand how many examples will be used for calibration.

4.4 Calibration under covariate shift

Here, we briefly consider the problem of calibration under covariate shift [41]. In this setting, calibration data $\{(X_i, Y_i)\}_{i\in[n]} \sim P^n$ is from a 'source' distribution P, while the test point is from a shifted 'target' distribution $(X_{n+1}, Y_{n+1}) \sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}$, meaning that the 'shift' occurs only in the covariate distribution while $P_{Y|X}$ does not change. We assume the likelihood ratio (LR)

$$w : \mathcal{X} \to \mathbb{R}; \quad w(x) := \mathrm{d}\widetilde{P}_X(x)/\mathrm{d}P_X(x)$$

is well-defined. The following is unambiguous: *if w is arbitrarily ill-behaved and unknown, the covariate shift problem is hopeless, and one should not expect any distribution-free guarantees*. Nevertheless, one can still make nontrivial claims using a 'modular' approach towards assumptions:

> Condition (A): $w(x)$ is known exactly and is bounded.
> Condition (B): an asymptotically consistent estimator $\widehat{w}(x)$ for $w(x)$ can be constructed.

We show the following: under Condition (A), a weighted estimator using w delivers approximate and asymptotic distribution-free calibration; under Condition (B), weighting with a plug-in estimator for w continues to deliver asymptotic distribution-free calibration. It is clear that Condition (B) will always require distributional assumptions: asymptotic consistency is nontrivial for ill-behaved w. Nevertheless, the above two-step approach makes it clear where the burden of assumptions lie: not with calibration step, but with the w estimation step. Estimation of w is a well studied problem in the covariate-shift literature and there is some understanding of what assumptions are needed to accomplish it, but there has been less work on recognizing the resulting implications for calibration. Luckily, many practical methods exist for estimating w given unlabeled samples from $\widetilde{P}_X$ [4, 16, 17]. In summary, if Condition (B) is possible, then distribution-free calibration is realizable, and if Condition (B) is not met (even with infinite samples), then it implies that w is probably very ill-behaved, and so distribution-free calibration is also likely to be impossible.

For a fixed partition $\{\mathcal{X}_b\}_{b\in[B]}$, one can use the labeled data from the source distribution to estimate $\mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b]$ (unlike $\mathbb{E}_P[Y \mid X \in \mathcal{X}_b]$ as before), given oracle access to w:

$$\breve{\pi}_b^{(w)} := \frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}. \tag{17}$$

As preluded to earlier, assume that

$$\text{for all } x \in \mathcal{X},\ L \leqslant w(x) \leqslant U \text{ for some } 0 < L \leqslant 1 \leqslant U < \infty. \tag{18}$$

The 'standard' i.i.d. assumption on the test point equivalently assumes w is known and $L = U = 1$. We now present our first claim: $\breve{\pi}_b^{(w)}$ satisfies a distribution-free approximate calibration guarantee. To show the result, we assume that the sample-space partition was constructed via uniform-mass binning (on the source domain) with sufficiently many points, as required by Theorem 5. This guarantees that all regions satisfy $|\{i : \mathcal{B}(X_i) = b\}| = \Omega(n/B)$ with high probability.

Theorem 7. *Assume w is known and bounded* (18)*. Then for an explicit universal constant $c > 0$, with probability at least $1-\alpha$,*

$$\left|\breve{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b]\right| \leqslant c\left(\frac{U}{L}\right)^2 \sqrt{\frac{B\ln(6B/\alpha)}{2n}}, \quad \textit{simultaneously for all } b \in [B],$$

as long as $n \geqslant c(U/L)^2 B \ln^2(6B/\alpha)$. Thus $\breve{\pi}_b^{(w)}$ is asymptotically calibrated at any $\alpha \in (0,1)$.

The proof is in Appendix D. Theorem 7 establishes distribution-free calibration under Condition (A). For Condition (B), using k *unlabeled* samples from the source and target domains, assume that we construct an estimator $\widehat{w}_k$ of w that is consistent, meaning

$$\sup_{x \in \mathcal{X}} |\widehat{w}_k(x) - w(x)| \xrightarrow{P} 0. \tag{19}$$

We now define an estimator $\breve{\pi}_b^{(\widehat{w}_k)}$ by plugging in $\widehat{w}_k$ for w in the right hand side of (17):

$$\breve{\pi}_b^{(\widehat{w}_k)} := \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)}.$$

Proposition 2. *If $\widehat{w}_k$ is consistent* (19)*, then $\breve{\pi}_b^{(\widehat{w}_k)}$ is asymptotically calibrated at any $\alpha \in (0, 1)$.*

In Appendix D, we illustrate through preliminary simulations that w can be estimated using unlabeled data from the target distribution, and consequently approximate calibration can be achieved on the target domain. Recently, Park et al. [35] also considered calibration under covariate shift through importance weighting, but they do not show validity guarantees in the same sense as Theorem 7. For real-valued regression, distribution-free prediction sets under covariate shift were constructed using conformal prediction [42] under Condition (A), and is thus a precursor to our modular approach.

5 Other related work

The problem of assessing the calibration of binary classifiers was first studied in the meteorological and statistics literature [5–7, 9, 10, 28–31, 39]; we refer the reader to the review by Dawid [8] for more details. These works resulted in two common ways of measuring calibration: reliability diagrams [9] and estimates of the squared expected calibration error (ECE) [39]: $\mathbb{E}(f(X) - \mathbb{E}[Y \mid f(X)])^2$. Squared ECE can easily be generalized to multiclass settings and some related notions such as absolute deviation ECE and top-label ECE have also been considered, for instance [12, 32]. ECE is typically estimated through binning, which provably leads to underestimation of ECE for calibrators with continuous output [21, 44]. Certain methods have been proposed to estimate ECE without binning [50, 53], but they require distributional assumptions for provability.

While these papers have focused on the difficulty of *estimating* calibration error, ours is the first formal impossibility result for *achieving* calibration. In particular, Kumar et al. [21, Theorem 4.1] show that the scaling-binning procedure achieves calibration error close to the best within a fixed, regular, injective parametric class. However, as discussed in Section 3.3 (after Theorem 3), we show that the best predictor in an injective parametric class itself cannot have a distribution-free guarantee. In summary, our results show not only that (some form of) binning is necessary for distribution-free calibration (Theorem 3), but also sufficient (Corollary 4).

Apart from classical methods for calibration [33, 36, 51, 52], some new methods have been proposed recently, primarily for calibration of deep neural networks [12, 18, 19, 22, 23, 27, 40, 43, 49]. These calibration methods perform well in practice but do not have distribution-free guarantees. A calibration framework that generalizes binning and isotonic regression is Venn prediction [24, 45–48]; we briefly discuss this framework and show some connections to our work in Appendix E.

Calibration has natural applications in numerous sensitive domains where uncertainty estimation is desirable (healthcare, finance, forecasting). Recently, calibrated classifiers have been used as a part of the pipeline for anomaly detection [13, 25] and label shift estimation [1, 11, 38].

6 Conclusion

We analyze calibration for binary classification problems from the standpoint of robustness to distributional assumptions. By connecting calibration to other ways of quantifying uncertainty, we establish that popular parametric scaling methods cannot provide provable informative calibration guarantees in the distribution-free setting. In contrast, we showed that a standard nonparametric method — histogram binning — satisfies approximate and asymptotic calibration guarantees without distributional assumptions. We also establish guarantees for the cases of streaming data and covariate shift.

Takeaway message. Recent calibration methods that perform binning on top of parametric methods (Platt-binning [21] and IROvA-TS [53]) have achieved strong empirical performance. In light of the theoretical findings in our paper, we recommend some form of binning as the last step of calibrated prediction due to the robust distribution-free guarantees provided by Theorem 4.

7 Broader Impact

Machine learning is regularly deployed in real-world settings, including areas having high impact on individual lives such as granting of loans, pricing of insurance and diagnosis of medical conditions. Often, instead of hard $0/1$ classifications, these systems are required to produce soft probabilistic predictions, for example of the probability that a startup may go bankrupt in the next few years (in order to determine whether to give it a loan) or the probability that a person will recover from a disease (in order to price an insurance product). Unfortunately, even though classifiers produce numbers between 0 and 1, these are well known to not be 'calibrated' and hence not be interpreted as probabilities in any real sense, and using them in lieu of probabilities can be both misleading (to the bank granting the loan) and unfair (to the individual at the receiving end of the decision).

Thus, following early research in meteorology and statistics, in the last couple of decades the ML community has embraced the formal goal of calibration as a way to quantify uncertainty as well as to interpret classifier outputs. However, there exist other alternatives to quantify uncertainty, such as confidence intervals for the regression function and prediction sets for the binary label. There is not much guidance on which of these should be employed in practice, and what the relationship between them is, if any. Further, while there are many post-hoc calibration techniques, it is unclear which of these require distributional assumptions to work and which do not—this is critical because making distributional assumptions (for convenience) on financial or medical data is highly suspect.

This paper explicitly relates the three aforementioned notions of uncertainty quantification without making distributional assumptions, describes what is possible and what is not. Importantly, by providing distribution-free guarantees on well-known variants of binning, we identify a conceptually simple and theoretically rigorous way to ensure calibration in high-risk real-world settings. Our tools are thus likely to lead to fairer systems, better estimates of risks of high-stakes decisions, and more human-interpretable outputs of classifiers that apply out-of-the-box in many real-world settings because of the assumption-free guarantees.

Acknowledgements

The authors would like to thank Tudor Manole, Charvi Rastogi, Michael Cooper Stanley, and the anonymous Neurips 2020 reviewers for comments on an initial version of this paper.

References

[1] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Adapting to label shift with bias-corrected calibration. In *International Conference on Machine Learning*, 2020.

[2] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *International Conference on Algorithmic Learning Theory*, 2007.

[3] Rina Foygel Barber. Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14(2):3487–3524, 2020.

[4] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *International Conference on Machine Learning*, 2007.

[5] Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[6] Jochen Bröcker. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate dynamics*, 39(3-4):655–667, 2012.

[7] A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

[8] A Philip Dawid. Probability forecasting. *Wiley StatsRef: Statistics Reference Online*, 2014.

[9] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

[10] Christopher AT Ferro and Thomas E Fricker. A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, 138(668):1954–1960, 2012.

[11] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, 2020.

[12] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

[13] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

[14] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[15] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, non-asymptotic confidence sequences. *The Annals of Statistics (to appear)*, 2020+.

[16] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J. Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, 2007.

[17] Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

[18] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 2017.

[19] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, 2018.

[20] Meelis Kull, Telmo M. Silva Filho, and Peter Flach. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11(2):5052–5080, 2017.

[21] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 2019.

[22] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, 2018.

[23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

[24] Antonis Lambrou, Ilia Nouretdinov, and Harris Papadopoulos. Inductive Venn prediction. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):181–201, 2015.

[25] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

[26] Jing Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.

[27] Dimitrios Milios, Raffaello Camoriano, Pietro Michiardi, Lorenzo Rosasco, and Maurizio Filippone. Dirichlet-based gaussian processes for large-scale calibrated classification. In *Advances in Neural Information Processing Systems*, 2018.

[28] Allan H Murphy. Scalar and vector partitions of the probability score: Part i. two-state situation. *Journal of Applied Meteorology*, 11(2):273–282, 1972.

[29] Allan H Murphy. Scalar and vector partitions of the probability score: Part ii. n-state situation. *Journal of Applied Meteorology*, 11(8):1183–1192, 1972.

[30] Allan H Murphy. A new vector partition of the probability score. *Journal of applied Meteorology*, 12(4):595–600, 1973.

[31] Allan H Murphy and Edward S Epstein. Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology*, 6(5):748–755, 1967.

[32] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

[33] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, 2005.

[34] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, 2002.

[35] Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[36] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

[37] Kostas Proedrou, Ilia Nouretdinov, Vladimir Vovk, and Alex Gammerman. Transductive confidence machines for pattern recognition. In *European Conference on Machine Learning*, 2002.

[38] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation*, 14(1):21–41, 2002.

[39] Frederick Sanders. On subjective probability forecasting. *Journal of Applied Meteorology*, 2 (2):191–201, 1963.

[40] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[41] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[42] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 2019.

[43] Gia-Lac Tran, Edwin V Bonilla, John Cunningham, Pietro Michiardi, and Maurizio Filippone. Calibrating deep convolutional gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[44] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B Schön. Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[45] Vladimir Vovk and Ivan Petej. Venn-Abers predictors. In *Conference on Uncertainty in Artificial Intelligence*, 2014.

[46] Vladimir Vovk, Glenn Shafer, and Ilia Nouretdinov. Self-calibrating probability forecasting. In *Advances in Neural Information Processing Systems*, 2004.

[47] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. 2005.

[48] Vladimir Vovk, Ivan Petej, and Valentina Fedorova. Large-scale probabilistic predictors with and without guarantees of validity. In *Advances in Neural Information Processing Systems*, 2015.

[49] Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[50] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: a unifying framework. In *Advances in Neural Information Processing Systems*, 2019.

[51] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *International Conference on Machine Learning*, 2001.

[52] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *International Conference on Knowledge Discovery and Data Mining*, 2002.

[53] Jize Zhang, Bhavya Kailkhura, and T Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, 2020.

Appendix

The Appendix contains proofs of results in the main paper ordered as they appear. Auxiliary results needed for some of the proofs are stated in Appendix F.

A Proof of Proposition 1

The 'if' part of the theorem is due to Vaicenavicius et al. [44, Proposition 1]; we reproduce it for completeness. Let $\sigma(g), \sigma(f)$ be the sub σ-algebras generated by g and f respectively. By definition of f, we know that f is $\sigma(g)$-measurable and, hence, $\sigma(f) \subseteq \sigma(g)$. We now have:

$$\begin{aligned}\mathbb{E}\left[Y \mid f(X)\right] &= \mathbb{E}\left[\mathbb{E}\left[Y \mid g(X)\right] \mid f(X)\right] && \text{(by tower rule since } \sigma(f) \subseteq \sigma(g)\text{)} \\ &= \mathbb{E}\left[f(X) \mid f(X)\right] && \text{(by property (5))} \\ &= f(X).\end{aligned}$$

The 'only if' part can be verified for $g = f$. Since f is perfectly calibrated,

$$\mathbb{E}\left[Y \mid f(X) = f(x)\right] = f(x),$$

almost surely P_X.

□

B Proofs of results in Section 3

B.1 Proof of Theorem 1

Assume that one is given a predictor f that is (ε, α)-approximately calibrated. Then the assertion follows from the definition of (ε, α)-approximate calibration since:

$$\left|\mathbb{E}\left[Y \mid f(X)\right] - f(X)\right| \leqslant \varepsilon(f(X)) \implies \mathbb{E}\left[Y \mid f(X)\right] \in C(f(X)).$$

Now we show the proof in the other direction. If m_C was injective, $\mathbb{E}\left[Y \mid m_C(f(X))\right] = \mathbb{E}\left[Y \mid f(X)\right]$ and thus if $\mathbb{E}\left[Y \mid f(X)\right] \in C(f(X))$ (which happens with probability at least $1-\alpha$), we would have $\mathbb{E}\left[Y \mid m_C(f(X))\right] \in C(f(X))$ and so

$$\left|\mathbb{E}\left[Y \mid m_C(f(X))\right] - m_C(f(X)\right| \leqslant \sup_{z \in \text{Range}(f)} \{|C(z)|/2\} = \varepsilon.$$

This serves as an intuition for the proof in the general case, when m_C need not be injective. Note that,

$$\begin{aligned}\left|\mathbb{E}\left[Y \mid m_C(f(X))\right] - m_C(f(X))\right| &= \left|\mathbb{E}\left[Y \mid m_C(f(X))\right] - \mathbb{E}\left[m_C(f(X)) \mid m_C(f(X))\right]\right| \\ &\overset{(1)}{=} \left|\mathbb{E}\left[\mathbb{E}\left[Y \mid f(X)\right] \mid m_C(f(X))\right] - \mathbb{E}\left[m_C(f(X)) \mid m_C(f(X))\right]\right| \\ &\overset{(2)}{=} \left|\mathbb{E}\left[\mathbb{E}\left[Y \mid f(X)\right] - m_C(f(X)) \mid m_C(f(X))\right]\right| \\ &\overset{(3)}{\leqslant} \mathbb{E}\left[\left|\mathbb{E}\left[Y \mid f(X)\right] - m_C(f(X))\right| \mid m_C(f(X))\right], \end{aligned} \tag{20}$$

where we use the tower rule in (1) (since m_C is a function of f), linearity of expectation in (2) and Jensen's inequality in (3). To be clear, the outermost expectation above is over $f(X)$ (conditioned on $m_C(f(X))$). Consider the event

$$A : \mathbb{E}\left[Y \mid f(X)\right] \in C(f(X)).$$

On A, by definition we have:

$$\left|\mathbb{E}\left[Y \mid f(X)\right] - m_C(f(X))\right| = \frac{u_C(f(X)) - l_C(f(X))}{2} \leqslant \sup_{z \in \text{Range}(f)} \left(\frac{|C(z)|}{2}\right) = \varepsilon.$$

By monotonicity property of conditional expectation, we also have that conditioned on A,

$$\mathbb{E}\left[\left|\mathbb{E}\left[Y \mid f(X)\right] - m_C(f(X))\right| \mid m_C(f(X))\right] \leqslant \mathbb{E}\left[\varepsilon \mid m_C(f(X))\right] = \varepsilon,$$

with probability 1. Thus by the relationship proved in the series of equations ending in (20), we have that conditioned on A, with probability 1,

$$|\mathbb{E}\left[Y \mid m_C(f(X))\right] - m_C(f(X))| \leqslant \varepsilon.$$

Since we are given that C is a $(1-\alpha)$-CI with respect to f, $\mathbb{P}(A) \geqslant 1-\alpha$. For any event B, it holds that $\mathbb{P}(B) \geqslant \mathbb{P}(B|A)\,\mathbb{P}(A)$. Setting

$$B : |\mathbb{E}\left[Y \mid m_C(f(X))\right] - m_C(f(X))| \leqslant \varepsilon,$$

we obtain:

$$\mathbb{P}\left(|\mathbb{E}\left[Y \mid m_C(f(X))\right] - m_C(f(X))| \leqslant \varepsilon\right) \geqslant 1-\alpha.$$

Thus, we conclude that $m_C(f(\cdot))$ is (ε, α)-approximately calibrated. □

B.2 Proof of Corollary 1

Let $\{f_n\}_{n\in\mathbb{N}}$ be asymptotically calibrated sequence with the corresponding sequence of functions $\{\varepsilon_n\}_{n\in\mathbb{N}}$ that satisfy $\varepsilon_n(f_n(X_{n+1})) = o_P(1)$. From Theorem 1, we can construct corresponding functions C_n that are $(1-\alpha)$-CI with respect to f_n and satisfy

$$|C_n(f_n(X_{n+1}))| = 2\varepsilon_n(f_n(X_{n+1})) = o_P(1).$$

This concludes the proof. □

B.3 Proof of Theorem 2

In the proof we write the test point as (X_{n+1}, Y_{n+1}). Suppose $\widehat{C}_n$ is a $(1-\alpha)$-CI with respect to f for all distributions P. We show that $\widehat{C}_n$ covers the label Y_{n+1} itself for distributions P such that $P_{f(X)}$ is nonatomic (and thus $\mathrm{disc}(\widehat{C}_n)$ would also cover the labels).

Let P be any distribution such that $P_{f(X)}$ is nonatomic. Fix a set of $m \geqslant n+1$ samples from the distribution P denoted as $\mathcal{T} = \{(A^{(j)}, B^{(j)})\}_{j\in[m]}$. Given $\mathcal{T}$, consider a distribution Q corresponding to the following sampling procedure for $(X, Y) \sim Q$:

> sample an index j uniformly at random from $[m]$ and set $(X, Y) = (A^{(j)}, B^{(j)})$.

The distribution function for Q is given by

$$m^{-1}\sum_{j=1}^{m} \delta_{(A^{(j)}, B^{(j)})}.$$

where $\delta_{(a,b)}$ denotes the points mass at (a, b). Note that Q is only defined conditional on $\mathcal{T}$. Observe the following facts about Q:

- $\mathrm{supp}(Q) = \{(A^{(j)}, B^{(j)})\}_{j\in[m]}$.
- Consider any $(x, y) \in \mathrm{supp}(Q)$. Let $(x, y) = (A^{(j)}, B^{(j)})$ for some $j \in [m]$. Then

$$\begin{aligned}\mathbb{E}_Q\left[Y \mid f(X) = f(x)\right] &= \mathbb{E}_Q\left[Y \mid f(X) = f(A^{(j)})\right] \\ &\overset{\xi_1}{=} \mathbb{E}_Q\left[Y \mid X = A^{(j)}\right] \\ &\overset{\xi_2}{=} B^{(j)} = y.\end{aligned}$$

  Above ξ_1 holds since $P_{f(X)}$ is nonatomic so that the $f(X^{(i)})$'s are unique almost surely. Note that $P_{f(X)}$ is nonatomic only if P_X itself is nonatomic. Thus the $A^{(j)}$'s are unique almost surely, and ξ_2 follow. In other words, if $(X, Y) \sim Q$, then we have

$$Y = \mathbb{E}_Q\left[Y \mid f(X)\right]. \tag{21}$$

Suppose the data distribution was Q, that is $\{(X_i, Y_i)\}_{i\in[n+1]} \sim Q^{n+1}$. Define the event that the CI guarantee holds as

$$E_1 : \mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right] \in \widehat{C}_n(f(X_{n+1})), \tag{22}$$

and the event that the PS guarantee holds as

$$E_2 : Y_{n+1} \in \widehat{C}_n(f(X_{n+1})). \tag{23}$$

Then due to (21), the events are exactly the same under Q:

$$E_1 \overset{Q}{\equiv} E_2. \tag{24}$$

In particular, this means

$$\mathbb{P}_{Q^{n+1}}(\mathbb{E}_Q\left[Y_{n+1} \mid f(X_{n+1})\right] \in \widehat{C}_n(f(X_{n+1}))) = \mathbb{P}_{Q^{n+1}}(Y_{n+1} \in \widehat{C}_n(f(X_{n+1}))). \tag{25}$$

If $\widehat{C}_n$ is a distribution-free CI, then $\mathbb{P}_{Q^{n+1}}(E_1) \geqslant 1-\alpha$ and thus $\mathbb{P}_{Q^{n+1}}(E_2) \geqslant 1-\alpha$. This shows that for Q, $\text{disc}(\widehat{C}_n)$ is a $(1-\alpha)$-PI. Note that Q corresponds to sampling *with replacement* from a fixed set $\mathcal{T}$ where each element is drawn with respect to P. Although $Q \neq P$, we expect that as $m \to \infty$ (while n is fixed), Q and P coincide. This would prove the result for general P. To formalize this intuition, we describe a distribution which is close to Q but corresponds to sampling *without replacement* from $\mathcal{T}$ instead.

For this, now suppose that $\{(X_i, Y_i)\}_{i\in[n+1]} \sim R^{n+1}$ where R^{n+1} corresponds to sampling without replacement from $\mathcal{T}$. Formally, to draw from R^{n+1}, we first draw a surjective mapping $\lambda : [n+1] \to [m]$ as

$$\lambda \sim \text{Unif } (n\text{-sized ordered subsets of } [m]),$$

and set $(X_i, Y_i) = (A^{(\lambda(i))}, B^{(\lambda(i))})$ for $i \in [n+1]$.

First we quantify precisely the intuition that as $m \to \infty$, Q^{n+1} and R^{n+1} are essentially identical. Consider the event $T :=$ no index is repeated in Q^{n+1}. Let $\mathbb{P}(T) = \tau_m$ for some m and note that $\lim_{m\to\infty} \tau_m = 1$. Now consider any probability event E over $\{(X_i, Y_i)\}_{i\in[n+1]}$ (such as E_1 or E_2). We have

$$\begin{aligned}\mathbb{P}_{Q^{n+1}}(E) &= \mathbb{P}_{Q^{n+1}}(E|T) \cdot \mathbb{P}(T) + \mathbb{P}_{Q^{n+1}}(E|T^c) \cdot \mathbb{P}(T^c) \\ &\in [\mathbb{P}_{Q^{n+1}}(E|T) \cdot \mathbb{P}(T), \mathbb{P}_{Q^{n+1}}(E|T) \cdot \mathbb{P}(T) + \mathbb{P}(T^c)].\end{aligned}$$

Now observe that $\mathbb{P}_{Q^{n+1}}(E|T) = \mathbb{P}_{R^{n+1}}(E)$ to conclude

$$\mathbb{P}_{Q^{n+1}}(E) \in [\mathbb{P}_{R^{n+1}}(E) \cdot \mathbb{P}(T), \mathbb{P}_{R^{n+1}}(E) \cdot \mathbb{P}(T) + \mathbb{P}(T^c)].$$

Since $m \geqslant n+1$, $\mathbb{P}(T) \neq 0$ so we can invert the above and substitute $\tau_m = \mathbb{P}(T)$ to get

$$\mathbb{P}_{R^{n+1}}(E) \in \left[\tau_m^{-1}(\mathbb{P}_{Q^{n+1}}(E) - (1-\tau_m)),\ \tau_m^{-1}\mathbb{P}_{Q^{n+1}}(E)\right]. \tag{26}$$

Consider $E = E_2$ defined in equation (23). We showed that $\mathbb{P}_{Q^{n+1}}(E_2) \geqslant 1-\alpha$. Thus from (26),

$$\mathbb{P}_{R^{n+1}}(E_2) \geqslant \tau_m^{-1}(1-\alpha-(1-\tau_m)).$$

The above is with respect to R^{n+1} which is conditional on a fixed draw $\mathcal{T}$. However since the right hand side is independent of $\mathcal{T}$, we can also include the randomness in $\mathcal{T}$ to say:

$$\mathbb{P}_{R^{n+1},\mathcal{T}}(E_2) \geqslant \tau_m^{-1}(1-\alpha-(1-\tau_m)). \tag{27}$$

Observe that if we consider the marginal distribution over R^{n+1} and $\mathcal{T}$ (that is we include the randomness in $\mathcal{T}$ as above), $\{(X_i, Y_i)\}_{i\in[n+1]} \overset{iid}{\sim} P$. This is not true if we do not marginalize over $\mathcal{T}$, in particular since the (X_i, Y_i)'s are not independent (due to sampling without replacement). Thus equation (27) can be restated as

$$\mathbb{P}_{P^{n+1}}(E_2) \geqslant \tau_m^{-1}(1-\alpha-(1-\tau_m)),$$

Since m can be set to any number and $\lim_{m\to\infty} \tau_m = 1$, we can indeed conclude

$$\mathbb{P}_{P^{n+1}}(E_2) \geqslant 1-\alpha.$$

Recall that E_2 is the event that $Y_{n+1} \in \widehat{C}_n(X_{n+1})$; equivalently $Y_{n+1} \in \text{disc}\widehat{C}_n(X_{n+1})$. Thus $\text{disc}(\widehat{C}_n)$ provides a $(1-\alpha)$-PI for P such that $P_{f(X)}$ is nonatomic. □

B.4 Proof of Corollary 2

Let P be any distribution such that $P_{f(X)}$ is nonatomic. By Theorem 2, $\widehat{C}_n$ must provide both a prediction set and a confidence interval for P:

$$\mathbb{P}(\mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right] \in \widehat{C}_n(f(X_{n+1}))) \geqslant 1-\alpha,$$

and

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

Thus by a union bound

$$\mathbb{P}_{P^{n+1}}(\{Y_{n+1}, \mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right]\} \subseteq \widehat{C}_n(f(X_{n+1}))) \geqslant 1-2\alpha. \tag{28}$$

Now consider a distribution P such that $P_{f(X)}$ is nonatomic and $\mathbb{P}(Y=1 \mid X) = 0.5$ a.s. P_X so that $\mathbb{E}\left[Y_{n+1} \mid f(X)\right] = 0.5$ a.s. $P_{f(X)}$. The inequality (28) is true for this P as well. If

$$\{Y_{n+1}, \mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right]\} \subseteq \widehat{C}_n(f(X_{n+1})),$$

then $|\widehat{C}_n(X_{n+1})| \geqslant |Y_{n+1} - \mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right]| \geqslant 0.5$. Thus

$$\mathbb{P}_{P^{n+1}}(|\widehat{C}_n(f(X_{n+1}))| \geqslant 0.5) \geqslant 1-2\alpha.$$

Consequently we have

$$\begin{aligned}\mathbb{E}_{P^{n+1}}|\widehat{C}_n(f(X_{n+1}))| &\geqslant 0.5(1-2\alpha)\\ &= 0.5-\alpha.\end{aligned}$$

This concludes the proof. □

B.5 Proof of Theorem 3

Suppose that $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated and satisfies

$$\limsup_{n\to\infty} \left|\mathcal{X}^{(f_n)}\right| > \aleph_0,$$

that is, for every $m \in \mathbb{N}$, there exists $n \geqslant m$ such that $\mathcal{X}^{(f_n)}$ is an uncountable set. We will show a contradiction using Corollary 2 for f_n and a certain C_n to be defined shortly.

First, we verify the condition of Corollary 2 for f_n if $\mathcal{X}^{(f_n)}$ is uncountable: we construct a distribution P such that $P_{(f_n(X))}$ is nonatomic. Let the range of f_n acting on $\mathcal{X}$ be denoted as $f_n(\mathcal{X})$, and for $z \in f_n(\mathcal{X})$ let the level set at value z be denoted as $\mathcal{X}_z^{(f_n)}$. Since the sets $\mathcal{X}^{(f_n)}$ are measurable, we can define $P(X)$ as follows:

$$P(f_n(X)) = \text{Unif}(f_n(\mathcal{X})); \quad P(X \mid f_n(X)) = \text{Unif}\left(\mathcal{X}_{f_n(X)}^{(f_n)}\right). \tag{29}$$

$P(X)$ along with any conditional probability function $P(Y \mid X)$ constitutes a valid probability distribution P. Further, from the construction, since $\mathcal{X}^{(f_n)}$ is uncountable, $P_{f_n(X)}$ is guaranteed to be nonatomic.

Next, since $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated, by Corollary 1, one can construct a sequence of functions $\{C_n\}_{n\in\mathbb{N}}$ such that each C_n is a $(1-\alpha)$-CI with respect to f_n for any distribution Q, and

$$|C_n(f_n(X_{n+1}))| = o_Q(1).$$

Thus there exists a constant m such that for $n \geqslant m$ and any distribution Q,

$$\mathbb{E}_{Q^{n+1}}|C_n(f_n(X_{n+1}))| < 0.5-\alpha. \tag{30}$$

However, since $\limsup_{n\to\infty}|\mathcal{X}^{(f_n)}| > \aleph_0$, there exists an $n \geqslant m$ such that $\mathcal{X}^{(f_n)}$ is uncountable. Hence the requirements of Corollary 2 are satisfied by $\widehat{C}_n$ and f_n: namely $\widehat{C}_n$ is a $(1-\alpha)$-CI with respect to f for all distributions P, and there exists a P such that $P_{f_n(X)}$ is nonatomic. Thus Corollary 2 yields that we can construct a distribution Q such that

$$\mathbb{E}_{Q^{n+1}}|C_n(f_n(X_{n+1}))| \geqslant 0.5-\alpha,$$

which is a contradiction to (30). Hence our hypothesis that $\limsup_{n\to\infty}|\mathcal{X}^{(f_n)}| > \aleph_0$ must be false, concluding the proof. □

C Proofs of results in Section 4 (other than Section 4.4)

C.1 Proof of Theorem 4

Let $E_{\mathcal{B}(x)}$ be the event that $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. On the event $E_{\mathcal{B}(x)}$, within each region $\mathcal{X}_b$, the number of point from the calibration set is known and the Y_i's in each bin represent independent Bernoulli random variables that share the same mean $\pi_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$. Consider any fixed region $\mathcal{X}_b$, $b \in [B]$. Using Theorem 10, we obtain that:

$$\mathbb{P}\left(|\pi_b - \widehat{\pi}_b| > \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b} \;\middle|\; E_{\mathcal{B}(x)} \right) \leqslant \alpha/B.$$

Applying union bound across all regions of the sample-space partition, we get that:

$$\mathbb{P}\left(\forall b \in [B] :\ |\pi_b - \widehat{\pi}_b| \leqslant \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b} \;\middle|\; E_{\mathcal{B}(x)} \right) \geqslant 1 - \alpha.$$

Because this is true for any $E_{\mathcal{B}(x)}$, we can marginalize to obtain the assertion of the theorem in unconditional form. □

C.2 Proof of Corollary 4

We show a calibration guarantee by using Theorem 1. Consider the scoring function as $\mathcal{B}$ with $\mathcal{Z} = [B]$. Then by Theorem 4, $C : [B] \to \mathcal{I}$ given by

$$C(b) = \left[\widehat{\pi}_b - \left(\sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b} \right), \widehat{\pi}_b + \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3\ln(3B/\alpha)}{N_b} \right],\ b \in [B],$$

provides a $(1-\alpha)$-CI with respect to $\mathcal{B}$. Let $b^\star = \min_{b \in [B]} N_b$. To apply Theorem 4, we define

$$\varepsilon = \sup_{b \in [B]} |C(b)|/2 = \sqrt{\frac{2\widehat{V}_{b^\star} \ln(3B/\alpha)}{N_{b^\star}}} + \frac{3\ln(3B/\alpha)}{N_{b^\star}},$$

and the mid-point function m_C for C is given by $m_C(b) = \widehat{\pi}_b$. Applying Theorem 1 gives the first part of the result.

Next, suppose some bin b has $\mathbb{P}(\mathcal{B}(X) = b) = 0$. Then, a test point X_{n+1} almost surely does not belong to the bin, and the bin can be ignored for our calibration guarantee. Thus without loss of generality, suppose every $b \in [B]$ satisfies

$$\mathbb{P}(\mathcal{B}(X) = b) > 0.$$

Let $\min_{b \in [B]} \mathbb{P}(\mathcal{B}(X) = b) = \tau > 0$. Then for a fixed number of samples n, any particular bin b, and any constant $\alpha \in (0, 1)$ we have by Hoeffding's inequality with probability $1 - \alpha/B$

$$N_b \geqslant n\tau - \sqrt{\frac{n \ln(B/\alpha)}{2}}.$$

Taking a union bound, we have with probability $1 - \alpha$, simultaneously for every $b \in [B]$,

$$N_b \geqslant n\tau - \sqrt{\frac{n \ln(B/\alpha)}{2}} = \Omega(n),$$

and in particular $N_{b^\star} = \Omega(n)$ where $b^\star = \arg\min_{b \in [B]} N_b$. Thus by the first part of this corollary, f_n is ε_n calibrated where $\varepsilon_n = O(\sqrt{n^{-1}}) = o(1)$. This concludes the proof. □

C.3 Proof of Theorem 5

Denote $|\mathcal{D}^2_{cal}| = n$. Let $p_j = \mathbb{P}(g(X) \in I_j)$ be the true probability that a random point falls into partition $\mathcal{X}_j$. Assume c is such that we can use Lemma 11 to guarantee that with probability at least $1-\alpha/2$, uniform mass binning scheme is 2-well-balanced. Hence, with probability at least $1-\alpha/2$:

$$\frac{1}{2B} \leqslant p_j \leqslant \frac{2}{B}, \ \forall j \in [B]. \tag{31}$$

Moreover, by Hoeffding's inequality we get that for any fixed region of sample-space partition, with probability at least $1-\alpha/2B$, for a fixed $j \in [B]$,

$$N_j \geqslant np_j - \sqrt{\frac{n\ln(2B/\alpha)}{2}}. \tag{32}$$

Hence, by union bound across applied accross all regions and using (31), we get that with probability at least $1-\alpha/2$:

$$N_{b^\star} \geqslant \frac{n}{2B} - \sqrt{\frac{n\ln(2B/\alpha)}{2}},$$

where the first term dominates asymptotically (for fixed B). Hence, we get that with probability at least $1-\alpha$, $N_{b^\star} = \Omega\left(n/B\right)$. By invoking the result of Corollary 4 and observing that $\widehat{V}_b \leqslant 1$, we conclude that uniform mass binning is (ε, α) approximately calibrated with $\varepsilon = O(\sqrt{B\ln(B/\alpha)/n})$ as desired. This also leads to asymptotic calibration by Corollary 4. □

C.4 Proof of Theorem 6

The proof is based on the result for an empirical-Bernstein confidence sequences for bounded observations [15]. We condition on the event $E^\infty_{\mathcal{B}(x)}$ defined as $(\mathcal{B}(X_1), \mathcal{B}(X_1), \dots) = (\mathcal{B}(x_1), \mathcal{B}(x_2), \dots)$, that is the random variables denoting which partition the infinite stream of samples fall in (thus allowing our bound to hold for every possible value of n). On $E^\infty_{\mathcal{B}(x)}$, the label values within each partition of the sample-space partition represent independent Bernoulli random variable that share the same mean $\pi_b = \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right], b \in [B]$. Consequently, the bound obtained can be marginalized over $E^\infty_{\mathcal{B}(x)}$ to obtain the assertion of the theorem in unconditional form. Now we show the bound that applies conditionally on $E^\infty_{\mathcal{B}(x)}$.

Consider any fixed region of the sample-space partition $\mathcal{X}_b$ and corresponding points $\left\{\left(X^b_i, Y^b_i\right)\right\}_{i=1}^{N_b}$. Then $S_t = \left(\sum_{i=1}^t Y^b_i\right) - t\pi_b$ is a sub-exponential process with variance process:

$$\widehat{V}^+_t = \sum_{i=1}^t \left(Y^b_i - \overline{Y}^b_{i-1}\right)^2.$$

Howard et al. [14, Proposition 2] implies that S_t is also a sub-gamma process with variance process $\widehat{V}_t$ and the same scale $c = 1$. Since the theorem holds for any sub-exponential uniform boundary, we choose one based on analytical convenience. Recall definition of the polynomial stitching function

$$\mathcal{S}_\alpha(v) := \sqrt{k_1^2 v l(v) + k_2^2 c^2 l^2(v)} + k_2 c l(v), \ \text{ where } \begin{cases} l(v) := \ln h(\ln_\eta(v/m)) + \ln(l_0/\alpha), \\ k_1 := (\eta^{1/4} + \eta^{-1/4})/\sqrt{2}, \\ k_2 := (\sqrt{\eta}+1)/\sqrt{2}. \end{cases}$$

where $l_0 = 1$ for the scalar case. Note that for $c > 0$ it holds that $\mathcal{S}_\alpha(v) \leqslant k_1\sqrt{vl(v)} + 2ck_2 l(v)$.

From Howard et al. [15, Theorem 1], it follows that $u(v) = \mathcal{S}_\alpha(v \vee m)$ is a sub-gamma uniform boundary with scale c and crossing probability α. Applying Theorem 9 with $h(k) \leftarrow (k+1)^s\zeta(s)$ where $\zeta(\cdot)$ is Riemann zeta function and parameters $\eta \leftarrow e$, $s \leftarrow 1.4$, $c \leftarrow 1$, $m \leftarrow 1$ and $\alpha \leftarrow \alpha/(2B)$, yields that $k_2 \leqslant 1.88, k_1 \leqslant 1.46$ and $l(v) = 1.4 \cdot \ln\ln(ev) + \ln(2\zeta(1.4)B/\alpha)$. Since Theorem 9 provides a bound that holds uniformly across time t, then it provides a guarantee for $t = N_b$, in particular. Hence, with probability at least $1-\alpha/B$,

$$|\pi_b - \widehat{\pi}_b| \leqslant \frac{1.46\sqrt{\widehat{V}^+_b \cdot 1.4 \cdot \ln\ln\left(e\left(\widehat{V}^+_b \vee 1\right)\right) + \ln(6.3B/\alpha)}}{N_b} + \frac{5.27 \cdot \ln\ln\left(e\left(\widehat{V}^+_b \vee 1\right)\right) + 3.76\ln(6.3B/\alpha)}{N_b}$$

$$\leqslant \frac{7\sqrt{\widehat{V}_b^+ \cdot \ln\ln\left(e\left(\widehat{V}_b^+ \vee 1\right)\right)} + 5.3\ln(6.3B/\alpha)}{N_b}.$$

using that $\sqrt{x+y} \leqslant \sqrt{x} + \sqrt{y}$ and $\ln\ln(ex) \leqslant \sqrt{x \ln\ln ex}$ for $x \geqslant 1$. Finally, we apply a union bound to get a guarantee that holds simultaneously for all regions of the sample-space partition. □

D Calibration under covariate shift (including proofs of results in Section 4.4)

The results from Section 4.4 are proved in Appendix D.1 (Theorem 7) and D.3 (Proposition 2). To show Theorem 7, we first propose and analyze a slightly different estimator than (39) that is unbiased for $\pi_b^{(w)}$, but needs additional oracle access to the parameters $\{m_b\}_{b\in[B]}$ defined as

$$m_b = \mathbb{P}_{P_X}(X \in \mathcal{X}_b) \,/\, \mathbb{P}_{\widetilde{P}_X}(X \in \mathcal{X}_b).$$

m_b denotes the 'relative mass' of region $\mathcal{X}_b$. (For simplicity, we assume that $\mathbb{P}_{\widetilde{P}}(X \in \mathcal{X}_b) > 0$ for every b since otherwise the test-point almost surely does not belong to $\mathcal{X}_b$ and estimation in that bin is not relevant for a calibration guarantee.) We then show that m_b can be estimated using w, which would lead to the proposed estimator $\breve{\pi}_b^{(w)}$. First, we establish the following relationship between $\mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]$ and $\mathbb{E}_P\left[Y \mid X \in \mathcal{X}_b\right]$.

Proposition 3. *Under the covariate shift assumption, for any $b \in [B]$*

$$\mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right] = m_b \cdot \mathbb{E}_P\left[w(X)Y \mid X \in \mathcal{X}_b\right].$$

Proof. Observe that

$$\frac{d\widetilde{P}(X \mid X \in \mathcal{X}_b)}{dP(X \mid X \in \mathcal{X}_b)} = \frac{d\widetilde{P}(X)}{dP(X)} \cdot \frac{\mathbb{P}_P\left(X \in \mathcal{X}_b\right)}{\mathbb{P}_{\widetilde{P}}(X \in \mathcal{X}_b)} = w(X) \cdot m_b.$$

Thus we have,

$$\begin{aligned}
\mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right] &\overset{(1)}{=} \mathbb{E}_{\widetilde{P}}\left[\mathbb{E}_{\widetilde{P}}\left[Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(2)}{=} \mathbb{E}_{\widetilde{P}}\left[\mathbb{E}_P\left[Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(3)}{=} \mathbb{E}_P\left[\frac{d\widetilde{P}(X \mid X \in \mathcal{X}_b)}{dP(X \mid X \in \mathcal{X}_b)} \cdot \mathbb{E}_P\left[Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(4)}{=} m_b \cdot \mathbb{E}_P\left[w(X)\mathbb{E}_P\left[Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(5)}{=} m_b \cdot \mathbb{E}_P\left[\mathbb{E}_P\left[w(X)Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(6)}{=} m_b \cdot \mathbb{E}_P\left[w(X)Y \mid X \in \mathcal{X}_b\right],
\end{aligned}$$

where in (1) we use the tower rule, in (2) we use the covariate shift assumption, (3) can be seen by using the integral form of the expectation, (4) uses the observation at the beginning of the proof, (5) uses that $w(X)$ is a function of X and finally, (6) uses the tower rule. □

Let N_b denote the number of calibration points from the source domain that belong to bin b. Given Proposition 3, a natural estimator for $\mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]$ is given by:

$$\widehat{\pi}_b^{(w)} := \frac{1}{N_b} \sum_{i:\mathcal{B}(X_i)=b} m_b w(X_i) Y_i. \tag{33}$$

Estimation properties of $\widehat{\pi}_b^{(w)}$ are given by the following theorem.

Theorem 8. *Assume that* $\sup_x w(x) = U < \infty$. *For any* $\alpha \in (0,1)$, *with probability at least* $1-\alpha$,

$$\left|\widehat{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| \leqslant \sqrt{\frac{2\widehat{V}_b^{(w)} \ln(3B/\alpha)}{N_b}} + \frac{3m_b U \ln(3B/\alpha)}{N_b}, \quad \textit{simultaneously for all } b \in [B],$$

where $\widehat{V}_b^{(w)} = \frac{1}{N_b}\sum_{i:\mathcal{B}(X_i)=b}(m_b w(X_i) Y_i - \widehat{\pi}_b^{(w)})^2$.

The proof is given in Appendix D.2. Next, we discuss a way of estimating m_b using likelihood ratio w instead of relying on oracle access. Observe that

$$\frac{d\widetilde{P}(X \mid X \in \mathcal{X}_b)}{dP(X \mid X \in \mathcal{X}_b)} = \frac{d\widetilde{P}(X)}{dP(X)} \cdot \frac{\mathbb{P}_P\left(X \in \mathcal{X}_b\right)}{\mathbb{P}_{\widetilde{P}}(X \in \mathcal{X}_b)} = w(X) \cdot m_b.$$

Thus we have,

$$\mathbb{E}_P\left[w(X) \mid X \in \mathcal{X}_b\right] = m_b^{-1}\mathbb{E}_P\left[\frac{d\widetilde{P}(X \mid X \in \mathcal{X}_b)}{dP(X \mid X \in \mathcal{X}_b)} \mid X \in \mathcal{X}_b\right] = m_b^{-1}, \tag{34}$$

which suggests a possible estimator for m_b given by

$$\widehat{m}_b = \left(\frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}{N_b}\right)^{-1}, \quad b \in [B]. \tag{35}$$

On substituting this estimate for m_b in (33), we get a new estimator

$$\frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)},$$

which is exactly $\breve{\pi}_b^{(w)}$. With this observation, we now prove Theorem 7.

D.1 Proof of Theorem 7

Let us define $r_b := 1/m_b$ and

$$\widehat{r}_b = \frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}{N_b}. \tag{36}$$

Step 1 (Uniform lower bound for N_b). Since the regions of the sample-space partition were constructed using uniform-mass binning, the guarantee of Theorem 5 holds. Precisely, we have that with probability at least $1-\alpha/3$, simultaneously for every $b \in [B]$,

$$N_b \geqslant \frac{n}{2B} - \sqrt{\frac{n \ln(6B/\alpha)}{2}}.$$

Step 2 (Approximating r_b). Observe that the estimator (36) is an average of N_b random variables bounded by the interval $[0, U]$. Let $E_{\mathcal{B}(x)}$ be the event that $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. On the event $E_{\mathcal{B}(x)}$, within each region $\mathcal{X}_b$, the number of point from the calibration set is known and the Y_i's in each bin represent independent Bernoulli random variables that share the same mean $\mathbb{E}\left[w(X) \mid X \in \mathcal{X}_b\right]$. Consider any fixed region $\mathcal{X}_b$, $b \in [B]$. By Hoeffding's inequality, it holds that

$$\mathbb{P}\left(|r_b - \widehat{r}_b| > \sqrt{\frac{U^2 \ln(6B/\alpha)}{2N_b}} \;\middle|\; E_{\mathcal{B}(x)}\right) \leqslant \alpha/(3B).$$

Applying union bound across all regions of the sample-space partition, we get that:

$$\mathbb{P}\left(\exists b \in [B] :\ |r_b - \widehat{r}_b| > \sqrt{\frac{U^2 \ln(6B/\alpha)}{2N_b}} \;\middle|\; E_{\mathcal{B}(x)}\right) \leqslant \alpha/3.$$

Because this is true for any $E_{\mathcal{B}(x)}$, we can marginalize to obtain that with probability at least $1-\alpha/3$,

$$\forall b \in [B],\ |r_b - \widehat{r}_b| \leqslant \sqrt{\frac{U^2 \ln(6B/\alpha)}{2N_b}}. \tag{37}$$

Step 3 (Going from r_b to m_b). Define $r^\star = \min_{b\in[B]} \mathbb{E}\left[w(X) \mid X \in \mathcal{X}_b\right]$. Suppose $\forall b \in [B]$, $|r_b - \widehat{r}_b| \leqslant \varepsilon$ and $\varepsilon < r^\star/2$. Then, we have with probability at least $1-\alpha/3$:

$$|m_b - \widehat{m}_b| = \left|\frac{1}{r_b} - \frac{1}{\widehat{r}_b}\right| = \left|\frac{r_b - \widehat{r}_b}{r_b \cdot \widehat{r}_b}\right| \leqslant \frac{\varepsilon}{r_b^2|1-\varepsilon/r_b|} \leqslant \frac{2\varepsilon}{r_b^2} = 2m_b^2\varepsilon, \quad \forall b \in [B]. \tag{38}$$

We now set $\varepsilon = \sqrt{\frac{U^2\ln(6B/\alpha)}{2N_b}}$ as specified in equation (37) and verify that $\varepsilon < r^\star/2$.

- First, from step 1, with probability at least $1-\alpha/3$, $N_{b^\star} = \Omega(n/B)$ and thus $N_b = \Omega(n/B)$ for every $b \in [B]$.
- By the condition in the theorem statement, for every $b \in [B]$,

$$\varepsilon = \sqrt{\frac{U^2\ln(6B/\alpha)}{2N_b}} = O\left(\sqrt{\frac{U^2 B\ln(6B/\alpha)}{n}}\right) = O\left(\sqrt{\frac{U^2 B\ln(6B/\alpha)}{\left(\frac{U^2 B\ln(6B/\alpha)}{L^2}\right)}}\right) = O\left(L\right).$$

  Finally recall that $L \leqslant r^\star$. Thus we can pick c in the theorem statement to be large enough such that $\varepsilon < L/2 \leqslant r^\star/2$.

Thus for $\varepsilon = \sqrt{\frac{U^2\ln(6B/\alpha)}{2N_b}}$, by a union bound over the event in (37) and step 1, the conditions for (38) are satisfied with probability at least $1 - 2\alpha/3$. Hence we have for some large enough constant $c > 0$,

$$|m_b - \widehat{m}_b| \leqslant cm_b^2 \cdot \sqrt{\frac{U^2 B\ln(6B/\alpha)}{2n}} \leqslant c \cdot \frac{U}{L^2}\sqrt{\frac{B\ln(6B/\alpha)}{2n}}.$$

The final inequality holds by observing that $m_b \leqslant 1/L$ which follows from relationship (34) and the assumption that $\inf_x w(x) \geqslant L$.

Step 4 (Computing the final deviation inequality for $\breve{\pi}_b^{(w)}$). Recall the definitions of the two estimators:

$$\widehat{\pi}_b^{(w)} := \frac{1}{N_b}\sum_{i:\mathcal{B}(X_i)=b} m_b w(X_i) Y_i,$$

and

$$\breve{\pi}_b^{(w)} := \frac{1}{N_b}\sum_{i:\mathcal{B}(X_i)=b} \widehat{m}_b w(X_i) Y_i,$$

which differ by replacing m_b by its estimator $\widehat{m}_b$ defined in (35). By triangle inequality,

$$|\breve{\pi}_b - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]| \leqslant \left|\breve{\pi}_b^{(w)} - \widehat{\pi}_b^{(w)}\right| + \left|\widehat{\pi}_b^{(w)} - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]\right|.$$

Theorem 8 bounds the term $\left|\widehat{\pi}_b^{(w)} - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]\right|$ with high probability. In the proof of Theorem 8, we can replace the empirical Bernstein's inequality by Hoeffding's inequality to obtain with probability at least $1-\alpha/3$,

$$\left|\widehat{\pi}_b^{(w)} - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]\right| \leqslant \sqrt{\frac{U^2\ln(6B/\alpha)}{2N_b}} \leqslant \left(\frac{U}{L}\right)^2\sqrt{\frac{\ln(6B/\alpha)}{2N_b}},$$

simultaneously for all $b \in [B]$ (the last inequality follows since $L \leqslant 1 \leqslant U$). To bound $\left|\widehat{\pi}_b^{(w)} - \breve{\pi}_b^{(w)}\right|$, first note that:

$$\begin{aligned}\left|\widehat{\pi}_b^{(w)} - \breve{\pi}_b^{(w)}\right| &= \left|\frac{1}{N_b}\sum_{i:\mathcal{B}(X_i)=b}\left(\widehat{m}_b - m_b\right) w(X_i) Y_i\right| \\ &\leqslant U \cdot \left|\frac{1}{N_b}\sum_{i:\mathcal{B}(X_i)=b}\left(\widehat{m}_b - m_b\right)\right|\end{aligned}$$

$$= U \cdot |\widehat{m}_b - m_b| .$$

Then we use the results from steps 1 and 3 to conclude that with probability at least $1 - 2\alpha/3$,

$$\left|\breve{\pi}_b^{(w)} - \widehat{\pi}_b^{(w)}\right| \leqslant c \cdot \left(\frac{U}{L}\right)^2 \sqrt{\frac{B \ln(6B/\alpha)}{2n}}, \text{ and } N_b \geqslant n/B - \sqrt{\frac{n \ln(6B/\alpha)}{2}}.$$

simultaneously for all $b \in [B]$. Thus by union bound, we get that it holds with probability at least $1 - \alpha$,

$$|\breve{\pi}_b - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]| \leqslant c \cdot \left(\frac{U}{L}\right)^2 \sqrt{\frac{B \ln(6B/\alpha)}{2n}},$$

simultaneously for all $b \in [B]$ and large enough absolute constant $c > 0$. This concludes the proof. □

D.2 Proof of Theorem 8

Consider the event $E_{\mathcal{B}(x)}$ defined as $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. Conditioned on $E_{\mathcal{B}(x)}$, since $\sup_x w(x) \leqslant U$, we get that $\widehat{\pi}_b^{(w)}$ is an average of independent non-negative random variables $m_b w(X_i) Y_i$ that are bounded by $m_b U$ and share the same mean $m_b \, \mathbb{E}_P\left[w(X)Y \mid X \in \mathcal{X}_b\right] = \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]$ (by Proposition 3).Using Theorem 10 for a fixed $b \in [B]$, we obtain:

$$\mathbb{P}\left(\left|\widehat{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| > \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3 m_b U \ln(3B/\alpha)}{N_b} \;\Big|\; E_{\mathcal{B}(x)}\right) \leqslant \alpha/B.$$

Applying a union bound over all $b \in [B]$, we get:

$$\mathbb{P}\left(\forall b \in [B] : \left|\widehat{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| \leqslant \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{N_b}} + \frac{3 m_b U \ln(3B/\alpha)}{N_b} \;\Big|\; E_{\mathcal{B}(x)}\right) \geqslant 1 - \alpha.$$

Because this is true for any $E_{\mathcal{B}(x)}$, we can marginalize to obtain the assertion of the theorem in unconditional form. □

D.3 Proof of Proposition 2

Fix any $\alpha \in (0, 1)$. For any $k \in \mathbb{N}$ observe that by triangle inequality,

$$\left|\breve{\pi}_b^{(\widehat{w}_k)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| \leqslant \left|\breve{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| + \left|\breve{\pi}_b^{(w)} - \breve{\pi}_b^{(\widehat{w}_k)}\right|.$$

Consider any $\varepsilon > 0$. Note that by Theorem 7, there exists sufficiently large n such that the first term is larger than $\varepsilon/2$ with probability at most $\alpha/2$ simultaneously for all $b \in [B]$. Hence, it suffices to show that there exists a large enough k such that the probability of the second term exceeding $\varepsilon/2$ is at most $\alpha/2$ simultaneously for all $b \in [B]$. While analyzing the second term, we treat n as a constant while leveraging the consistency of $\widehat{w}_k$ as $k \to \infty$. For simplicity, denote $\Delta_k = \sup_x |w(x) - \widehat{w}_k(x)|$. Then for any $b \in [B]$:

$$\begin{aligned}
\left|\breve{\pi}_b^{(w)} - \breve{\pi}_b^{(\widehat{w}_k)}\right| &= \left|\frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)} - \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)}\right| \\
&\overset{(1)}{\leqslant} \left|\frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)} - \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}\right| \\
&\quad + \left|\frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)} - \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)}\right| \\
&\overset{(2)}{\leqslant} n \cdot \Delta_k \cdot \left|\frac{1}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}\right|
\end{aligned}$$

$$+\left|\frac{1}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)} - \frac{1}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)}\right| \left|\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i) Y_i\right|$$

$$\overset{(3)}{\leqslant} \frac{n}{L} \cdot \Delta_k + \left(\frac{n \cdot \Delta_k}{(L-\Delta_k)L}\right) \cdot ((U+\Delta_k) \cdot n),$$

where (1) is due to the triangle inequality, (2) is due to the facts that the number of points in any bin is at most n and that absolute difference between $\widehat{w}$ and w is at most Δ_k, (3) combines the aforementioned reasons in (2) and the assumptions: $L \leqslant \inf_x w(x) \leqslant \sup_x w(x) \leqslant U$. Since $\Delta_k \xrightarrow{P} 0$, clearly there exists a large enough k such that:

$$\mathbb{P}\left(\left|\breve{\pi}_b^{(w)} - \breve{\pi}_b^{(\widehat{w}_k)}\right| \geqslant \varepsilon/2\right) \leqslant \alpha/2.$$

Thus we conclude that $\breve{\pi}_b^{(\widehat{w}_k)}$ is asymptotically calibrated at level α. □

D.4 Preliminary simulations

This section is structured as follows. We first describe the overall procedure for calibration under covariate shift. The finite-sample calibration guarantee of Theorem 7 holds for oracle w whereas in our experiments we will estimate w; to assess the loss in calibration due to this approximation, we introduce some standard techniques used in literature. The preliminary experiments are performed with simulated data which are described after this. Finally, we propose a modified estimator $\widetilde{\pi}_b^{(\widehat{w})}$ of $\mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b]$ which appears natural but has poor performance in practice.

Procedure. We describe how to construct approximately calibrated predictions practically. This involves approximating the importance weights w and the relatives mass terms $\{m_b\}_{b\in[B]}$. The summarized calibration procedure consists of the following steps:

1. Split the calibration set into two parts and use the first to perform *uniform mass* binning
2. Given unlabeled examples from both source and target domain, estimate $\widehat{w}$. The unconstrained Least-Squares Importance Fitting (uLSIF) procedure [17] is used for this.
3. Compute for every $b \in [B]$, the estimator as per (17), replacing w with $\widehat{w}$:

$$\breve{\pi}_b^{(\widehat{w})} := \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}(X_i)}. \tag{39}$$

4. On a new test point from the target distribution, output the calibrated estimate $\breve{\pi}^{(\widehat{w})}_{\mathcal{B}(X_{n+1})}$.

Assessment through reliability diagrams and ECE. Given a test set (from the target distribution) of size m: $\{(X_i', Y_i')\}_{i\in[m]}$ and a function $g : \mathcal{X} \to [0,1]$ that outputs approximately calibrated probabilities, we consider the reliability diagram to estimate its calibration properties. A reliability diagram is constructed using splitting the unit interval $[0,1]$ into non-overlapping intervals $\{I_b\}_{b\in[B']}$ for some B' as

$$I_i = \left[\frac{i-1}{B'}, \frac{i}{B'}\right), \; i = 1, \ldots, B'-1 \;\text{ and }\; I_{B'} = \left[\frac{B'-1}{B'}, 1\right].$$

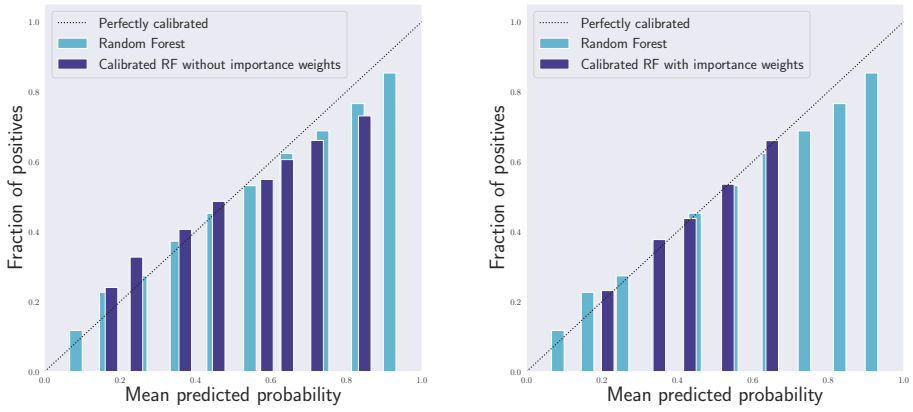
Let $\mathcal{B}' : [0,1] \to [B']$ denote the binning function that corresponds to this binning. We then compute the following quantities for each bin $b \in [B']$:

$$\mathrm{FP}(I_b) = \frac{\sum_{i:\mathcal{B}'(X_i')=b} Y_i'}{|\{i : \mathcal{B}'(X_i') = b\}|} \qquad \text{(fraction of positives in a bin)},$$

$$\mathrm{MP}(I_b) = \frac{\sum_{i:\mathcal{B}'(X_i')=b} g(X_i')}{|\{i : \mathcal{B}'(X_i') = b\}|} \qquad \text{(mean predicted probability in a bin)}.$$

If g is perfectly calibrated, the reliability diagram is diagonal. Define the proportion of points that fall into various bins as:

$$\widehat{p}_b = \frac{|\{i : \mathcal{B}'(X_i') = b\}|}{m}, \quad b \in [B'].$$



(a) (b)

Figure 2: In Figure 2a uncalibrated Random Forest (ECE ≈ 0.023) is compared with calibration that does not take the covariate shift into account (ECE ≈ 0.047). In Figure 2b uncalibrated Random Forest is compared with calibration that takes the covariate shift into account (ECE ≈ 0.017).

Then ECE (or ℓ_1-ECE) is defined as:

$$\mathrm{ECE}(g) = \sum_{b \in [B']} \widehat{p}_b \cdot |\mathrm{MP}(I_b) - \mathrm{FP}(I_b)| \,.$$

ECE can also be defined in the ℓ_p sense and for multiclass problems but we limit our attention to the ℓ_1-ECE for binary problems.

Simulations with synthetic data. We illustrate the performance of our proposed estimator (17) using the following simulated example, for which we can explicitly control the covariate shift. Consider the following data generation pipeline: for the source domain each component of the feature vector is drawn from $\mathrm{Beta}(\alpha, \beta)$ where $\alpha = \beta = 1$, which corresponds to uniform draws from the unit cube. For the target distribution each component can be drawn independently from $\mathrm{Beta}(\alpha', \beta')$. If the dimension is d, the true likelihood ratio is given as

$$w(x) = \frac{d\widetilde{P}_X(x)}{dP_X(x)} = \frac{B^d(\alpha;\beta)}{B^d(\alpha';\beta')} \prod_{i=1}^{d} \frac{(x_{(i)})^{\alpha'-1}(1-x_{(i)})^{\beta'-1}}{(x_{(i)})^{\alpha-1}(1-x_{(i)})^{\beta-1}},$$

where $x_{(i)}$ are the coordinates of feature vector x. We set $d = 3$ and $\alpha' = 2, \beta' = 1$ so that $w(x) = 8 \cdot x_{(1)}x_{(2)}x_{(3)}$. The labels for both source and target distributions are assigned according to:

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{1}{2}\left(1 + \sin\left(\omega\left(x_{(1)}^2 + x_{(2)}^2 + x_{(3)}^2\right)\right)\right),$$

for $\omega = 20$. As the underlying classifier we use a Random Forest with 100 trees (from `sklearn`). 14700 data points were used to train the underlying Random Forest classifier, 2000 data points from both source and target were used for the estimation of importance weights. The parameters σ and λ for uLSIF were tuned by leave-one-out cross-validation: we considered 25 equally spaced values on a log-scale in range $(10^{-2}, 10^2)$ for σ and 100 equally spaced values on a log-scale in range $(10^{-3}, 10^3)$ for λ. Uniform mass binning was performed with 10 bins and 1940 data points from the source domain were used to estimate the quantiles. 7840 source data points were used for the calibration and finally, 28000 data points from the target domain were used for evaluation purposes. We note that this simulation is a 'proof-of-concept'; the sample sizes we used are not necessarily optimal can presumably be improved.

We compare the unweighted estimator (12) which corresponds to weighing points in each bin equally as we would do if there was no covariate shift, and the estimator (17) that uses an estimate of w to account for covariate shift. The reliability diagrams are presented in Figure 2, with the ECE reported in the caption. For the ECE estimation and reliability diagrams, we used $B' = 10$.

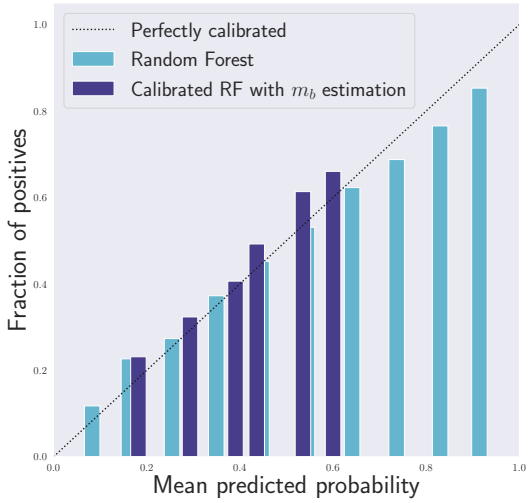


Figure 3: Calibration of Random Forest with m_b estimated as per equation (35) (ECE ≈ 0.05).

Alternative estimator for m_b. Estimator (35) is one way of estimating m_b using the w values, that leads to (17). However, there exists another natural estimator which we propose and show some preliminary empirical results for. Suppose we have access to additional unlabeled data from the source and target domains ($\{X_i^s\}_{i\in[n_s]}$, and $\{X_i^t\}_{i\in[n_t]}$ respectively). From the definition of $m_b = \mathbb{P}_{P_X}(X \in \mathcal{X}_b)/\mathbb{P}_{\tilde{P}_X}(X \in \mathcal{X}_b)$, a natural estimator is,

$$\widehat{m}_b = \frac{\frac{1}{n_s}\left|\{i \in [n_s] : \mathcal{B}(X_i^s) = b\}\right|}{\frac{1}{n_t}\left|\{i \in [n_t] : \mathcal{B}(X_i^t) = b\}\right|}, \quad b \in [B]. \tag{40}$$

In this case, the estimator (33) reduces to:

$$\widetilde{\pi}_b^{(\widehat{w})} = \frac{\widehat{m}_b}{N_b} \sum_{i:\mathcal{B}(X_i)=b} \widehat{w}(X_i) Y_i.$$

We show experimental results with this estimation procedure. We used 8500 data points from the source domain and 8000 points from the target domain to compute (40). The reliability diagram and ECE with this estimator is reported in Figure 3. On our simulated dataset, we observe that the estimators $\widetilde{\pi}_b^{(\widehat{w})}$ perform significantly worse than the estimators $\widehat{\pi}_b^{(\widehat{w})}$. While this is only a single experimental setup, we outline some drawbacks of this estimation method that may lead to poor performance in general.

1. $\widetilde{\pi}_b^{(\widehat{w})}$ requires access to additional unlabeled data from the source and target domains without leading to increase in performance.
2. The denominator of $\widehat{m}_b$ could be badly behaved if the number of points from the target domain in bin b are small. We could perform uniform-mass binning on the target domain to avoid this, but in this case N_b may be small which would lead to the estimator $\widetilde{\pi}_b^{(\widehat{w})}$ performing poorly.

Our overall recommendation through these preliminary experiments is to use the estimator $\widehat{\pi}_b^{(\widehat{w})}$ as proposed in Section 4.4 instead of $\widetilde{\pi}_b^{(\widehat{w})}$.

E Venn prediction

Venn prediction [24, 45–47] is a calibration framework that provides distribution-free guarantees, which are different from the ones in Definitions 1 and 2. For a multiclass problem with L labels, Venn prediction produces L predictions, one of which is guaranteed to be perfectly calibrated

(although it is impossible to know which one). These are called multiprobabilistic predictors, formally defined as a collection of predictions $(f_1, f_2, \dots f_L)$ where each $f_i \in \{\mathcal{X} \to \Delta_{L-1}\}$ (here Δ_{L-1} is the boundary of the ℓ_1 ball in the non-negative orthant of $\mathbb{R}^L$, corresponding to all possible distributions over $\{1, 2, \dots, L\}$). Vovk and Petej [45] defined two calibration guarantees for multiprobabilistic predictors, the first being oracle calibration.

Definition 4 (Oracle calibration)**.** $(f_1, f_2, \dots f_L)$ is oracle calibrated if there exists an oracle selector S such that f_S is perfectly calibrated.

Venn predictors satisfy oracle calibration [45, Theorem 1] with $S = Y$. In the binary case, this means that when $Y = 1$, $f_1(X)$ is perfectly calibrated but we do not have any guarantee on $f_0(X)$; on the other hand if $Y = 0$, $f_0(X)$ is perfectly calibrated but we know nothing about $f_1(X)$. Since Y is unknown, oracle calibration seems to us to primarily serve as theoretical guidance, but does not give a clear prescription on what to output and what theoretical guarantee that output satisfies. In practice, it seems reasonable to suspect that if $f_0(X)$ and $f_1(X)$ are close, then their average should be approximately calibrated in the sense of Definition 1, but to the best of our knowledge, such results have not been shown formally (other aggregate functions apart from average are also suggested (without formal guarantees) by Vovk and Petej [45, Section 4]). For instance, it may be tempting to think that oracle calibration of a multiprobabilistic predictor leads to approximate calibration in the following way. Consider the prediction function

$$f(X) = \frac{\min f_i(X) + \max f_i(X)}{2},$$

and the radius of the interval $[\min f_i(X), \max f_i(X)]$:

$$\varepsilon(X) = \frac{\max f_i(X) - \min f_i(X)}{2}.$$

Since Venn predictors satisfy oracle calibration, one might conjecture that f is (ε, α) approximately calibration (per Definition 1) for the given function ε and for any $\alpha \in (0, 1)$. We examined this claim but were unable to prove such a guarantee formally. In fact, it seems that no general calibration guarantee should be possible with the size of the calibration interval being $O(\varepsilon(X))$; we evidence this through the following construction.

Consider a setup, with no covariates and only label values Y, and a single bin that contains all points (in the Venn prediction language: a taxonomy under which all points are equivalent). For a test-point Y_{n+1} and any predictor f, note that $\mathbb{E}\left[Y_{n+1} \mid f\right]$ is simply equal to $\mathbb{E}\left[Y_{n+1}\right]$ since any information used to construct f is independent of Y_{n+1}. To ensure calibration, we may look for a guarantee of the following form for some δ:

$$\left|\mathbb{E}\left[Y_{n+1} \mid f\right] - f\right| = \left|\mathbb{E}\left[Y_{n+1}\right] - f\right| \leqslant \delta.$$

In essence, f is an estimator for the parameter $\mathbb{E}\left[Y\right]$ with a corresponding deviation bound of δ. Without distributional assumptions, we only expect to estimate such a parameter with error at best $\delta = O(1/\sqrt{n})$ for a fixed constant probability of failure. On the other hand, the Venn prediction interval $[\min f_i, \max f_i]$ often has radius $O(1/n)$. Thus for valid approximate calibration, we would need to provide a larger interval than $[\min f_i, \max f_i]$, even though one of the f_i's is perfectly calibrated. Given this example, our conjecture is that it might be possible to show that there always exists an $f_i(X)$ that is $\left(n^{-0.5}\text{polylog}\left(1/\alpha\right)), \alpha\right)$ calibrated. Without knowing which $f_i(X)$ to pick, perhaps one can show that an aggregate point in the interval $[\min f_i, \max f_i]$ is $\left((\max f_i - \min f_i) + n^{-0.5}\text{polylog}\left(1/\alpha\right), \alpha\right)$ approximately calibrated. In Section 4, we showed such a result for histogram binning (which can be interpreted as a Venn predictor). It would be interesting to study if such results can be shown for general Venn predictors.

Another guarantee for multiprobabilistic predictors is calibration in the large.

Definition 5 (Calibration in the large)**.** $(f_1, f_2, \dots f_L)$ is calibrated in the large if the following is satisfied: $\mathbb{E}\left[Y\right] \in [\mathbb{E}\min f_i(X), \mathbb{E}\max f_i(X)]$.

Vovk and Petej [45, Theorem 2] show that Venn predictors satisfy calibration in the large. Due to the expectation signs and the coverage of the marginal probability $\mathbb{E}\left[Y\right]$, calibration in the large does not lead to a clear interpretable guarantee for uncertainty quantification, but rather a minimum requirement that serves as a guiding principle.

F Auxiliary results

F.1 Concentration inequalities

Theorem 9 (Howard et al. [15], Theorem 4)**.** *Suppose $Z_t \in [a, b]$ a.s. for all t. Let $(\widehat{Z}_t)$ be any $[a, b]$-valued predictable sequence, and let u be any sub-exponential uniform boundary with crossing probability α for scale $c = b - a$. Then:*

$$\mathbb{P}\left(\forall t \geqslant 1 : \left|\overline{Z}_t - \mu_t\right| < \frac{u\left(\sum_{i=1}^{t}\left(Z_i - \widehat{Z}_i\right)^2\right)}{t}\right) \geqslant 1 - 2\alpha.$$

Theorem 10 (Partial statement of Audibert et al. [2], Theorem 1)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables bounded in $[0, s]$, for some $s > 0$. Let $\mu = \mathbb{E}[X_1]$ be their common expected value. Consider the empirical mean $\overline{X}_n$ and variance V_n defined respectively by*

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}, \quad \text{and} \quad V_n = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{n}.$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\left|\overline{X}_n - \mu\right| \leqslant \sqrt{\frac{2V_n \log(3/\delta)}{n}} + \frac{3s \log(3/\delta)}{n}.$$

F.2 Uniform-mass binning

Kumar et al. [21] defined well-balanced binning and showed that uniform mass-binning is well-balanced.

Definition 6 (Well-balanced binning)**.** A binning scheme $\mathcal{B}$ of size B is β-well-balanced ($\beta \geqslant 1$) for some classifier g if

$$\frac{1}{\beta B} \leqslant \mathbb{P}\left(g(X) \in I_b\right) \leqslant \frac{\beta}{B},$$

simultaneously for all $b \in [B]$.

To perform uniform-mass binning labeled examples are required at the stage of training the base classifier $g(\cdot)$. We denote this data as $\mathcal{D}^1_{\text{cal}}$. Procedures based on uniform-mass binning are well-balanced if $\left|\mathcal{D}^1_{\text{cal}}\right|$ is sufficiently large.

Lemma 11 (Kumar et al. [21], Lemma 4.3)**.** *For a universal constant $c > 0$, if $\left|\mathcal{D}^1_{cal}\right| \geqslant cB \ln(B/\alpha)$, then with probability at least $1 - \alpha$, the uniform mass binning scheme $\mathcal{B}$ is 2-well-balanced.*

The calibration guarantees in Section 4 depend on the minimum number of training points $N_{b^\star}$ in any bin. Uniform mass-binning guarantees that $N_{b^\star} = \Omega(n/B)$. This is used in the proof of Theorem 5.