Distribution-Free Calibration Guarantees for Histogram Binning without Sample Splitting

Chirag Gupta [1] **Aaditya Ramdas** [1]

Abstract

We prove calibration guarantees for the popular histogram binning (also called uniform-mass binning) method of Zadrozny and Elkan (2001). Histogram binning has displayed strong practical performance, but theoretical guarantees have only been shown for sample split versions that avoid 'double dipping' the data. We demonstrate that the statistical cost of sample splitting is practically significant on a credit default dataset. We then prove calibration guarantees for the original method that double dips the data, using a certain Markov property of order statistics. Based on our results, we make practical recommendations for choosing the number of bins in histogram binning. In our illustrative simulations, we propose a new tool for assessing calibration—validity plots—which provide more information than an ECE estimate.

1. Introduction

In classification, the goal is to learn a model that uses observed feature measurements to make a class prediction on the categorical outcome. However, for safety-critical areas such as medicine and finance, a single class prediction might be insufficient and reliable measures of confidence or certainty may be desired. Such uncertainty quantification is often provided by predictors that produce not just a class label, but a probability distribution over the labels. If the predicted probability distribution is consistent with observed empirical frequencies of labels, the predictor is said to be calibrated (Dawid, 1982).

In this paper we study the problem of calibration for binary classification; let $\mathcal{X}$ and $\mathcal{Y} = \{0, 1\}$ denote the feature and label spaces. We focus on the recalibration or post-hoc calibration setting, a standard statistical setting where the goal is to recalibrate existing ('pre-learnt') classifiers that are powerful and (statistically) efficient for classification accuracy, but do not satisfy calibration properties out-of-the-box. This setup is popular for recalibrating pre-trained deep nets. For example, Guo et al. (2017, Figure 4) demonstrated that a pre-learnt ResNet is initially miscalibrated, but can be effectively post-hoc calibrated. In the case of binary classification, the pre-learnt model can be an arbitrary predictor function that provides a classification 'score' $g \in \mathcal{G}$, where $\mathcal{G}$ is the space of all measurable functions from $\mathcal{X} \to [0, 1]$. Along with g, we are given access to a calibration dataset of size $n \in \mathbb{N}$, $\mathcal{D}_n = \{(X_i, Y_i)\}_{i \in [n]}$, drawn independently from a distribution $P \equiv P_X \times P_{Y|X}$. The goal is to define a calibrator $H : \mathcal{G} \times (\mathcal{X} \times [0, 1])^n \to \mathcal{G}$, that 'recalibrates' g to an approximately calibrated predictor $H(g, \mathcal{D}_n)$ (formally defined shortly). We denote $H(g, \mathcal{D}_n)$ as h. *All probabilities in this paper are conditional on g and thus conditional on the data on which g is learnt.*

Let $\mathbb{E}[\cdot]$ denote the expectation operator associated with P, interpreted marginally or conditionally depending on the context. The predictor h is said to be perfectly calibrated if $\mathbb{E}[Y \mid h(X)] = h(X)$ (almost surely). While perfect calibration is impossible in finite samples, we desire a framework to make transparent claims about how close h is to being perfectly calibrated. The following notion proposed by Gupta et al. (2020) defines a calibrator that provides *probably approximate calibration* for chosen levels of approximation $\varepsilon \in (0, 1)$ and failure $\alpha \in (0, 1)$. For brevity, we skip the qualification 'probably approximate'.

Definition 1 (Marginal calibration[1])**.** A calibrator $H : (g, \mathcal{D}_n) \mapsto h$ is said to be (ε, α)-marginally calibrated if for every predictor $g \in \mathcal{G}$ and distribution P over $\mathcal{X} \times [0, 1]$,

$$P(|\mathbb{E}[Y|h(X)] - h(X)| \leqslant \varepsilon) \geqslant 1 - \alpha. \qquad (1)$$

The above probability is taken over both X and $\mathcal{D}_n$ since $h = H(g, \mathcal{D}_n)$ contains the randomness of $\mathcal{D}_n$. The qualification *marginal* signifies that the inequality $|\mathbb{E}[Y \mid h(X)] - h(X)| \leqslant \varepsilon$ may not hold conditioned on

[1]Carnegie Mellon University. Correspondence to: Chirag Gupta <chiragg@cmu.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]This definition is unrelated to that of Gneiting et al. (2007, Definition 1c), where marginal calibration refers to an asymptotic notion of calibration in the regression setting.

X or $h(X)$, but holds only on *average*. We now define a more stringent conditional notion of calibration, which requires that approximate calibration hold simultaneously (or conditionally) for every value of the prediction.

Definition 2 (Conditional calibration)**.** A calibrator $H : (g, \mathcal{D}_n) \mapsto h$ is (ε, α)-conditionally calibrated if for every predictor $g \in \mathcal{G}$ and distribution P over $\mathcal{X} \times [0, 1]$,

$$P(\forall r \in \text{Range}(h), |\mathbb{E}\left[Y \mid h(X) = r\right] - r| \leqslant \varepsilon) \geqslant 1 - \alpha. \quad (2)$$

In contrast to (1), the Pr above is only over $\mathcal{D}_n$. Evidently, if H is conditionally calibrated, it is also marginally calibrated. The conditional calibration property (2) has a PAC-style interpretation: with probability $1 - \alpha$ over $\mathcal{D}_n$, h satisfies the following deterministic property:

$$\forall r \in \text{Range}(h), |\mathbb{E}\left[Y \mid h(X) = r\right] - r| \leqslant \varepsilon. \quad (3)$$

Marginal calibration does not have such an interpretation; we cannot infer from (1) a statement of the form "with probability $1 - \gamma$ over $\mathcal{D}_n$, h satisfies $\cdots$ ".

Marginal and conditional calibration assess the truth of the event $\mathbb{1}\{|\mathbb{E}\left[Y \mid h(X)\right] - h(X)| \leqslant \varepsilon\}$ for a given ε. Instead we can consider bounding the expected value of $|\mathbb{E}\left[Y \mid h(X)\right] - h(X)|$ for $X \sim P_X$. This quantity is known as the expected calibration error.

Definition 3 (Expected Calibration Error (ECE))**.** For $p \in [1, \infty)$, the ℓ_p-ECE of a predictor h is

$$\ell_p\text{-ECE}(h) = \left(\mathbb{E}_X|\mathbb{E}\left[Y \mid h(X)\right] - h(X)|^p\right)^{1/p}. \quad (4)$$

Note that the expectation above is only over $X \sim P_X$ and not over $\mathcal{D}_n$. We can ask for bounds on the ECE of $h = H(g, \mathcal{D}_n)$ that hold with high-probability or in-expectation over the randomness in $\mathcal{D}_n$. The conditional calibration property (3) for h implies a bound on the ℓ_p-ECE for every p, as formalized by the following proposition which also relates ℓ_p-ECE for different p.

Proposition 1. *For any predictor h and $1 \leqslant p \leqslant q < \infty$,*

$$\ell_p\text{-ECE}(h) \leqslant \ell_q\text{-ECE}(h). \quad (5)$$

Further, if (3) holds, then ℓ_p-ECE$(h) \leqslant \varepsilon, \forall p \in [1, \infty)$.

The proof (in Appendix A) is a straightforward application of Hölder's inequality. Informally, one can interpret the L.H.S. of (3) as the ℓ_∞-ECE of h so that (5) holds for $1 \leqslant p \leqslant q \leqslant \infty$. Thus conditional calibration is the strictest calibration property we consider: if H is (ε, α)-conditionally calibrated, then (a) H is (ε, α)-marginally calibration and (b) with probability $1 - \alpha$, ℓ_p-ECE$(h) \leqslant \varepsilon$.

Example 1. We verify Proposition 1 on a simple example, which also helps build intuition for the various notions of calibration. Suppose h takes just two values: $P(h(X) = 0.2) = 0.9$ and $P(h(X) = 0.8) = 0.1$. Let $\mathbb{E}\left[Y \mid h(X) = 0.2\right] = 0.3$ and $\mathbb{E}\left[Y \mid h(X) = 0.8\right] = 0.6$. Then ℓ_1-ECE$(h) = 0.11 < \ell_2$-ECE$(h) \approx 0.114$. Marginal calibration (1) for $H(\cdot, \cdot) \equiv h$ is satisfied for $(\varepsilon \geqslant 0.1, \alpha \leqslant 0.9)$, while the conditional calibration requirement (3) is only satisfied for $\varepsilon \geqslant 0.2$.

In this paper, we show that the histogram binning method of Zadrozny and Elkan (2001), described shortly, is calibrated in each of the above senses (marginal and conditional calibration; high-probability and in-expectation bounds on ECE), if the number of bins is chosen appropriately.

Some safety-critical domains may require calibration methods that are robust to the data-generating distribution. We refer to Definitions 1 and 2 as distribution-free (DF) guarantees since they are required to hold for all distributions over (X, Y) without restriction. This paper is in the DF setting: the only assumption we make is that the calibration data $\mathcal{D}_n$ and (X, Y) are independent and identically distributed (i.i.d.). Gupta et al. (2020, Theorem 3) showed that if H is DF marginally calibrated with a meaningful value of ε (formally, ε can be driven to zero as sample size grows to infinity), then H must necessarily produce only discretized predictions (formally, Range(h) must be at most countable). We refer to such H as 'binning methods' — this emphasizes that H essentially partitions the sample-space into a discrete number of 'bins' and provides one prediction per bin (see Proposition 1 (Gupta et al., 2020)). Since our goal is DF calibration, we focus on binning methods.

1.1. Prior Work on Binning

Binning was initially introduced in the calibration literature for assessing calibration. Given a continuous scoring function h, if we wish to plot a reliability diagram (Niculescu-Mizil and Caruana, 2005; Sanders, 1963) or compute an ECE estimate (Miller, 1962; Naeini et al., 2015; Sanders, 1963), then h must first be discretized using binning. A common binning scheme used for this purpose is 'fixed-width binning', where $[0, 1]$ is partitioned into $B \in \mathbb{N}$ intervals (called bins) of width $1/B$ each and a single prediction is assumed for every bin. For example, if $B = 10$, then the width of each bin is 0.1, and if (say) $h(x) \in [0.6, 0.7)$ then the prediction is assumed to be 0.65.

Gupta et al. (2020, Theorem 3) showed that some kind of binning is in fact necessary to *achieve* DF calibration. The first binning method for calibration was proposed by Zadrozny and Elkan (2001) to calibrate a naive Bayes classifier. Their procedure is as follows. First, the interval $[0, 1]$ is partitioned into $B \in \mathbb{N}$ bins using the histogram of the $g(X_i)$ values, to ensure that each bin has the same number of calibration points (plus/minus one). Thus the bins have nearly 'uniform (probability) mass'. Then, the calibration points are assigned to bins depending on the interval to which the score $g(X_i)$ belongs to, and the probability that $Y = 1$ is estimated for each bin as the average of the

observed Y_i-values in that bin. This average estimates the 'bias' of the bin. The binning scheme and the bias estimates together define h. A slightly modified version of this procedure is formally described in Algorithm 1.

While Algorithm 1 was originally called histogram binning, it has also been referred to as uniform-mass binning in some works. In the rest of this paper, we use the latter terminology. Specifically, we refer to it as UMD, short for Uniform-Mass-Double-dipping. This stresses that the same data is used twice, both to determine inter-bin boundaries and to calculate intra-bin biases. UMD continues to remain a competitive benchmark in empirical work (Guo et al., 2017; Naeini et al., 2015; Roelofs et al., 2020), but no finite-sample calibration guarantees have been shown for it. Some asymptotic consistency results for a histogram regression algorithm closely related to UMD were shown by Parthasarathy and Bhattacharya (1961) (see also the work by Lugosi and Nobel (1996)). Zadrozny and Elkan (2002) proposed another popular binning method based on isotonic regression, for which some non-DF analyses exist (see Dai et al. (2020) and references therein). Recently, two recalibration methods closely related to UMD have been proposed, along with some theoretical guarantees that rely on sample-splitting — scaling-binning (Kumar et al., 2019) and sample split uniform-mass binning (Gupta et al., 2020).

In the scaling-binning method, the binning is performed on the output of another continuous recalibration method (such as Platt scaling (Platt, 1999)), and the bias for each bin is computed as the average of the output of the scaling procedure in that bin. This is unlike other binning methods, where the bias of each bin is computed as the average of the true outputs Y_i in that bin. Kumar et al. (2019, Theorem 4.1) showed that under some assumptions on the scaling class (which includes injectivity), the ECE of the sample split scaling-binning procedure is ε-close to $\sqrt{2}\,\ell_2$-ECE of the scaling procedure if, roughly, $n = \Omega(\log B/\varepsilon^2)$. However, the results of Gupta et al. (2020, Section 3.3) imply that there exist data distributions on which any injective scaling procedure itself has trivial ECE.

In sample split uniform-mass binning, the first split of the data is used to define the bin boundaries so that the bins are balanced. The second split of the data is used for estimating the bin biases, using the average of the Y_i-values in the bin. We refer to this version as UMS, for Uniform-Mass-Sample-splitting. Gupta et al. (2020, Theorem 5) showed that UMS is (ε, α)-marginally calibrated if (roughly) $n = \Omega(B \log(B/\alpha)/\varepsilon^2)$. To the best of our knowledge, this is the only known DF guarantee for a calibration method. However, in Section 2 we demonstrate that the constants in this guarantee are quite conservative, and the loss in performance due to sample splitting is practically significant on a real dataset.

1.2. Our Contribution

We show tight DF calibration guarantees for the original method proposed by Zadrozny and Elkan (2001), UMD. While the existing theoretical analyses rely on sample splitting (Gupta et al., 2020; Kumar et al., 2019), it has been observed in experiments that double dipping to perform both bin formation and bias estimation on the same data leads to excellent practical performance (Guo et al., 2017; Kumar et al., 2019; Roelofs et al., 2020; Zadrozny and Elkan, 2001). Our work fills this gap in theory and practice.

We exploit a certain Markov property of order statistics, which are a set of classical, elegant results that are not well known outside of certain subfields of statistics (for one exposition of the Markov property, see Arnold et al. (2008, Chapter 2.4)). The strength of these probabilistic results is not widely appreciated — judging by their non-appearance in the ML literature — nor have they had implications for any modern AI applications that we are aware of. Thus, we consider it a central contribution of this work to have recognized that these mathematical tools can be brought to bear in order to shed light on a contemporary ML algorithm.

A simplified version of the Markov property is as follows: for order statistics $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)}$ of samples $\{Z_i\}_{i\in[n]}$ drawn i.i.d from any absolutely continuous distribution Q, and any indices $1 < i < j \leqslant n$, we have that

$$Z_{(j)} \perp Z_{(i-1)}, Z_{(i-2)}, \ldots, Z_{(1)} \mid Z_{(i)}.$$

For example, given the empirical median M, the points to its left are conditionally independent of the points to its right. Further each of these have a distribution that is identical to that of i.i.d. draws from $Z \sim Q$ when restricted to $Z < M$ (or $Z > M$). The implication is that if we form bins using the order statistics of the scores as the bin boundaries, then (a) the points within any bin are independent of the points outside that bin, and (b) conditioned on being in a given bin, say B_i, the points in the bin are i.i.d. with distribution $Q_{Z|Z\in B_i}$. When we split a calibration sample $\mathcal{D}$ and use one part $\mathcal{D}_1$ for binning and the other $\mathcal{D}\backslash\mathcal{D}_1$ for estimating bin probabilities, the points in $\mathcal{D}\backslash\mathcal{D}_1$ that belong to B_i are also conditionally i.i.d. with distribution $Q_{Z|Z\in B_i}$, which is exactly what we accomplished without sample splitting. In short, the Markov property allows us to 'double dip' the data, i.e., use the same data for binning and estimating within-bin probabilities.

Organization. Section 2 motivates our research problem by showing that UMS is sample-inefficient both in theory and practice. Empirical evidence is provided through a novel diagnostic tool called validity plots (Section 2.1). Section 3 presents UMD formally along with its analysis (main results in Theorems 3 and 4). Section 4 contains illustrative simulations. Proofs are in the supplement.

2. Sample Split Uniform-Mass Binning is Inefficient

The DF framework encourages development of algorithms that are robust to arbitrarily distributed data. At the same time, the hope is that the DF guarantees are adaptive to real data and give meaningful bounds in practice. In this section, we assess if the practical performance of uniform-mass-sample-splitting (UMS) is well explained by its DF calibration guarantee (Gupta et al., 2020). As far as we know, this is the only known DF guarantee for a calibration method. However, we demonstrate that the guarantee is quite conservative. Further, we demonstrate that sample splitting leads to a drop in performance on a real dataset.

Suppose we wish to guarantee $(\varepsilon, \alpha) = (0.1, 0.1)$-marginal calibration with $B = 10$ bins using UMS. We unpacked the DF calibration bound for UMS, and computed that to guarantee $(0.1, 0.1)$-marginal calibration with 10 bins, roughly $n \geqslant 17500$ is required. The detailed calculations can be found in Appendix B. This sample complexity seems conservative for a binary classification problem. In Section 2.2, we use an illustrative experiment to show that the n required to achieve the desired level of calibration is indeed much lower than 17500. Our experiment uses a novel diagnostic tool called validity plots, introduced next.

2.1. Validity Plots

Validity plots assess the marginal calibration properties of a calibration method by displaying estimates of the LHS of (1) as ε varies. Define the function $V : [0,1] \to [0,1]$ given by $V(\varepsilon) = \Pr(|\mathbb{E}[Y \mid h(X)] - h(X)| \leqslant \varepsilon)$. By definition of V, H is $(\varepsilon, 1 - V(\varepsilon))$-marginally calibrated for every ε. For this reason, we call the graph of V, $\{(\varepsilon, V(\varepsilon)) : \varepsilon \in [0,1]\}$, as the 'validity curve'. (The term "curve" is used informally since V may have jumps.) Note the following neat relationship between the ℓ_1-ECE and the area-under-the-curve (AUC) of the validity curve:

$$\begin{aligned}\mathbb{E}[\ell_1\text{-ECE}(h)] &= \mathbb{E}[|\mathbb{E}[Y \mid h(X)] - h(X)|] \\ &= \int_0^1 P(|\mathbb{E}[Y \mid h(X)] - h(X)| > \varepsilon)\, d\varepsilon \\ &= 1 - \int_0^1 P(|\mathbb{E}[Y \mid h(X)] - h(X)| \leqslant \varepsilon)\, d\varepsilon \\ &= 1 - \int_0^1 V(\varepsilon)\, d\varepsilon = 1 - \text{AUC(validity curve)}.\end{aligned}$$

A validity plot is a finite sample estimate of the validity curve on a single calibration set $\mathcal{D}_n$ and test set $\mathcal{D}_{\text{test}}$. We now outline the steps for constructing a validity plot. First, h is learned using $\mathcal{D}_n$ and g. Next, if h is not a binning method, it must be discretized through binning in order to enable estimation of $\mathbb{E}[Y \mid h(X)]$. This is identical to the binning step required by plugin ECE estimators and reliability diagrams. For example, one can use fixed-width binning as described in the first paragraph of Section 1.1. In this paper, we empirically assess only binning methods, and so an additional binning step is not necessary. Next, the empirical distribution on $\mathcal{D}_{\text{test}}$ is used as a proxy for the true distribution of (X, Y), to estimate $V(\varepsilon)$:

$$\widehat{V}(\varepsilon) = \frac{\sum_{(X_i,Y_i)\in\mathcal{D}_{\text{test}}} \mathbb{1}\{|\mathbb{E}_{\widehat{P}}[Y|h(X)=h(X_i)] - h(X_i)| \leqslant \varepsilon\}}{|\mathcal{D}_{\text{test}}|}, \text{ where}$$
$$\mathbb{E}_{\widehat{P}}[Y \mid h(X) = h(x)] \equiv \frac{\sum_{(X_i,Y_i)\in\mathcal{D}_{\text{test}}} Y_i \mathbb{1}\{h(X_i)=h(x)\}}{\sum_{(X_i,Y_i)\in\mathcal{D}_{\text{test}}} \mathbb{1}\{h(X_i)=h(x)\}}. \quad (6)$$

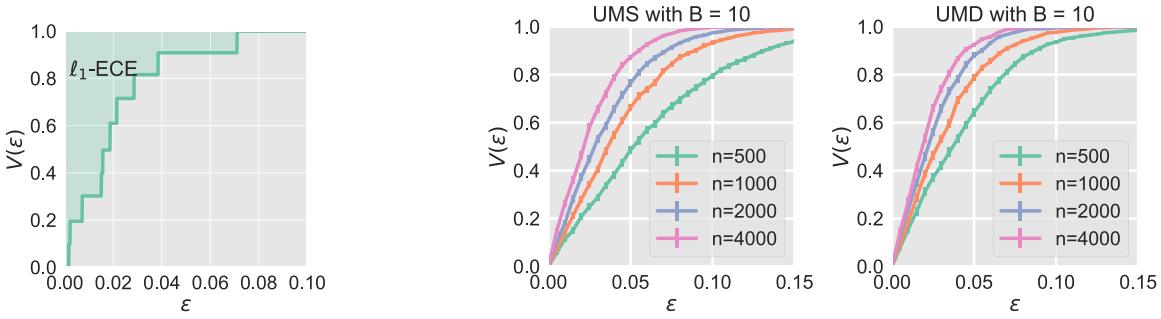
For different values of $\varepsilon \in [0,1]$ on the X-axis, the estimate of $V(\varepsilon)$ is plotted on the Y-axis to form the validity plot. Like the AUC of a validity curve corresponds to $\mathbb{E}[\ell_1\text{-ECE}]$, the AUC of a validity plot corresponds to the plugin ℓ_1-ECE estimate (Naeini et al., 2015). (There may be small differences in practice since we draw the validity plot for a finite grid of values in $[0,1]$.) Thus validity plots convey the ℓ_1-ECE estimate and more.

Figure 1a displays an illustrative validity plot for a binning method with $B = 10$. V is a right-continuous step function with at most $|\text{Range}(h)| \leqslant B$ many discontinuities. Each ε for which there is a discontinuity in V corresponds to a bin that has $|\mathbb{E}[Y \mid h(X) = r] - r| = \varepsilon$, and the incremental jump in the value of V, $V(\varepsilon) - V(\varepsilon^-)$, corresponds to the fraction of test points in that bin. Figure 1a was created using UMD, and thus each jump corresponds to roughly a $1/B = 0.1$ fraction of the test points. The ε values for the bins are approximately $10^{-3} \cdot (1.5, 2, 8, 16, 17, 19, 22, 29, 39, 71)$.

Unlike reliability diagrams (Niculescu-Mizil and Caruana, 2005), validity plots do not convey the predictions $h(X)$ to which the ε values correspond to, or the direction of miscalibration (whether $h(X)$ is higher or lower than $\mathbb{E}[Y \mid h(X)]$). On the other hand, validity plots convey the bin frequencies for every bin without the need for a separate histogram (such as the top panel in Niculescu-Mizil and Caruana (2005, Figure 1)). In our view, validity plots also 'collate' the right entity; we can easily read off from a validity plot practically meaningful statements such as "for 90% of the test points, the miscalibration is at most 0.04".

We can create a smoother validity plot that better estimates V by using multiple runs based on subsampled or bootstrapped data. To do this, for every $\varepsilon \in [0,1]$, $\widehat{V}(\varepsilon)$ is computed separately for each run and the mean value is plotted as the estimate of $V(\varepsilon)$. In our simulations, we always perform multiple runs, and also show $\pm$std-dev-of-mean in the plot. Figure 1b displays such validity plots (further details presented in the following subsection).

It is well known that plugin ECE estimators for a binned method are biased towards slightly overestimating the ECE (e.g., see Bröcker (2012); Kumar et al. (2019); Widmann et al. (2019)). For the same reasons, $\widehat{V}(\varepsilon)$ is a biased un-



(a) An illustrative validity plot. We can read off that marginal calibration is achieved for $(\varepsilon, \alpha) = (0.04, 0.1)$ and $(0.03, 0.2)$. The ℓ_1-ECE estimate is roughly 0.023.

(b) Validity plots comparing UMD and UMS on the CREDIT dataset. The plots show that UMD has higher validity $V(\varepsilon)$ for the same values of n, ε, and thus lower ℓ_1-ECE. For example, for $n = 1000$ and $\varepsilon = 0.05$, UMS has $V(\varepsilon) \approx 0.63$, while UMD has $V(\varepsilon) \approx 0.79$.

Figure 1. Validity plots display estimates of $V(\varepsilon) = P(|\mathbb{E}[Y \mid h(X)] - h(X)| \leqslant \varepsilon)$ as ε varies. Validity plots are described in Section 2.1. The experimental setup for Figure 1b is presented in Section 2.2.

derestimate of $V(\varepsilon)$. In other words, the validity plot is on average below the true validity curve. The reason for this bias is that to estimate ECE as well as to create validity plots, we compute $\left|\mathbb{E}_{\hat{P}}[Y \mid h(X)] - h(X)\right|$ which can be written as $|\mathbb{E}[Y \mid h(X)] + \text{mean-zero-noise} - h(X)|$. On average, the noise term will lead to overestimating $|\mathbb{E}[Y \mid h(X)] - h(X)|$. However, the noise term is small if there is enough test data (if n_b is the number of test points in bin b, then the noise term is $O(\sqrt{1/n_b})$ w.h.p.). Further, it is highly unlikely that the noise will help some methods and hurts others. Thus validity plots can be reliably used to make inferences on the relative performance of different calibration methods. While there exist unbiased estimators for $(\ell_2\text{-ECE})^2$ (Bröcker, 2012; Widmann et al., 2019), we are not aware of any unbiased ℓ_1-ECE estimators. If such an estimator is proposed in the future, the same technique will also improve validity plots.

2.2. Comparing UMS and UMD using Validity Plots

Figure 1b uses validity plots to assess UMS and UMD on CREDIT, a UCI credit default dataset[2]. The task is to accurately predict the probability of default. The experimental protocol is as follows. The entire feature matrix is first normalized[3]. CREDIT has 30K (30,000) samples which are randomly split (once for the entire experiment) into splits (A, B, C) = (10K, 5K, 15K). First, g is formed by training a logistic regression model on split A and then re-scaling the learnt model using Platt scaling on split B (Platt scaling before binning was suggested by Kumar et al. (2019); we also observed that this helps in practice). Next, the calibration set $\mathcal{D}_n$ is formed by randomly subsampling n ($\leqslant$10K) points from split C (without replacement). From the remaining points in split C, a test set of size 5K is subsampled (without replacement). The entire subsampling from split C is repeated 100 times to create 100 different calibration and test sets. For a given subsample, UMS/UMD with $B = 10$ is trained on the calibration set (with 50:50 sample splitting for UMS), and $\widehat{V}(\varepsilon)$ for every ε is estimated on the test set. Finally, the (mean$\pm$std-dev-of-mean) of $\widehat{V}(\varepsilon)$ is plotted with respect to ε. This experimental setup assesses marginal calibration for a fixed g, in keeping with our post-hoc calibration setting.

The validity plot in Figure 1b (left) indicates that the desired $(0.1, 0.1)$-marginal calibration is achieved by UMS with just $n = 1000$. Contrast this to $n \geqslant 17500$ required by the theoretical bound, as computed in Appendix B. In fact, $n = 4000$ nearly achieves $(0.05, 0.1)$-marginal calibration. This gap occurs because the analysis of UMS is complex, with constants stacking up at each step.

Next, consider the validity plot for UMD in Figure 1b (right). By avoiding sample splitting, UMD achieves $(0.1, 0.1)$-marginal calibration at $n = 500$. In Section 3 we show that $n \geqslant 1500$ is provably sufficient for $(0.1, 0.1)$-marginal calibration and $n \geqslant 2900$ is sufficient for $(0.1, 0.1)$-conditional calibration. Some gap in theory and practice is expected since the theoretical bound is DF, and thus applies no matter how anomalous the data distribution is. However, the gap is much smaller compared to UMS, due to a clean analysis. In Section 4, we illustrate that the gap nearly vanishes for larger n. Section 4 also introduces the related concept of *conditional* validity plots that assess conditional calibration.

3. Distribution-Free Analysis of UMD

Define the random variables $S = g(X)$; $S_i = g(X_i)$ for $i \in [n]$, called scores. Let $(S, Y) \sim Q$ and

[2] Yeh and Lien (2009); https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[3] using Python's sklearn.preprocessing.scale

$S \sim Q_S$. In binning, we wish to use the calibration data $\{(S_i, Y_i)\}_{i\in[n]} \sim Q^n$ to (a) define a binning function $\mathcal{B} : [0,1] \to [B]$ for some number of bins $B \in \mathbb{N}$, and (b) estimate the biases in the bins $\{\Pi_b := \mathbb{E}[Y \mid \mathcal{B}(S) = b]\}_{b\in[B]}$. We denote the bias estimates as $\widehat{\Pi}_b$. The approximately calibrated function is then defined as $h(\cdot) = \widehat{\Pi}_{\mathcal{B}(\cdot)}$.

Suppose the number of recalibration points is $n \approx 150$. In the absence of known properties of the data (i.e., in the DF setting), it seems reasonable to have $B = 1$ and define $H(g, \mathcal{D}_n)$ as the constant function $h(\cdot) := n^{-1}\sum_{i=1}^n Y_i$. Formally, $n = 150$ leads to the following Hoeffding-based confidence interval: with probability at least 0.9, $\left|n^{-1}\sum_{i=1}^n Y_i - \mathbb{E}Y\right| \leqslant \sqrt{\log(2/0.1)/(2\cdot 150)} \approx 0.1$. In other words, if $n = 150$, H satisfies $(0.1, 0.1)$-marginal calibration. Of course, having a single bin completely destroys sharpness of h, but it's an instructive special case.

Suppose now that $n \approx 300$, and we wish to learn a non-constant h using two bins. If g is informative, we hope that $\mathbb{E}[Y \mid g(X) = \cdot]$ is roughly a monotonically increasing function. In light of this belief, it seems reasonable to choose a threshold t and identify the two bins as: $g(X) \leqslant t$ and $g(X) > t$. A natural choice for t is $M = \text{Median}(S_1, \ldots, S_n)$ since this ensures that both bins get the same number of points (plus/minus one). This is the motivation for UMD. In this case, h and $\widehat{\Pi}$ are defined as,

$$h(\cdot) := \begin{cases} \widehat{\Pi}_1 := \text{Average}(Y_i : S_i \leqslant M) \text{ if } g(\cdot) \leqslant M \\ \widehat{\Pi}_2 := \text{Average}(Y_i : S_i > M) \text{ if } g(\cdot) > M. \end{cases} \tag{7}$$

Suppose M were the true median of Q_S instead of the empirical median. Then h has a calibration guarantee obtained by applying a Bernoulli concentration inequality separately for both bins and using a union bound (this is done formally by Gupta et al. (2020, Theorem 4)). In UMS, we try to emulate the true median case by using one split of the data to estimate the median. $\widehat{\Pi}$ is then computed on the second (independent) split of the data, and concentration inequalities can be used to provide calibration guarantees.

UMD does not sample split: in equation (7) above, M is computed using the same data that is later used to estimate $\widehat{\Pi}$. On the face of it, this double dipping eliminates the independence of the Y_i values required to apply a concentration inequality. However, we show that the independence structure can be retained if UMD is slightly modified. This subtle modification is to remove a single point from the bias estimation, namely the Y_i corresponding to the median M. (In comparison, in UMS we typically remove a fixed ratio of n.) The informal argument is as follows.

For simplicity, suppose Q_S is absolutely continuous (with respect to the Lebesgue measure), so that the S_i's are almost surely distinct, and suppose that the number of samples is odd: $n = 2m + 1$. Denote the ordered scores as $S_{(1)} < S_{(2)} < \ldots < S_{(n)}$ and let $Y_{(i)}$ denote the label corresponding to the score $S_{(i)}$. Thus $\widehat{\Pi}_1 = m^{-1}\sum_{i=1}^m Y_{(i)}$ and $M = S_{(m+1)}$. Clearly, $(S_{(i)}, Y_{(i)})$ is not independent of $S_{(m+1)}$ for any i. However, it turns out that the following property is true: conditioned on $S_{(m+1)}$, the unordered values $\{(S_{(i)}, Y_{(i)})\}_{i\in[m]}$ can be viewed as m *independent* samples identically distributed as (S, Y), given $S < S_{(m+1)}$. (Note that (S, Y) is an unseen and independent random variable.) Thus, we can use Hoeffding's inequality to assert: $P(|\mathbb{E}[Y \mid M, S < M] - \widehat{\Pi}_1| \geqslant \varepsilon \mid M, S < M) \leqslant 2\exp\left(-2m\varepsilon^2\right)$. This can be converted to a calibration guarantee on the first bin. The same bound can be shown if $S > M$, for the estimate $\widehat{\Pi}_2 = m^{-1}\sum_{i=m+1}^{2m+1} Y_{(i)}$. Using a union bound gives a calibration guarantee that holds for both bins simultaneously, which in turn gives conditional calibration.

In the following subsection, we show some key lemmas regarding the order statistics of the S_i's. These lemmas formalize what was argued above: *careful double dipping does not eliminate the independence structure*. In Section 3.2, we formalize the modified UMD algorithm, and prove that it is DF calibrated. Based on the guarantee for the modified version, Corollary 1 finally shows that the original UMD itself is DF calibrated.

Simplifying assumption. In the following analysis, we assume that $g(X)$ is absolutely continuous with respect to the Lebesgue measure, and thus has a probability density function (pdf). This assumption is made at no loss of generality, for reasons discussed in Appendix C.1.

3.1. Key Lemmas on Order Statistics

Consider two indices $i, j \in [n]$. The score S_i is not independent of the order statistic $S_{(j)}$. However, it turns out that conditioned on $S_{(j)}$, the distribution of S_i given $S_i < S_{(j)}$, is identical to the distribution of an unseen score S, given $S < S_{(j)}$. The following lemmas (both proved in Appendix A) state versions of this fact that are useful for our analysis of UMD.

We first set up some notation. S is assumed to have a pdf, denoted as f. For some $1 \leqslant l < u \leqslant n$, consider the set of indices $\{i : S_{(l)} < S_i < S_{(u)}\}$, and index them arbitrarily as $\{t_1, t_2, \ldots, t_{u-l-1}\}$. This is just an indexing and not an ordering; in particular it is not necessary that $S_{t_1} = S_{(l+1)}$. For $j \in \{l+1, \ldots, u-1\}$, define $S_{\{j\}} = S_{t_{j-l}}$. Thus the set $\{S_{\{j\}} : j \in \{l+1, \ldots, u-1\}\}$ corresponds to the *unordered* S_i values between $S_{(l)}$ and $S_{(u)}$.

Lemma 1. *Fix $l, u \in [n]$ such that $l < u$. The conditional density of the unordered S_i values between the order statistics $S_{(l)}, S_{(u)}$, $f(S_{\{l+1\}}, \ldots, S_{\{u-1\}} \mid S_{(l)}, S_{(u)})$, is identical to the density of independent $S'_i \sim Q_S$, conditional on lying between $S_{(l)}, S_{(u)}$:*

$$f(S'_1, \ldots, S'_{u-l-1} \mid S_{(l)}, S_{(u)}, S_{(l)} < \{S'_i\}_{i\in[u-l-1]} < S_{(u)}).$$

Algorithm 1 UMD: Uniform-mass binning without sample splitting

1: **Input:** Scoring function $g : \mathcal{X} \to [0, 1]$, #bins B, calibration data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$
2: **Output:** Approximately calibrated function h
3: $(S_1, S_2, \ldots, S_n) \leftarrow (g(X_1), g(X_2), \ldots, g(X_n))$
4: $(S_{(1)}, S_{(2)}, \ldots, S_{(n)}) \leftarrow$ order-stats$(S_1, S_2, \ldots, S_n)$
5: $(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}) \leftarrow (Y_1, Y_2, \ldots, Y_n)$ ordered as per the ordering of $(S_{(1)}, S_{(2)}, \ldots, S_{(n)})$
6: $\Delta \leftarrow (n + 1)/B$
7: $\widehat{\Pi} \leftarrow$ empty array of size B
8: $A \leftarrow$ 0-indexed array$([0, \lceil\Delta\rceil, \lceil 2\Delta\rceil, \ldots, n + 1])$
9: **for** $b \leftarrow 1$ **to** B **do**
10: $\quad l \leftarrow A_{b-1}$
11: $\quad u \leftarrow A_b$
12: $\quad \widehat{\Pi}_b \leftarrow$ Mean$(Y_{(l+1)}, Y_{(l+2)}, \ldots, Y_{(u-1)})$
13: **end for**
14: $(S_{(0)}, S_{(n+1)}) \leftarrow (0, 1)$
15: $h(\cdot) \leftarrow \sum_{b=1}^{B} \mathbb{1}\left\{S_{(A_{b-1})} \leqslant g(\cdot) < S_{(A_b)}\right\} \widehat{\Pi}_b$

In the final analysis, $S_{(l)}$ and $S_{(u)}$ will represent the scores at consecutive bin boundaries, which define the binning scheme. Lemma 2 is similar to Lemma 1, but with conditioning on all bin boundaries (order statistics) simultaneously. To state it concisely, define $S_{(0)} := 0$ and $S_{(n+1)} := 1$ as fixed hypothetical 'order statistics'.

Lemma 2. *Fix any $B - 1$ indices $k_1, k_2, \ldots k_{B-1}$ such that $0 = k_0 < k_1 < \ldots < k_{B-1} < k_B = n + 1$. For any $b \in [B]$, the conditional density of the unordered S_i values between the order statistics $S_{(k_{b-1})}, S_{(k_b)}$, $f(S_{\{k_{b-1}+1\}}, \ldots, S_{\{k_b-1\}} \mid S_{(k_0)}, \ldots, S_{(k_B)})$, is identical to the conditional density*

$$f(S'_1, \ldots, S'_{k_b - k_{b-1} - 1} \mid S_{(k_0)}, \ldots, S_{(k_B)}, \quad \text{for every } i \in [k_b - k_{b-1} - 1],\ S_{(k_{b-1})} < S'_i < S_{(k_b)}))$$

of independent random variables $S'_i \sim Q_S$.

3.2. Main Results

UMD is described in Algorithm 1 (in the description, $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote the floor and ceiling operators respectively). UMD takes input $(g, \mathcal{D}_n)$ and outputs h. There is a small difference between UMD as stated and the proposal by Zadrozny and Elkan (2001). The original version also uses the calibration points that define the bin boundaries for bias estimation — this corresponds to replacing line 12 with

line 12: $\widehat{\Pi}_b \leftarrow$ Mean$(Y_{(l+1)}, \ldots, Y_{(u-1)}, \boldsymbol{Y_{(u)}})$, for $b < B$.

The two algorithms are virtually the same; after stating the calibration guarantee for UMD, we show the result for the original proposal as a corollary.

By construction, every bin defined by UMD has at least $\lfloor n/B\rfloor - 1$ many points for mean estimation. Thus, UMD effectively 'uses' only $B - 1$ points for bin formulation using quantile estimation. We prove the following calibration guarantee for UMD in Appendix A.

Theorem 3. *Suppose $g(X)$ is absolutely continuous with respect to the Lebesgue measure and $n \geqslant 2B$. UMD is (ε, α)-conditionally calibrated for any $\alpha \in (0, 1)$ and*

$$\varepsilon = \sqrt{\frac{\log(2B/\alpha)}{2(\lfloor n/B\rfloor - 1)}}. \tag{8}$$

Further, for every distribution P, w.p. $1 - \alpha$ over the calibration data $\mathcal{D}_n$, for all $p \in [1, \infty)$, ℓ_p-ECE$(h) \leqslant \varepsilon$.

Note that since UMD is (ε, α)-conditionally calibrated, it is also (ε', α)-conditionally calibrated for any $\varepsilon' \in (\varepsilon, 1)$. The absolute continuity requirement for $g(X)$ can be removed with a randomization trick discussed in Section C.1, to make the result fully DF. The proof sketch is as follows. Given the bin boundaries, the scores in each bin are independent, as shown by Lemma 2. We use this to conclude that the Y_i values in each bin b are independent and distributed as Bern$(\mathbb{E}[Y \mid \mathcal{B}(X) = b])$. The average of the Y_i values thus concentrates around $\mathbb{E}[Y \mid \mathcal{B}(X) = b]$. Since each bin has at least $(\lfloor n/B\rfloor - 1)$ points, Hoeffding's inequality along with a union bound across bins gives conditional calibration for the value of ε in (8).

The convenient property that every bin has at least $\lfloor n/B\rfloor - 1$ calibration points for mean estimation is not satisfied deterministically even if we used the true quantiles of $g(X)$. In fact, as long as $B = o(n)$, the ε in (8) approaches the ε we would get if all the data was used for bias estimation, with at least $\lfloor n/B\rfloor$ points in each bin:

$$\text{if } B = o(n),\ \lim_{n\to\infty} \left|\sqrt{\frac{\log(2B/\alpha)}{2(\lfloor n/B\rfloor - 1)}} - \sqrt{\frac{\log(2B/\alpha)}{2(\lfloor n/B\rfloor)}}\right| = 0.$$

In comparison to the clean proof sketch above, UMS requires a tedious multi-step analysis:

1. Suppose the sizes of the two splits are n_1 and n_2. Performing reliable quantile estimation on the first split of the data requires $n_1 = \Omega(B \log(B/\alpha))$ (Kumar et al., 2019, Lemma 4.3)).
2. The estimated quantiles have the guarantee that the *expected* number of points falling into a bin, on the second split is $\geqslant n_2/2B$. A high probability bound is used to lower bound the actual number of points in each bin. This lower bound is $(n_2/2B) - \sqrt{n_2 \log(2B/\alpha)/2}$ (Gupta et al., 2020, Theorem 5).

This multi-step analysis leads to a loose bound due to constants stacking up, as discussed in Section 2.

A guarantee for the original UMD procedure follows as an immediate corollary of Theorem 3. This is because the modification to line 12 can change every estimate $\widehat{\Pi}_b$ by

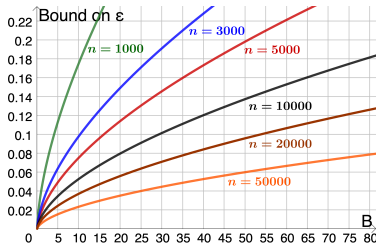


Figure 2. Plots displaying the relationship (8) between ε and B for $\alpha = 0.1$ and different values of n. Some indicative suggestions based on the plot: if $n = 1$K, choose $B = 5$ (gives $\varepsilon \leqslant 0.12$); if $n = 5$K, choose $B = 10$ (gives $\varepsilon \leqslant 0.08$); if $n = 20$K, choose $B = 22$ (gives $\varepsilon \leqslant 0.06$).

at most $1/(\lfloor n/B \rfloor)$ due to the following fact regarding averages: for any $b \in \mathbb{N}, a \in \{0, 1, \ldots, b\}$,

$$\max\left(\left|\tfrac{a}{b+1} - \tfrac{a}{b}\right|, \left|\tfrac{a+1}{b+1} - \tfrac{a}{b}\right|\right) \leqslant \tfrac{1}{b+1}. \tag{9}$$

Using (9), we prove the following corollary in Appendix A.

Corollary 1. *Suppose $g(X)$ is absolutely continuous with respect to the Lebesgue measure and $n \geqslant 2B$. The original UMD algorithm (Zadrozny and Elkan, 2001) is (ε, α)-conditionally calibrated for any $\alpha \in (0, 1)$ and*

$$\varepsilon = \sqrt{\frac{\log(2B/\alpha)}{2(\lfloor n/B \rfloor - 1)}} + \frac{1}{\lfloor n/B \rfloor}. \tag{10}$$

Further, for every distribution P, w.p. $1 - \alpha$ over the calibration data $\mathcal{D}_n$, for all $p \in [1, \infty)$, ℓ_p-ECE$(h) \leqslant \varepsilon$.

As claimed in Section 2.2, if $(n, \alpha, B) = (2900, 0.1, 10)$, (10) gives $\varepsilon < 0.1$. The difference between (10) and (8) is small. For example, we computed that if $\varepsilon \leqslant 0.1$, $\alpha \leqslant 0.5$, $B \geqslant 5$, then (8) requires $n/B \geqslant 150$, and thus the additional term in (10) is at most 0.007. Likewise, in practice, we expect both versions to perform similarly.

At the end of the day, a practitioner may ask: "Given n points for recalibration, how should I use Theorem 3 to decide B?" Smaller B gives better bounds on ε, but larger B implicitly means that the h learnt is sharper. As n becomes higher, one may like to have higher sharpness (higher B), but at the same time more precise calibration (lower ε and thus lower B). We provide a (subjective) discussion on how to balance these two requirements.

First, we suggest fixing a rough domain-dependent probability of failure α. Since the dependence of ε on α in (8) is $\log(1/\alpha)$, small changes in α do not affect ε too much. Typically, 10-20% failure rate is acceptable, so let us set $\alpha = 0.1$. (For a highly sensitive domain, one can set $\alpha = 0.01$.) Then, constraint (8) roughly translates to $\varepsilon = \sqrt{B \log(20B)/2n}$. For a fixed n, this is a relationship between ε and B, that can be plotted as a curve with B as the independent parameter and ε as the dependent parameter. Finally, one can eyeball the curve to identify a B. We plot such curves in Figure 2 for a range of values of n. The caption shows examples of how one can choose B to balance calibration (small ε) and sharpness (high B).

While (ε, α)-conditional calibration implies (ε, α)-marginal calibration, we expect to have marginal calibration with smaller ε. Such an improved guarantee can be shown if the bin biases $\widehat{\Pi}_b$ estimated by Algorithm 1 are distinct. In Appendix C, we propose a randomized version of UMD (Algorithm 2) which guarantees uniqueness of the bin biases. Algorithm 2 satisfies the following calibration guarantee (proved in Appendix A).

Theorem 4. *Suppose $n \geqslant 2B$ and let $\delta > 0$ be an arbitrarily small randomization parameter. Algorithm 2 is (ε_1, α)-marginally and (ε_2, α)-conditionally calibrated for any $\alpha \in (0, 1)$,*

$$\varepsilon_1 = \sqrt{\tfrac{\log(2/\alpha)}{2(\lfloor n/B \rfloor - 1)}} + \delta,\ \varepsilon_2 = \sqrt{\tfrac{\log(2B/\alpha)}{2(\lfloor n/B \rfloor - 1)}} + \delta. \tag{11}$$

Further, for every distribution P, (a) w.p. $1 - \alpha$ over the calibration data $\mathcal{D}_n$, for all $p \in [1, \infty)$, ℓ_p-ECE$(h) \leqslant \varepsilon_2$, and (b) $\mathbb{E}_{\mathcal{D}_n}[\ell_p\text{-ECE}(h)] \leqslant \sqrt{B/2n} + \delta$ for all $p \in [1, 2]$.

In the proof, we use the law of total expectation to avoid taking a union bound in the marginal calibration result; this gives a $\sqrt{\log(2/\alpha)}$ term in ε_1 instead of the $\sqrt{\log(2B/\alpha)}$ in ε_2. Theorem 4 also does not require absolute continuity of $g(X)$. As claimed in Section 2.2, if $(n, \alpha, B) = (1500, 0.1, 10)$, (11) gives $\varepsilon_1 < 0.1$ (for small enough δ).

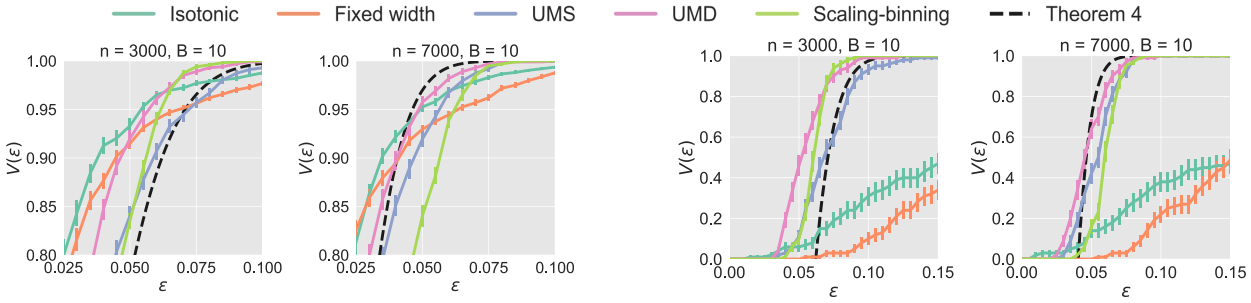
4. Simulations

We perform illustrative simulations on the `CREDIT` dataset with two goals: (a) to compare the performance of UMD to other binning methods and (b) to show that the guarantees we have shown are reasonably tight, and thus, practically useful.[4] In addition to validity plots, which assess marginal calibration, we use conditional validity plots, that assess conditional calibration. Let $V : [0, 1] \to [0, 1]$ be given by $V(\varepsilon) = P(\forall r \in \text{Range}(h), |\mathbb{E}[Y \mid h(X) = r] - r| \leqslant \varepsilon)$. Given a test set $\mathcal{D}_{\text{test}}$, we first compute $\mathbb{E}_{\widehat{P}}[Y \mid h(X) = h(x)]$ (defined in (6)), and then estimate $V(\varepsilon)$ as

$$\widehat{V}(\varepsilon) = \mathbb{1}\{\max_{(X_i, Y_i) \in \mathcal{D}_{\text{test}}} \left|\mathbb{E}_{\widehat{P}}[Y \mid h(X) = h(X_i)] - h(X_i)\right| \leqslant \varepsilon\}.$$

For a single $\mathcal{D}_n$ and $\mathcal{D}_{\text{test}}$, $\widehat{V}(\varepsilon)$ is either 0 or 1. Thus to estimate $V(\varepsilon)$, we average $\widehat{V}(\varepsilon)$ across multiple calibration and test sets. The mean$\pm$std-dev-of-mean of the $\widehat{V}(\varepsilon)$ values are plotted as ε varies. This gives us a conditional validity plot. It is easy to see that the conditional validity plot is uniformly dominated by the (marginal) validity plot.

[4] Relevant code can be found at https://github.com/aigen/df-posthoc-calibration



(a) Marginal validity plots.

(b) Conditional validity plots.

Figure 3. UMD performs competitively on the CREDIT dataset. The guarantee of Theorem 4 closely matches empirical behavior.

The experimental protocol for CREDIT is described in Section 2.2. In our experiments, we used the randomized version of UMD (Algorithm 2). Figure 3 presents validity plots for UMD, UMS, fixed-width binning, isotonic regression, scaling-binning, along with the Theorem 4 curve for $n = 3K$ and $n = 7$K. In Appendix D, we also present plots for $n = 1K$ and $n = 5$K. Fixed-width binning refers to performing binning with equally spaced bins $([0, 1/B), \ldots, [1-1/B, 1])$. UMS uses a 50:50 split of the calibration data. We do not rescale in scaling-binning, since it is already done on split B (for all compared procedures) — instead the comparison is between averaging the predictions of the scaling method (as is done in scaling-binning), against averaging the true outputs in each bin (as is done by all other methods). To have a fair comparison, we use double dipping for scaling-binning (thus scaling-binning and UMD are identical except what is being averaged). We make the following observations:

- Isotonic regression and fixed-width binning perform well for marginal calibration, but fail for conditional calibration. This is because both these methods tend to have bins with skewed masses, leading to small ε in bins with many points, and high ε in bins with few points.
- Scaling-binning is competitive with UMD for $n = 3$K, $\varepsilon > 0.05$. If $n = 7$K or $\varepsilon \leqslant 0.05$, UMD outperforms scaling-binning. In Appendix D, we show that for $n = 1$K, scaling-binning is nearly the best method.
- UMD always performs better than UMS, and the performance of UMD is almost perfectly explained by the theoretical guarantee. Paradoxically, for $n = 7$K, the theoretical curve *crosses* the validity plot for UMD. This can occur since validity plots are based on a finite sample estimate of $\mathbb{E}[Y \mid h(X)]$, and the estimation error leads to slight *underestimation* of validity. This phenomenon is the same as the bias of plugin ECE estimators, and is discussed in detail in the last paragraph of Section 2.1. The curve-crossing shows that Theorem 4 is so precise that 5K test points are insufficient to verify it.

Overall, our experiment indicates that UMD performs competitively in practice and our theoretical guarantee closely explains its performance.

5. Conclusion

We used the Markov property of order statistics to prove distribution-free calibration guarantees for the popular uniform-mass binning method of Zadrozny and Elkan (2001). We proposed a novel assessment tool called validity plots, and used this tool to demonstrate that our theoretical bound closely tails empirical performance on a UCI credit default dataset. To the best of our knowledge, we demonstrated for the first time that it is possible to show informative calibration guarantees for binning methods that double dip the data (to both estimate bins and the probability of $Y = 1$ in a bin). Popular calibration methods such as isotonic regression (Zadrozny and Elkan, 2002), probability estimation trees (Provost and Domingos, 2003), random forests (Breiman, 2001) and Bayesian binning (Naeini et al., 2015) perform exactly this style of double dipping. We thus open up the exciting possibility of providing DF calibration guarantees for one or more of these methods.

Another recent line of work for calibration in data-dependent groupings, termed as multicalibration, uses a discretization step similar to fixed-width binning (Hébert-Johnson et al., 2018). Our uniform-mass binning techniques can potentially be extended to multicalibration. A number of non-binned methods for calibrating neural networks have displayed good performance on some tasks (Guo et al., 2017; Kull et al., 2017; Lakshminarayanan et al., 2017). However, the results of Gupta et al. (2020) imply that these methods cannot have DF guarantees. Examining whether they have guarantees under some (weak) distributional assumptions is also interesting future work.

Acknowledgments

We wish to thank Sasha Podkopaev, Anish Sevekari, Saurabh Garg, Elan Rosenfeld, and the anonymous ICML reviewers, for comments on an earlier version of the paper. CG also thanks the instructors and students in the 'Writing in Statistics' course at CMU for valuable feedback.

References

Mohammad Ahsanullah, Valery B Nevzorov, and Mohammad Shakil. *An introduction to order statistics*, volume 8. Springer, 2013.

Barry C Arnold, Narayanaswamy Balakrishnan, and Haikady Navada Nagaraja. *A first course in order statistics*. SIAM, 2008.

Leo Breiman. Random forests. *Machine learning*, 45(1): 5–32, 2001.

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

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Ran Dai, Hyebin Song, Rina Foygel Barber, and Garvesh Raskutti. The bias of isotonic regression. *Electronic journal of statistics*, 14(1):801, 2020.

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

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

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

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *Advances in Neural Information Processing Systems*, 2020.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, 2018.

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.

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

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

Gábor Lugosi and Andrew Nobel. Consistency of data-driven histogram methods for density estimation and classification. *Annals of Statistics*, 24(2):687–706, 1996.

Robert G Miller. Statistical prediction by discriminant analysis. In *Statistical Prediction by Discriminant Analysis*, pages 1–54. Springer, 1962.

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

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

KR Parthasarathy and PK Bhattacharya. Some limit theorems in regression theory. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 91–102, 1961.

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.

Foster Provost and Pedro Domingos. Tree induction for probability-based ranking. *Machine learning*, 52(3): 199–215, 2003.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. Mitigating bias in calibration error estimation. *arXiv preprint arXiv:2012.08668*, 2020.

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

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

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

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

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