Verified Uncertainty Calibration

Ananya Kumar, Percy Liang, Tengyu Ma
Department of Computer Science
Stanford University
{ananya, pliang, tengyuma}@cs.stanford.edu

Abstract

Applications such as weather forecasting and personalized medicine demand models that output calibrated probability estimates—those representative of the true likelihood of a prediction. Most models are not calibrated out of the box but are recalibrated by post-processing model outputs. We find in this work that popular recalibration methods like Platt scaling and temperature scaling, are (i) less calibrated than reported and (ii) current techniques cannot estimate how miscalibrated they are. An alternative method, histogram binning, has measurable calibration error but is sample inefficient—it requires $O(B/\epsilon^2)$ samples, compared to $O(1/\epsilon^2)$ for scaling methods, where B is the number of distinct probabilities the model can output. To get the best of both worlds, we introduce the scaling-binning calibrator, which first fits a parametric function that acts like a baseline for variance reduction and then bins the function values to actually ensure calibration. This requires only $O(1/\epsilon^2 + B)$ samples. We then show that methods used to estimate calibration error are suboptimal—we prove that an alternative estimator introduced in the meteorological community requires fewer samples—samples proportional to $\sqrt{B}$ instead of B. We validate our approach with multiclass calibration experiments on CIFAR-10 and ImageNet, where we obtain a 35% lower calibration error than histogram binning and, unlike scaling methods, guarantees on true calibration.

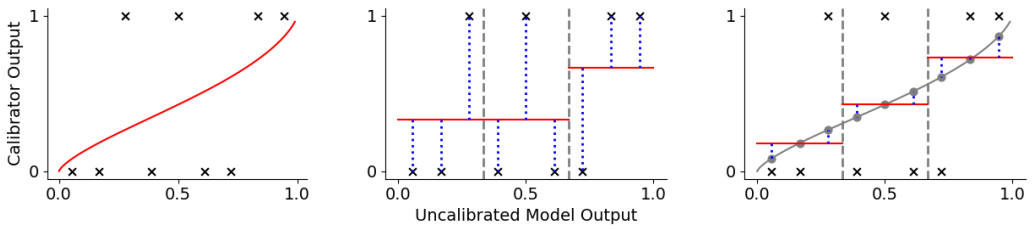
1 Introduction

The probability that a system outputs for an event should reflect the true frequency of that event: if an automated diagnosis system says 1,000 patients have cancer with probability 0.1, approximately 100 of them should indeed have cancer. In this case we say the model is uncertainty calibrated. The importance of this notion of calibration has been emphasized in personalized medicine [1], meteorological forecasting [2, 3, 4, 5, 6] and natural language processing applications [7, 8]. As most modern machine learning models, such as neural networks, do not output calibrated probabilities out of the box [9, 10, 11], recalibration methods take the output of an uncalibrated model, and transform it into a calibrated probability. *Scaling* approaches for recalibration—Platt scaling [12], isotonic regression [13], and temperature scaling [9]—are widely used and require very few samples, but do they actually produce calibrated probabilities?

We discover that these methods are less calibrated than reported. Past work approximates a model's calibration error using a finite set of bins. We show that by using more bins, we can uncover a higher calibration error for models on CIFAR-10 and ImageNet. We show that a fundamental limitation with approaches that output a continuous range of probabilities is that their true calibration error may never be measurable with a finite number of bins (Example 3.2).

An alternative approach, histogram binning [10], outputs probabilities from a finite set. Histogram binning can produce a model that is calibrated, and unlike scaling methods we can measure its calibration error, but it is sample inefficient. In particular, the number of samples required to calibrate scales linearly with the number of distinct probabilities the model can output, B [14], which can

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.



(a) Platt scaling. (b) Histogram binning. (c) The scaling-binning calibrator.

Figure 1: Visualization of the 3 recalibration approaches. The black crosses are the ground truth labels, and the red lines are the output of the recalibration methods. Platt Scaling (Figure 1a) fits a function to the recalibration data, but its calibration error is not measurable. Histogram binning (Figure 1b) outputs the average label in each bin. The scaling-binning calibrator (Figure 1c) fits a function $g \in \mathcal{G}$ to the recalibration data and then *takes the average of the function values (the gray circles)* in each bin. The function values have lower variance than the labels, as visualized by the blue dotted lines, which is why our approach has lower variance.

be large particularly in the multiclass setting where B typically scales with the number of classes. Recalibration sample efficiency is crucial—we often want to recalibrate our models in the presence of domain shift [15] or recalibrate a model trained on simulated data, and may have access to only a small labeled dataset from the target domain.

To get the sample efficiency of Platt scaling and the verification guarantees of histogram binning, *we propose the scaling-binning calibrator* (Figure 1c). Like scaling methods, we fit a simple function $g \in \mathcal{G}$ to the recalibration dataset. We then bin the input space so that an equal number of inputs land in each bin. In each bin, we output the average of the g values in that bin—these are the gray circles in Figure 1c. In contrast, histogram binning outputs the average of the label values in each bin (Figure 1b). The motivation behind our method is that the g values in each bin are in a narrower range than the label values, so when we take the average we incur less of an estimation error. If $\mathcal{G}$ is well chosen, our method requires $O(\frac{1}{\epsilon^2} + B)$ samples to achieve calibration error ϵ instead of $O(\frac{B}{\epsilon^2})$ samples for histogram binning, where B is the number of model outputs (Theorem 4.1). Note that in prior work, binning the outputs of a function was used for evaluation and without any guarantees, whereas in our case it is used for the method itself and we show improved sample complexity.

Next, we turn to the question of estimating the calibration error. Prior work in machine learning [7, 9, 15, 16, 17] directly estimates each term in the calibration error from samples (Definition 5.1). The sample complexity of this plugin estimator scales linearly with B. An alternative estimator introduced in the meteorological literature [18, 19] reduces the bias of the plugin estimator; *we show that it has sample complexity that scales with* $\sqrt{B}$. We prove that it achieves this by leveraging error cancellations across bins.

We run multiclass calibration experiments on CIFAR-10 [20] and ImageNet [21]. The objective is to minimize the mean-squared error, also known as the Brier score [22], subject to a calibration error budget [5]. We show that the scaling-binning calibrator achieves a better calibration error than histogram binning, while allowing us to measure the true calibration error. For example, we get a *35% lower calibration error on CIFAR-10* and a *5x lower calibration error on ImageNet* than histogram binning if $B = 100$.

2 Setup and background

2.1 Binary classification

Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ be the label space where $\mathcal{Y} = \{0, 1\}$ in the binary classification setting. Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be random variables denoting the input and label, given by an unknown joint distribution $P(X, Y)$. Expectations are taken over all random variables unless otherwise specified.

Suppose we have a model $f : \mathcal{X} \to [0, 1]$ where the output of the model represents the model's confidence that the label is 1. As f may not be calibrated, we define the calibration error, which

examines the difference between the model's probability and the true probability given the model's output. If the calibration error is 0 then the model is perfectly calibrated.

Definition 2.1 (Calibration error)**.** *For $p \geq 1$, the ℓ_p calibration error of $f : \mathcal{X} \to [0, 1]$ is given by:*

$$\ell_p\text{-CE}(f) = \Big(\mathbb{E}\big[\ \mid f(X) - \mathbb{E}[Y \mid f(X)] \mid^p \big]\Big)^{1/p} \tag{1}$$

The L^2 calibration error, $(\ell_2\text{-CE}(f))^2$, [2, 3, 4, 7, 15, 16, 17, 18] is most commonly used but the ℓ_1 and ℓ_∞ calibration errors are also used in the literature [9, 23, 24].

Calibration alone is not sufficient: consider an image dataset containing 50% dogs and 50% cats. If f outputs 0.5 on all inputs, f is calibrated but not very useful. We often also wish to minimize the mean-squared error—also known as the Brier score—as defined below.

Definition 2.2. *The mean-squared error of $f : \mathcal{X} \to [0, 1]$ is given by* $\text{MSE}(f) = \mathbb{E}[(f(X) - Y)^2]$.

We often want to minimize the MSE subject to a calibration budget [5, 25]. Of course, these are not orthogonal because $\text{MSE} = 0$ implies perfect calibration—in fact the MSE is the sum of the L^2 calibration error and a "sharpness" term [2, 4, 16].

2.2 Multiclass classification

While calibration in binary classification is well-studied, it's less clear what to do for multiclass, where multiple definitions abound, differing in their strength. In the multiclass setting, $\mathcal{Y} = [K]$, where $[K] = \{1, \ldots, K\}$ and $f : \mathcal{X} \to [0, 1]^K$ outputs a confidence measure for each class in $[K]$.

Definition 2.3 (Top-label calibration error)**.** *The L^2 top-label calibration error examines the difference between the model's probability for its top prediction and the true probability of that prediction given the model's output:*

$$L^2\text{-TCE}(f) = \mathbb{E}\Big[\Big(\mathbb{P}\big(Y = \underset{j \in [K]}{\textit{argmax}}\, f(X)_j \mid \max_{j \in [K]} f(X)_j\big) - \max_{j \in [K]} f(X)_j\Big)^2\Big] \tag{2}$$

We would often like the model to be calibrated on less likely predictions as well—imagine that a medical diagnosis system says there is a 50% chance a patient has a benign tumor, a 10% chance she has an aggressive form of cancer, and a 40% chance she has one of a long list of other conditions. We would like the model to be calibrated on all of these predictions so we define the marginal calibration error which examines, *for each class*, the difference between the model's probability and the true probability of that class given the model's output.

Definition 2.4 (Marginal calibration error)**.** *Let $w_k \in [0, 1]$ denote how important calibrating class k is, where $w_k = \mathbb{P}(Y = k)$ if all classes are equally important. The L^2 marginal calibration error is:*

$$L^2\text{-MCE}(f) = \sum_{k=1}^{K} w_k \mathbb{E}\big[(f(X)_k - \mathbb{P}(Y = k \mid f(X)_k))^2\big] \tag{3}$$

Note that prior works [9, 15, 17] often claim to perform multiclass calibration but only measure top-label calibration—[24] shows that temperature scaling [9] scores worse than vector scaling on a marginal calibration metric, even though it has lower top-label calibration error.

For notational simplicity, our theory focuses on the binary classification setting. We can transform top-label calibration into a binary calibration problem—the model outputs a probability corresponding to its top prediction, and the label represents whether the model gets it correct or not. Marginal calibration can be transformed into K one-vs-all binary calibration problems where for each $k \in [K]$ the model outputs the probability associated with the k-th class, and the label represents whether the input belongs to the k-th class or not [13]. We look at both top-label calibration and marginal calibration in our experiments. Other notions of multiclass calibration include joint calibration (which requires the entire probability *vector* to be calibrated) [2, 6] and event-pooled calibration [16].

2.3 Recalibration

Since most machine learning models do not output calibrated probabilities out of the box [9, 10] recalibration methods take the output of an uncalibrated model, and transform it into a calibrated

probability. That is, given a trained model $f : \mathcal{X} \to \mathcal{Z}$, let $Z = f(X)$. We are given recalibration data $T = \{(z_i, y_i)\}_{i=1}^n$ independently sampled from $P(Z, Y)$, and we wish to learn a recalibrator $g : \mathcal{Z} \to [0, 1]$ such that $g \circ f$ is well-calibrated.

Scaling methods, for example Platt scaling [12], output a function $g = \text{argmin}_{g \in \mathcal{G}} \sum_{(z,y) \in T} \ell(g(z), y)$, where $\mathcal{G}$ is a model family, $g \in \mathcal{G}$ is continuous, and l is a loss function, for example the log loss or mean-squared error. The advantage of such methods is that they converge very quickly since they only fit a single parameter.

Histogram binning first constructs a set of bins (intervals) that partitions $[0, 1]$, formalized below.

Definition 2.5 (Binning schemes)**.** *A binning scheme $\mathcal{B}$ of size B is a set of B intervals $I_1, \ldots, I_B$ that partitions $[0, 1]$. Given an input $z \in [0, 1]$ if $z \in I_j$ let $\beta(z) = j$ be the interval z lands in.*

The bins are typically chosen such that either $I_1 = [0, \frac{1}{B}], I_2 = (\frac{1}{B}, \frac{2}{B}], \ldots, I_B = (\frac{B-1}{B}, 1]$ or so that each bin contains an equal number of z_i values in the recalibration data [10, 9]. Histogram binning then outputs the average y_i value in each bin.

3 Is Platt scaling calibrated?

In this section, we show why methods like Platt scaling and temperature scaling are (i) less calibrated than reported and (ii) it is difficult to tell how miscalibrated they are. The issue is that it is very difficult to measure the calibration error of models that output a continuous range of values. We show, both theoretically and with experiments on CIFAR-10 and ImageNet, why the calibration error of such models is *underestimated*. We defer proofs to Appendix C.

The key to estimating the calibration error is estimating the conditional expectation $\mathbb{E}[Y \mid f(X)]$ where $f(X)$ is a continuous value; without smoothness assumptions on $\mathbb{E}[Y \mid f(X)]$ (that cannot be verified in practice), this is impossible. This is analogous to the difficulty of measuring the mutual information between two continuous signals [26].

To approximate the ℓ_p calibration error, prior work bins the output of f into B intervals. The calibration error in each bin is estimated as the difference between the average value of $f(X)$ and Y in that bin. Note that the binning here is for evaluation only, whereas in histogram binning is used for the recalibration method itself. We formalize the notion of this binned calibration error below.

Definition 3.1. *The binned version of f outputs the average value of f in each bin I_j:*

$$f_{\mathcal{B}}(x) = E[f(X) \mid f(X) \in I_j] \qquad \text{where } x \in I_j \tag{4}$$

Given $\mathcal{B}$, the binned calibration error of f is simply ℓ_p-CE$(f_{\mathcal{B}})$. A simple example shows that using binning to estimate the ℓ_p calibration error can severely underestimate the true ℓ_p calibration error, even for $p = 1$, the average calibration error.

Example 3.2. *For any binning scheme $\mathcal{B}$, $p \in \mathbb{Z}^+$, and continuous bijective function $f : [0, 1] \to [0, 1]$, there exists a distribution P over X, Y s.t. ℓ_p-CE$(f_{\mathcal{B}}) = 0$ but ℓ_p-CE$(f) \geq 0.49$. Note that for all f, $0 \leq \ell_p$-CE$(f) \leq 1$.*

The intuition is that in each interval I_j in $\mathcal{B}$, the model could underestimate the true probability $\mathbb{E}[Y \mid f(X)]$ half the time, and overestimate the probability half the time. So if we average over the entire bin the model appears to be calibrated, even though it is very uncalibrated. The formal construction is in Appendix C.

Next, we show that given a function f, its binned version always has lower calibration error.

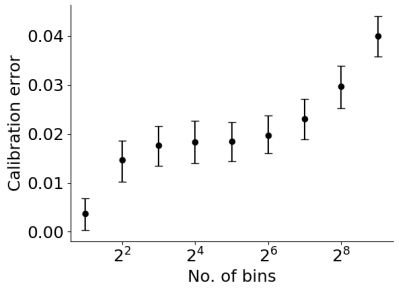
Proposition 3.3 (Binning underestimates error)**.** *Given binning scheme $\mathcal{B}$ and model $f : \mathcal{X} \to [0, 1]$, we have:*

$$\ell_p\text{-CE}(f_{\mathcal{B}}) \leq \ell_p\text{-CE}(f)$$

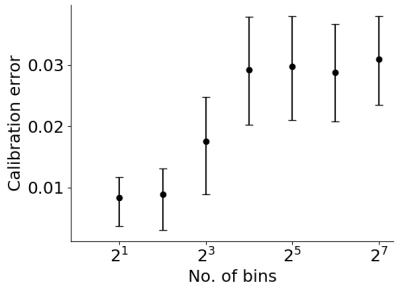
The proof is by Jensen's inequality (Appendix C). Intuitively, averaging a model's prediction within a bin allows errors at different parts of the bin to cancel out with each other.

3.1 Experiments

Our experiments on ImageNet and CIFAR-10 suggest that previous work reports numbers which are lower than the actual calibration error of their models. Recall that binning lower bounds the



(a) ImageNet.



(b) CIFAR-10.

Figure 2: Binned ℓ_2 calibration errors of a recalibrated VGG-net model on CIFAR-10 and ImageNet with 90% confidence intervals. The binned calibration error increases as we increase the number of bins. This suggests that binning cannot be reliably used to measure the true calibration error.

calibration error. We cannot compute the actual calibration error but if we use a 'finer' set of bins then we get a tighter lower bound on the calibration error.

As in [9], our model's objective was to output the top predicted class and a confidence score associated with the prediction. For ImageNet, we started with a trained VGG16 model with an accuracy of 64.3%. We split the validation set into 3 sets of size $(20000, 5000, 25000)$. We used the first set of data to recalibrate the model using Platt scaling, the second to select the binning scheme $\mathcal{B}$ so that each bin contains an equal number of points, and the third to measure the binned calibration error. We calculated 90% confidence intervals for the binned calibration error using 1,000 bootstrap resamples and performed the same experiment with varying numbers of bins.

Figure 2a shows that as we increase the number of bins on ImageNet, the measured calibration error is higher and this is statistically significant. For example, if we use 15 bins as in [9], we would think the ℓ_2 calibration error is around 0.02 when the calibration error is at least twice as high. Figure 2b shows similar findings for CIFAR-10, and in Appendix B we show that our findings hold even if we use ℓ_1 calibration error and alternative binning strategies.

4 The scaling-binning calibrator

In the previous section we saw that the problem with scaling methods is we cannot estimate their calibration error. The upside of scaling methods is that if the function family has at least one function that can achieve calibration error ϵ, they require $O(1/\epsilon^2)$ samples to reach calibration error ϵ, while histogram binning requires $O(B/\epsilon^2)$ samples where B can be large. Can we get a method that is sample efficient to calibrate and one where it's possible to estimate the calibration error? To achieve this, we propose the scaling-binning calibrator (Figure 1c) where we first fit a scaling function, and then bin the outputs of the scaling function.

4.1 Algorithm

We split the recalibration data T of size n into 3 sets: T_1, T_2, T_3. The scaling-binning calibrator, illustrated in Figure 1, outputs $\hat{g_{\mathcal{B}}}$ such that $\hat{g_{\mathcal{B}}} \circ f$ has low calibration error:

Step 1 (Function fitting): The first step is to select $g = \text{argmin}_{g \in \mathcal{G}} \sum_{(z,y) \in T_1} (y - g(z))^2$.

Step 2 (Bin construction): The second step is to construct a suitable binning scheme. We choose the bins so that an equal number of $g(z_i)$ in T_2 land in each bin I_j for each $j \in \{1, \ldots, B\}$, instead of equal width bins used in [9]–our choice of bins is essential for our bounds.

Step 3 (Discretization): The third step is to discretize g, by outputting the average g value in each bin—these are the gray circles in Figure 1c. Let $\mu(S) = \frac{1}{|S|} \sum_{s \in S} s$ denote the mean of a set of values S. Let $\hat{\mu}[j] = \mu(\{g(z_i) \mid g(z_i) \in I_j \wedge (z_i, y_i) \in T_3\})$ be the mean of the $g(z_i)$ values that landed in the j-th bin. Recall that if $z \in I_j$, $\beta(z) = j$ is the interval z lands in. Then we set $\hat{g_{\mathcal{B}}}(z) = \hat{\mu}[\beta(g(z))]$—that is we simply output the mean value in the bin that $g(z)$ falls in.

4.2 Analysis

We now show that the scaling-binning calibrator requires $O(B + 1/\epsilon^2)$ samples to calibrate, and in Section 5 we show that we can efficiently measure its calibration error. For the main theorem, we make some standard regularity assumptions on $\mathcal{G}$ which we formalize in Appendix D. Our result is a generalization result—we show that if $\mathcal{G}$ contains some g^* with low calibration error, then our method is *at least* almost as well-calibrated as g^* after a certain number of samples.

Theorem 4.1 (Calibration bound)**.** *Assume regularity conditions on $\mathcal{G}$ (finite parameters, injectivity, Lipschitz-continuity, consistency, twice differentiability). Given $\delta \in (0, 1)$, there is a constant c such that* for all *B, ϵ, with $n \geq c\Big(B \log B + \frac{\log B}{\epsilon^2}\Big)$ samples, the scaling-binning calibrator finds $\hat{g}_{\mathcal{B}}$ with L^2-CE$(\hat{g}_{\mathcal{B}}) \leq 2 \min_{g \in \mathcal{G}} L^2$-CE$(g) + \epsilon^2$, with probability $\geq 1 - \delta$.*

Note that our method can potentially be better calibrated than g^*, because we bin the outputs of the scaling function, which reduces its calibration error (Proposition 3.3). While binning worsens the 'sharpness' and can increase the mean-squared error of the model, in Proposition D.4 we show that if we use many bins, binning the outputs cannot increase the mean-squared error by much.

We prove Theorem 4.1 in Appendix D but give a sketch here. Step 1 of our algorithm is Platt scaling, which simply fits a function g to the data—standard results in asymptotic statistics show that g converges in $O(\frac{1}{\epsilon^2})$ samples. As usual with asymptotic analysis, the constant c can depend arbitrarily on δ and $\mathcal{G}$—making these dependencies explicit is an important direction for future research.

Step 3, where we bin the outputs of g, is the main step of the algorithm. If we had infinite data, Proposition 3.3 showed that the binned version $g_{\mathcal{B}}$ has lower calibration error than g, so we would be done. However we do not have infinite data—the core of our proof is to show that the empirically binned $\hat{g_{\mathcal{B}}}$ converges to $g_{\mathcal{B}}$ in $O(B + \frac{1}{\epsilon^2})$ samples, instead of $O(B + \frac{B}{\epsilon^2})$ samples in histogram binning. The intuition is in Figure 1—the $g(z_i)$ values in each bin (gray circles in Figure 1c) are in a narrower range than the y_i values (black crosses in Figure 1b) so when we take the average, we incur less estimation error. The perhaps surprising part is that we are estimating B numbers with $\widetilde{O}(1/\epsilon^2)$ samples. In fact, there may be a small number of bins where the $g(z_i)$ values are not in a narrow range, but our proof still shows that the overall estimation error is small.

In Section 3 we showed that current techniques cannot measure the calibration error of scaling methods, in contrast we show that we *can* measure the calibration error of our calibrator. Recall that we chose our bins so that each bin had an equal proportion of points in the recalibration set. Lemma 4.3 will show that this property approximately holds in the population as well. This will allow us to estimate the calibration error efficiently (Theorem 5.4).

Definition 4.2 (Well-balanced binning)**.** *Given a binning scheme $\mathcal{B}$ of size B, and $\alpha \geq 1$. We say $\mathcal{B}$ is α-well-balanced if for all j,*

$$\frac{1}{\alpha B} \leq \mathbb{P}(Z \in I_j) \leq \frac{\alpha}{B}$$

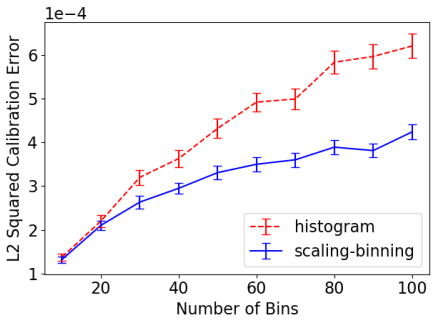
Lemma 4.3. *For universal constant c, if $n \geq c\big(B \log \frac{B}{\delta}\big)$, with probability at least $1 - \delta$, the binning scheme $\mathcal{B}$ we chose is 2-well-balanced.*

While the way we choose bins is not novel [10], we believe the guarantees around it are—not all binning schemes in the literature allow us to efficiently estimate the calibration error, for example the binning scheme in [9] does not. Our proof of Lemma 4.3 is in Appendix D. The core challenge is that applying Chernoff bounds or a standard VC dimension argument gives us a loose bound and tells us we need $O(B^2 \log \frac{B}{\delta})$ samples. We use a discretization argument to prove the result.

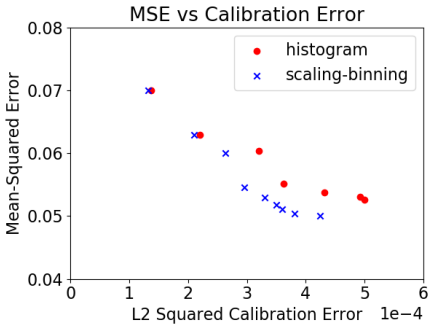
4.3 Experiments

Our experiments on CIFAR-10 and ImageNet show that in the low-data regime, for example when we use ≤ 1000 data points to recalibrate, the scaling-binning calibrator produces models with much lower calibration error than histogram binning. The model's objective was to output a confidence score associated with each class, where we calibrated each class separately as in [13], used B bins per class and evaluated calibration using the marginal calibration error (Definition 2.4).

We describe our experimental protocol for CIFAR-10. The CIFAR-10 validation set has 10,000 data points. We sampled, with replacement, a recalibration set of 1,000 points. We ran either the



(a) Effect of number of bins on L^2 calibration error.



(b) Tradeoff between calibration and MSE.

Figure 3: (**Left**) Recalibrating using 1,000 data points on CIFAR-10, the scaling-binning calibrator achieves lower L^2 calibration error than histogram binning, especially when the number of bins B is large. (**Right**) For a fixed calibration error, the scaling-binning calibrator allows us to use more bins. This results in models with more predictive power which can be measured by the mean-squared error. Note the Y-axis range is $[0.04, 0.08]$ to zoom into the relevant region.

scaling-binning calibrator (we fit a sigmoid in the function fitting step) or histogram binning and measured the marginal calibration error on the entire set of 10K points. We repeated this entire procedure 100 times and computed mean and 90% confidence intervals, and we repeated this varying the number of bins B. Figure 3a shows that the scaling-binning calibrator produces models with lower calibration error, for example 35% lower calibration error when we use 100 bins per class.

Using more bins allows a model to produce more fine-grained predictions, e.g. [18] use $B = 51$ bins, which improves the quality of predictions which can be measured by the mean-squared error – Figure 3b shows that our method achieves better mean-squared errors for any given calibration constraint. More concretely, the figure shows a scatter plot of the mean-squared error and L^2 calibration error for histogram binning and the scaling-binning calibrator when we vary the number of bins. For example, if we want our models to have an L^2 calibration error ≤ 0.0004 (equivalently, an ℓ_2 'average' calibration error of 2%) we get a 9% better mean-squared error. In Appendix E we show that we get *5x lower top-label calibration error on ImageNet*, and give further experiment details.

Validating theoretical bounds: We ran synthetic experiments to validate the bound in Theorem 4.1. Our theory predicts that $n \lesssim 1/\epsilon^2 + B$ for our method (if the model family is well-specified) but for histogram binning $n \lesssim B/\epsilon^2$. We set the ground truth $P(Y = 1|Z = z) = g(z)$ where g is from the Platt scaling family G. In the first experiment, we fix B and vary n, computing 90% confidence intervals from 1000 trials—we see that $1/\epsilon^2$ is approximately linear in n for both calibrators. For example, when $B = 10$ if we increase from $n = 1000$ to $n = 2000$ the L^2-CE of histogram binning decreases by 2.00 ± 0.06 times, and the L^2-CE of our method decreases by 1.98 ± 0.09 times. In the second experiment, we fix n and vary B—as predicted by the theory, for the scaling-binning calibrator $1/\epsilon^2$ is nearly constant, but for histogram binning $1/\epsilon^2$ scales close to $1/B$. When $n = 2000$ and we increase from 5 to 20 bins, our method's L^2-CE decreases by $2\% \pm 7\%$ but for histogram binning it increases by 3.71 ± 0.15 *times*. We give experimental details and plots in Appendix E.

5 Verifying calibration

Before deploying our model we would like to check that it has calibration error below some desired threshold ϵ^2. In this section we show that we can efficiently estimate the calibration error of binned models, if the binning scheme is 2-well-balanced. Recent work in machine learning typically estimates each term in the calibration error directly from samples [7, 15, 16, 17]. Older work in meteorology [18, 19] notices that this leads to a biased estimate, and proposes a 'debiased' estimator that subtracts off an approximate correction term to reduce the bias. Our contribution is to show that while the naive estimator requires samples proportional to B to estimate the calibration error, the debiased estimator requires samples proportional to $\sqrt{B}$. To our knowledge we are the first to show an *improved sample complexity*—prior work only showed that the naive estimator is biased.

Suppose we wish to measure the L^2 calibration error E^{*2} of a binned model $f : \mathcal{X} \to S$ where $S \subseteq [0, 1]$ and $|S| = B$. Suppose we get an evaluation set $T_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Past work typically estimates the calibration error by directly estimating each term from samples:

Definition 5.1 (Plugin estimator)**.** *Let L_s denote the y_j values where the model outputs s: $L_s = \{y_j \mid (x_j, y_j) \in T_n \wedge f(x_j) = s\}$. Let $\hat{p}_s$ be the estimated probability of f outputting s: $\hat{p}_s = \frac{|L_s|}{n}$. Let $\hat{y}_s$ be the empirical average of Y when the model outputs s: $\hat{y}_s = \sum_{y \in L_s} \frac{y}{|L_s|}$*

The plugin estimate for the L^2 calibration error is the weighted squared difference between $\hat{y}_s$ and s:

$$\hat{E}^2_{\text{pl}} = \sum_{s \in S} \hat{p}_s (s - \hat{y}_s)^2$$

Alternatively, [18, 19] propose to subtract an approximation of the bias from the estimate:

Definition 5.2 (Debiased estimator)**.** *The debiased estimator for the L^2 error is:*

$$\hat{E}^2 = \sum_{s \in S} \hat{p}_s \Big[(s - \hat{y}_s)^2 - \frac{\hat{y}_s(1 - \hat{y}_s)}{\hat{p}_s n - 1} \Big]$$

We are interested in analyzing the number of samples required to estimate the calibration error within a constant multiplicative factor, that is to give an estimate $\hat{E}^2$ such that $|\hat{E}^2 - E^{*2}| \leq \frac{1}{2} E^{*2}$ (where $\frac{1}{2}$ can be replaced by any constant r with $0 < r < 1$). Our main result is that the plugin estimator requires $\widetilde{O}(\frac{B}{\epsilon^2})$ samples (Theorem 5.3) while the debiased estimator requires $\widetilde{O}(\frac{\sqrt{B}}{\epsilon^2})$ samples (Theorem 5.4), where $\epsilon^2 = E^{*2}$.

Theorem 5.3 (Plugin bound)**.** *Suppose we have a binned model with L^2 calibration error $E^{*2} = \epsilon^2$, where the binning scheme is 2-well-balanced, that is for all $s \in S$, $\mathbb{P}(f(X) = s) \geq \frac{1}{2B}$.* [1] *If $n \geq c \frac{B}{\epsilon^2} \log \frac{B}{\delta}$ for some universal constant c then for the plugin estimator: $\frac{1}{2} E^{*2} \leq \hat{E}^2_{\text{pl}} \leq \frac{3}{2} E^{*2}$ with probability at least $1 - \delta$.*

Theorem 5.4 (Debiased bound)**.** *Suppose we have a binned model with L^2 calibration error $E^{*2} = \epsilon^2$ and for all $s \in S$, $\mathbb{P}(f(X) = s) \geq \frac{1}{2B}$. If $n \geq c \frac{\sqrt{B}}{\epsilon^2} \log \frac{B}{\delta}$ for some universal constant c then for the debiased estimator: $\frac{1}{2} E^{*2} \leq \hat{E}^2 \leq \frac{3}{2} E^{*2}$ with probability at least $1 - \delta$.*

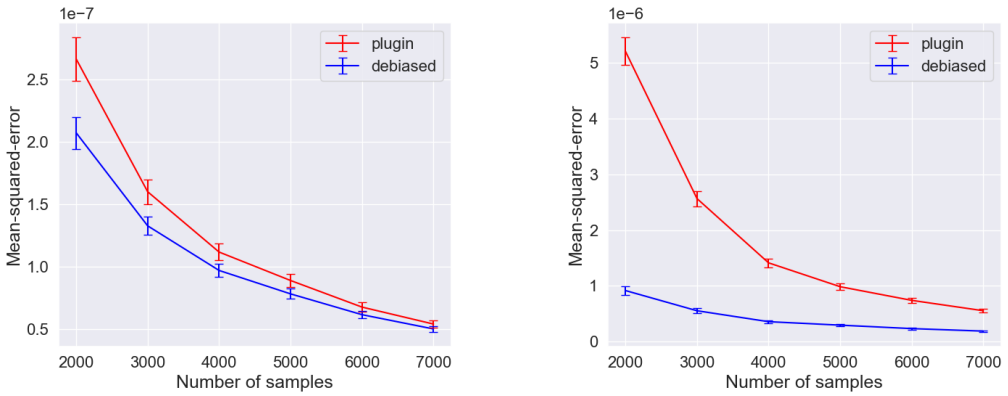
The proof of both theorems is in Appendix F. The idea is that for the plugin estimator each term in the sum has bias $1/n$. These biases accumulate, giving total bias B/n. The debiased estimator has much lower bias and the estimation variance cancels across bins—this intuition is captured in Lemma F.8 which requires careful conditioning to make the argument go through.

5.1 Experiments

We run a multiclass marginal calibration experiment on CIFAR-10 which suggests that the debiased estimator produces better estimates of the calibration error than the plugin estimator. We split the validation set of size 10,000 into two sets S_C and S_E of sizes 3,000 and 7,000 respectively. We use S_C to re-calibrate and discretize a trained VGG-16 model. We calibrated each of the $K = 10$ classes seprately as described in Section 2 and used $B = 100$ or $B = 10$ bins per class. For varying values of n, we sample n points with replacement from S_E, and estimate the calibration error using the debiased estimator and the plugin estimator. We then compute the squared deviation of these estimates from the calibration error measured on the entire set S_E. We repeat this resampling 1,000 times to get the mean squared deviation of the estimates from the ground truth and confidence intervals. Figure 4a shows that the debiased estimates are much closer to the ground truth than the plugin estimates—the difference is especially significant when the number of samples n is small or the number of bins B is large. Note that having a perfect estimate corresponds to 0 on the Y-axis.

In Appendix G, we include histograms of the absolute difference between the estimates and ground truth for the plugin and debiased estimator, over the 1,000 resamples. We also show that for any desired calibration error, the debiased estimator enables us to pick out models with a lower mean-squared error. For example, if we use the debiased estimator we would select a model with 13% lower mean-squared error if we want the ℓ_2-CE to be less than 1.5%.

[1] We do not need the upper bound of the 2-well-balanced property.



(a) $B = 10$ (b) $B = 100$

Figure 4: Mean-squared errors of debiased and plugin estimators on a recalibrated VGG16 model on CIFAR-10 with 90% confidence intervals (lower values better). The debiased estimator is closer to the ground truth, which corresponds to 0 on the y-axis, especially when B is large and/or the number of samples used by the estimator, n, is small.

6 Additional related work

Calibration, including the L^2 calibration error, has been studied in many fields such as meteorology [2, 3, 4, 5, 6], medicine [1, 27, 28], reinforcement learning [29], natural language processing [7, 8], speech recognition [30], econometrics [25], psychology [31], and machine learning [9, 10, 11, 13, 14, 15, 17]. Besides the calibration error metric, prior work also uses the Hosmer-Lemeshov test [32] and reliability diagrams [4, 33] to evaluate calibration. Besides calibration, other ways of producing and quantifying uncertainties include Bayesian methods [34] and conformal prediction [35, 36].

Recalibration is related to (conditional) density estimation [37, 38] as the goal is to estimate $\mathbb{E}[Y \mid f(X)]$. The L_2 loss function used in the density estimation literature is closely related to the L^2 calibration error. Algorithms and analysis in density estimation typically assume the true density is $L-$Lipschitz, while in calibration applications, the calibration error of the final model should be measurable from data, without making untestable assumptions on L.

Bias is a common issue with statistical estimators, for example for the sample standard deviation. It has also long been known that the mean-squared error, measured on samples, gives a biased estimate—the seminal work by Stein [39] investigates and fixes this bias. However, debiasing an estimator does not typically lead to *an improved sample complexity*, as it does in our case.

7 Conclusion and future work

In this paper we had three contributions: 1. We showed that the calibration error of continuous methods is underestimated; 2. We introduced the first method, to our knowledge, that has better sample complexity than histogram binning but has a *measurable calibration error*, giving us the best of both worlds of scaling and binning; and 3. We showed that an alternative estimator has better sample complexity than the commonly used plugin estimator. There are many exciting avenues for future work:

1. **Dataset shifts**: Can we maintain calibration under dataset shifts? When the dataset shifts (for example, train on MNIST, but evaluate on SVHN) it is difficult to get very high accuracies on the target dataset, but can we at least know when our model is accurate and when it is not, or in other words can we ensure our model is calibrated on the new dataset? When a small amount of labeled *target* data is available, we can simple re-calibrate using the labeled target data since recalibration simply involves tuning a small number of parameters and therefore requires very few samples. However, can we maintain calibration when we do not have labeled examples from the target dataset?

2. **Measuring calibration**: Can we come up with alternative metrics that still capture a notion of calibration, but are measurable for scaling methods?
3. **Multiclass calibration**: Most papers on calibration focus on top-label calibration—can we come up with better methods for marginal calibration? Can we achieve stronger notions of calibration, such as joint calibration, efficiently?
4. **Better binning**: Can we devise better ways of binning? While binning makes the calibration error measurable, it leads to an increase in the mean-squared error. In Proposition D.4 we bound the increase in MSE for the well-balanced binning scheme. However, in Section D.3 we give intuition for why other binning schemes may have a smaller increase in MSE.
5. **Finite sample guarantees**: Can we prove finite sample guarantees for Theorem 4.1 which show the dependency on the dimension and probability of failure?
6. **Estimating calibration**: The plugin estimator for the ℓ_1 calibration error (called the average calibration error in [9, 24]) is also biased—can we come up with a more sample efficient estimator for the ℓ_1 calibration error? For the L^2 calibration error, can we devise an even better estimator? If not, can we prove minimax lower bounds on this?

8 Acknowledgements

The authors would like to thank the Open Philantropy Project, Stanford Graduate Fellowship, and Toyota Research Institute for funding. Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

We are grateful to Pang Wei Koh, Chuan Guo, Anand Avati, Shengjia Zhao, Weihua Hu, Yu Bai, John Duchi, Dan Hendrycks, Jonathan Uesato, Michael Xie, Albert Gu, Aditi Raghunathan, Fereshte Khani, Stefano Ermon, Eric Nalisnick, and Pushmeet Kohli for insightful discussions. We thank the anonymous reviewers for their thorough reviews and suggestions that have improved our paper. We would also like to thank Pang Wei Koh, Yair Carmon, Albert Gu, Rachel Holladay, and Michael Xie for their inputs on our draft, and Chuan Guo for providing code snippets from their temperature scaling paper.

References

[1] X. Jiang, M. Osl, J. Kim, and L. Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19 (2):263–274, 2012.

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

[3] A. H. Murphy and R. L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26: 41–47, 1977.

[4] M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32:12–22, 1983.

[5] T. Gneiting and A. E. Raftery. Weather forecasting with ensemble methods. *Science*, 310, 2005.

[6] J. Brocker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009.

[7] K. Nguyen and B. O'Connor. Posterior calibration and exploratory analysis for natural language processing models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1587–1598, 2015.

[8] D. Card and N. A. Smith. The importance of calibration for estimating proportions from annotations. In *Association for Computational Linguistics (ACL)*, 2018.

[9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

[10] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning (ICML)*, pages 609–616, 2001.

[11] V. Kuleshov, N. Fenner, and S. Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning (ICML)*, 2018.

[12] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

[13] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 694–699, 2002.

[14] M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Binary classifier calibration: Non-parametric approach. *arXiv*, 2014.

[15] D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2019.

[16] V. Kuleshov and P. Liang. Calibrated structured prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[17] D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning (ICML)*, 2019.

[18] J. Brocker. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate Dynamics*, 39:655–667, 2012.

[19] C. A. T. Ferro and T. E. Fricker. A bias-corrected decomposition of the brier score. *Quarterly Journal of the Royal Meteorological Society*, 138(668):1954–1960, 2012.

[20] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[21] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[22] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[23] M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2015.

[24] J. V. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. Measuring calibration in deep learning. *arXiv*, 2019.

[25] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[26] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253, 2003.

[27] C. S. Crowson, E. J. Atkinson, and T. M. Therneau. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 25:1692–1706, 2017.

[28] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[29] A. Malik, V. Kuleshov, J. Song, D. Nemer, H. Seymour, and S. Ermon. Calibrated model-based deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2019.

[30] D. Yu, J. Li, and L. Deng. Calibration of confidence measures in speech recognition. *Trans. Audio, Speech and Lang. Proc.*, 19(8):2461–2473, 2011.

[31] S. Lichtenstein, B. Fischhoff, and L. D. Phillips. *Judgement under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.

[32] D. W. Hosmer and S. Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9:1043–1069, 1980.

[33] J. Bröcker and L. A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661, 2007.

[34] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC Chapman and Hall/CRC, 1995 1995.

[35] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research (JMLR)*, 9:371–421, 2008.

[36] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111, 2016.

[37] Larry Wasserman. Density estimation. `http://www.stat.cmu.edu/~larry/=sml/densityestimation.pdf`, 2019.

[38] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[39] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.

[40] François Chollet. keras. `https://github.com/fchollet/keras`, 2015.

[41] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[42] Yonatan Geifman. cifar-vgg. `https://github.com/geifmany/cifar-vgg`, 2015.

[43] A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

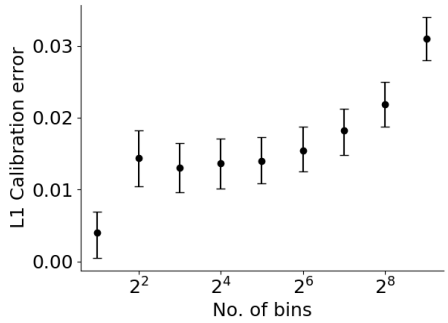
[44] J. H. Hubbard and B. B. Hubbard. *Vector Calculus, Linear Algebra, And Differential Forms*. Prentice Hall, 1998.

[45] M. Kull, T. M. S. Filho, and P. Flach. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11: 5052–5080, 2017.

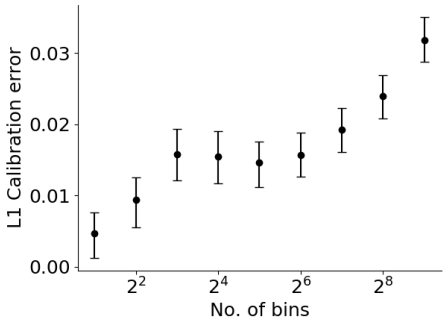
A Model and code details

The VGG16 model we used for ImageNet experiments is from the Keras [40] module in the TensorFlow [41] library and we used pre-trained weights supplied by the library. The VGG16 model for CIFAR-10 was obtained from an open-source implementation on GitHub [42], and we used the pre-trained weights there. We independently verified the accuracies of these models.

We include our code for the experiments in the supplementary material.

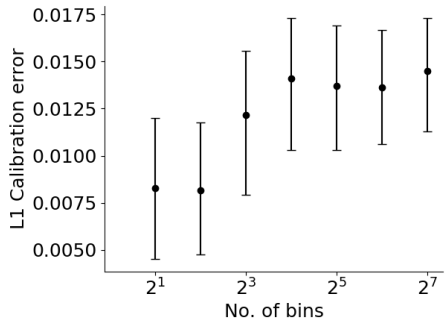


(a) ImageNet, ℓ_1-CE

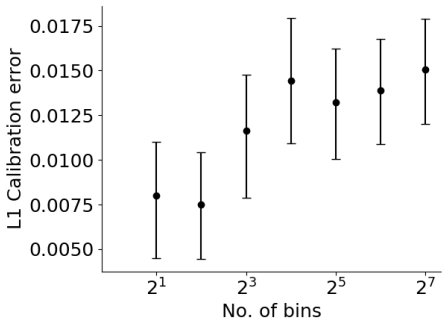


(b) ImageNet, ℓ_1-CE, equal-probability binning

Figure 5: Binned ℓ_1 calibration errors of a recalibrated VGG-net model on ImageNet with 90% confidence intervals. The binned calibration error increases as we increase the number of bins. This suggests that binning cannot be reliably used to measure the ℓ_1-CE.



(a) CIFAR-10, ℓ_1-CE



(b) CIFAR-10, ℓ_1-CE, equal-probability binning

Figure 6: Binned ℓ_1 calibration errors of a recalibrated VGG-net model on CIFAR-10 with 90% confidence intervals. The results are not as conclusive here because the error bars are large, however it seems to suggest that the binned calibration error increases as we increase the number of bins.

B Ablations for Section 3

Here we present additional experiments for Section 3. Recall that the experiments in section 3 showed that binning underestimates the calibration error of a model—we focused on the ℓ_2-CE and selected bins so that each bin has an equal number of data points. Figure 5a shows that binning is also unreliable at measuring the ℓ_1-CE on ImageNet—using more bins uncovers a higher calibration error than we would otherwise detect with fewer bins. Figure 5b shows that the same conclusion holds on ImageNet if we look at the ℓ_1-CE *and* use an alternative approach to selecting bins used in [9] that we call *equal-probability binning*. Here, the B bins are selected to be $I_1 = [0, \frac{1}{B}], I_2 = (\frac{1}{B}, \frac{2}{B}], \ldots, I_B = (\frac{B-1}{B}, 1]$. The experimental protocol is the same as in section 3.

We repeated both of these experiments on CIFAR-10 as well, and plot the results in Figure 6. Here the results are inconclusive because the error bars are large. This is because the CIFAR-10 dataset is smaller than ImageNet, and the accuracy of the CIFAR-10 model is 93.1%, so the calibration error that we are trying to measure is much smaller.

We provide details on the dataset split for CIFAR-10. For CIFAR-10, we used a VGG16 model and split the test set into 3 sets of size $(1000, 1000, 8000)$, where used the first set of data to recalibrate the model using Platt scaling, the second to select the binning scheme, and the third to measure the binned calibration error. As stated in the main body of the paper, for ImageNet we used a split of $(20000, 5000, 25000)$

C Proofs for Section 3

Restatement of Example 3.2. *For any binning scheme $\mathcal{B}$, $p \in \mathbb{Z}^+$, and continuous bijective function $f : [0,1] \to [0,1]$, there exists a distribution P over X, Y s.t. ℓ_p-CE$(f_{\mathcal{B}}) = 0$ but ℓ_p-CE$(f) \geq 0.49$. Note that for all f, $0 \leq \ell_p$-CE$(f) \leq 1$.*

Proof. As stated in the main text, the intuition is that in each interval I_j in $\mathcal{B}$, the model could underestimate the true probability $\mathbb{E}[Y \mid f(X)]$ half the time, and overestimate the probability half the time. So if we average over the entire bin the model appears to be calibrated, even though it is very uncalibrated. The proof simply formalizes this intuition.

Since f is bijective and continuous we can select data distribution P s.t. $f(X) \sim \text{Uniform}[0.5 - \epsilon, 0.5 + \epsilon]$ for any $\epsilon > 0$. To see this, first note that from real analysis since $f : [0,1] \to [0,1]$ and f is bijective and continuous, f^{-1} is also bijective and continuous. Then we can let $X \sim f^{-1}(\text{Uniform}[0.5 - \epsilon, 0.5 + \epsilon])$ which has the desired property and has a density.

Now, consider each interval I_j in binning scheme $\mathcal{B}$, and let $A_j = I_j \cap \text{Uniform}[0.5 - \epsilon, 0.5 + \epsilon]$. If $A_j = \emptyset$ then $P(f(X) \in A_j) = 0$ so we can ignore this interval (since $f(X)$ will never land in this bin). Let $p_j = \mathbb{E}[f(X) \mid f(X) \in A_j]$. Note that $\mathbb{E}[f(X) \mid f(X) \in A_j] = \mathbb{E}[f(X) \mid f(X) \in I_j]$. Since $f(X) \in [0.5 - \epsilon, 0.5 + \epsilon]$, $p_j \in [0.5 - \epsilon, 0.5 + \epsilon]$ as well. We will choose $P(Y)$ so that Y is 1 whenever $f(X)$ lands in the first p_j fraction of interval A_j, and 0 whenever $f(X)$ lands in the latter $1 - p_j$ fraction of A_j. Then $\mathbb{E}[Y \mid f(X) \in A_j] = p_j$, so the binned calibration error is 0. But notice that for all $s \in [0.5 - \epsilon, 0.5 + \epsilon]$, $\mathbb{E}[Y \mid f(X) = s]$ is either 0 or 1. So we have:

$$|\mathbb{E}[Y \mid f(X) = s] - s| \geq 0.5 - \epsilon$$

That is, at every point the model is actually very miscalibrated at each s. By taking ϵ very small, we then get that ℓ_p-CE$(p) \geq 0.5 - \epsilon'$ for any $\epsilon' > 0$, which completes the proof. □

Restatement of Proposition 3.3. *Given binning scheme $\mathcal{B}$ and model $f : \mathcal{X} \to [0,1]$, we have:*

$$\ell_p\text{-CE}(f_{\mathcal{B}}) \leq \ell_p\text{-CE}(f)$$

Proof. It suffices to prove the claim for the ℓ_p^p error:

$$(\ell_p\text{-CE}(f_{\mathcal{B}}))^p \leq (\ell_p\text{-CE}(f))^p$$

This is because if $p > 0$ then $a \leq b \Leftrightarrow a^p \leq b^p$.

For $p \geq 1$, let $l(a,b) = (|a - b|)^p$. We note that l is convex in both arguments. The proof is now a simple result of Jensen's inequality and convexity of l. Suppose that $\mathcal{B}$ is given by intervals $I_1, ..., I_B$. Let $Z = f(X)$—note that Z is a random variable.

We can write $(\ell_p\text{-CE}(f_{\mathcal{B}}))^p$ as:

$$(\ell_p\text{-CE}(f_{\mathcal{B}}))^p = \sum_{j=1}^{B} P(Z \in I_j)\; l\Big(\mathbb{E}[Z \mid Z \in I_j], \mathbb{E}[Y \mid Z \in I_j]\Big)$$

We can write $(\ell_p\text{-CE}(f))^p$ as:

$$(\ell_p\text{-CE}(f))^p = \sum_{j=1}^{B} P(Z \in I_j)\; \mathbb{E}\Big[l\big(Z, \mathbb{E}[Y \mid Z]\big) \mid Z \in I_j\Big]$$

Fix some bin $I_j \in \mathcal{B}$. By Jensen's inequality,

$$l\Big(\mathbb{E}[Z \mid Z \in I_j], \mathbb{E}[Y \mid Z \in I_j]\Big) \leq \mathbb{E}\Big[l\big(Z, \mathbb{E}[Y \mid Z]\big) \mid Z \in I_j\Big]$$

Since this inequality holds for each term in the sum, it holds for the whole sum:

$$(\ell_p\text{-CE}(f_{\mathcal{B}}))^p \leq (\ell_p\text{-CE}(f))^p$$

Note that the proof also implies that finer binning schemes give a better lower bound. That is, given $\mathcal{B}'$ suppose for all $I'_j \in \mathcal{B}'$, $I'_j \subseteq I_k$ for some $I_k \in \mathcal{B}$. Then ℓ_p-CE$(f_{\mathcal{B}}) \leq \ell_p$-CE$(f_{\mathcal{B}'}) \leq \ell_p$-CE$(f)$. This is because $f_{\mathcal{B}'}$ can be seen as a binned version of $f_{\mathcal{B}}$.

□

D Proofs for section 4

Our analysis of the sample complexity of the scaling-binning calibrator requires some assumptions on the function family $\mathcal{G}$:

1. (Finite parameters). Let $\mathcal{G} = \{g_\theta : \mathcal{Z} \to [0,1] \mid \theta \in A\}$ where $A \subseteq \mathbb{R}^d$ and A is open.
2. (Injective). For all $g_\theta \in \mathcal{G}$ we assume g_θ is injective.
3. (Consistency) Intuitively, consistency means that given infinite data, the estimated parameters should converge to the unique optimal parameters in A. More formally, suppose $\theta^* = \text{argmin}_{\theta \in A} \text{MSE}(g_\theta)$. Then the parameters $\hat{\theta}_n$ estimated by minimizing the empirical MSE with n samples in step 1 of the algorithm, converges in distribution to θ^*, that is, $\hat{\theta}_n \to_D \theta^*$ as $n \to \infty$. Note that consistency inherently assumes identifiability, that there is a unique minimizer θ^* in the open set A.
4. (Regularity). We assume that regularity conditions in Theorem 5.23 of [43] hold, which require the loss to be twice differentiable with symmetric, non-singular Hessian, and that $g_\theta(x)$ is Lipschitz in θ for all x. We will also require the second derivative to be continuous.

We assume that $\mathcal{G}$ satisfies these assumptions in the rest of this section. Note that aside from injectivity, the remaining conditions are only required for the (fairly standard) analysis of step 1 of the algorithm, which says that a parametric scaling method with a small number of parameters will quickly converge to its optimal error.

D.1 Calibration bound (Proof of Theorem 4.1)

The goal is to prove the following theorem from Section 4, which we restate:

Restatement of Theorem 4.1. *Assume regularity conditions on $\mathcal{G}$ (finite parameters, injectivity, Lipschitz-continuity, consistency, twice differentiability). Given $\delta \in (0,1)$, there is a constant c such that* for all *B, ϵ, with $n \geq c\left(B \log B + \frac{\log B}{\epsilon^2}\right)$ samples, the scaling-binning calibrator finds $\hat{g}_\mathcal{B}$ with L^2-CE$(\hat{g}_\mathcal{B}) \leq 2 \min_{g \in \mathcal{G}} L^2$-CE$(g) + \epsilon^2$, with probability $\geq 1 - \delta$.*

We will analyze each step of our algorithm and then combine the pieces to get Theorem 4.1. As we mention in the main text, step 3 is the main step, so Lemma D.2 is one of the core parts of our proof. Step 2 is where we construct a binning scheme so that each bin has an equal number of points—we show that this property holds approximately in the population (Lemma 4.3). This is important as well, particularly to ensure we can estimate the calibration error. Step 1 is basically Platt scaling, and the asymptotic analysis is fairly standard.

Step 3: Our proofs will require showing convergence in ℓ_2 and ℓ_1 norm in function space, which we define below:

Definition D.1 (Distances between functions)**.** *Given $f, g : \mathcal{Z} \to [0,1]$, for the ℓ_2 norm we define $||f-g||_2^2 = \mathbb{E}[(f(Z)-g(Z))^2]$ and $||f-g||_2 = \sqrt{||f-g||_2^2}$. For the ℓ_1 norm we define $||f-g||_1 = \mathbb{E}[|f(Z)-g(Z)|]$*

Recall that we showed that in the limit of infinite data the binned version of g, $g_\mathcal{B}$, has lower calibration error than g (Proposition 3.3). However our method uses n data points to empirically bin g, giving us $\hat{g_\mathcal{B}}$. We now show the key lemma that allows us to bound the calibration error and later the mean-squared error. That is, we show that the empirically binned function $\hat{g_\mathcal{B}}$ quickly converges to $g_\mathcal{B}$ in both ℓ_2 and ℓ_1 norms.

Lemma D.2 (Empirical binning)**.** *There exist constants c_B, c_1, c_2 such that the following is true. Given $g : \mathcal{Z} \to [0,1]$, binning set $T_3 = \{(z_i, y_i)\}_{i=1}^n$ and a 2-well-balanced binning scheme $\mathcal{B}$ of size B. Given $0 < \delta < 0.5$, suppose that $n \geq c_B B \log \frac{B}{\delta}$. Then with probability at least $1-\delta$, $||\hat{g_\mathcal{B}} - g_\mathcal{B}||_2 \leq \frac{c_2}{\sqrt{n}} \sqrt{\log \frac{B}{\delta}}$ and $||\hat{g_\mathcal{B}} - g_\mathcal{B}||_1 \leq \frac{c_1}{\sqrt{nB}} \sqrt{\log \frac{B}{\delta}}$*

Proof. Recall that the intuition is in Figure 1 of the main text—the $g(z_i)$ values in each bin (gray circles in Figure 1c) are in a narrower range than the y_i values (black crosses in Figure 1b) so when

we take the average we incur less of an estimation error. Now, there may be a small number of bins where the $g(z_i)$ values are not in a narrow range, but we will use the assumption that $\mathcal{B}$ is 2-well-balanced to show that these effects average out and the overall estimation error is small.

Define R_j to be the set of $g(z_i)$ that fall into the j-th bin, given by $R_j = \{g(z_i) \mid g(z_i) \in I_j \wedge (z_i, y_i) \in T_3\}$ (recall that T_3 is the data we use in step 3). Let p_j be the probability of landing in bin j, given by $p_j = \mathbb{P}(g(Z) \in I_j)$. Since $\mathcal{B}$ is 2-well-balanced, $p_j \geq \frac{1}{2B}$. Since $n \geq c_B B \log \frac{B}{\delta}$, by the multiplicative Chernoff bound, for some large enough c_B, with probability at least $1 - \frac{\delta}{2}$, $|R_j| \geq \frac{p_j}{2}$.

Consider each bin j. Let μ_j be the expected output of g in bin j, given by $\mu_j = \mathbb{E}[g(Z) \mid g(Z) \in I_j]$. $\mu(R_j)$, the mean of the values in R_j, is the empirical average of $|R_j|$ such values, each bounded between b_{j-1} and b_j where $I_j = [b_{j-1}, b_j]$. So $\hat{\mu}(R_j)$ is sub-Gaussian with parameter:

$$\sigma^2 = \frac{(b_j - b_{j-1})^2}{4|R_j|} \leq \frac{(b_j - b_{j-1})^2}{2p_j n}$$

Then by the sub-Gaussian tail bound, for any $1 \leq j \leq B$, with probability at least $1 - \frac{\delta}{2B}$, we have:

$$(\mu_j - \hat{\mu}(R_j))^2 \leq \frac{(b_j - b_{j-1})^2}{p_j n} \log \frac{4B}{\delta} \tag{5}$$

So by union bound with probability at least $1 - \frac{\delta}{2}$ the above holds for all $1 \leq j \leq B$ simultaneously. We then bound the ℓ_2-error.

$$\begin{aligned}
||\hat{g_{\mathcal{B}}} - g_{\mathcal{B}}||_2 &= \sqrt{\sum_{j=1}^{B} p_j (\mu_j - \hat{\mu}(R_j))^2} \\
&\leq \sqrt{\sum_{j=1}^{B} p_j \frac{(b_j - b_{j-1})^2}{p_j n} \log \frac{4B}{\delta}} && \text{(by equation (5))} \\
&= \sqrt{\frac{1}{n} \log \frac{4B}{\delta} \sum_{j=1}^{B} (b_j - b_{j-1})^2} \\
&\leq \sqrt{\frac{1}{n} \log \frac{4B}{\delta} \sum_{j=1}^{B} (b_j - b_{j-1})} && \text{(because } 0 \leq b_j - b_{j-1} \leq 1\text{)} \\
&\leq \sqrt{\frac{1}{n} \log \frac{4B}{\delta}} \\
&\leq c_2 \frac{1}{\sqrt{n}} \sqrt{\log \frac{B}{\delta}}
\end{aligned}$$

Similarly, we can also bound the ℓ_1-error. Here we also use the fact that $p_j \leq \frac{2}{B}$ since $\mathcal{B}$ is 2-well-balanced.

$$\begin{aligned}
||\hat{g_{\mathcal{B}}} - g_{\mathcal{B}}||_1 &= \sum_{j=1}^{B} p_j |\mu_j - \hat{\mu}(R_j)| \\
&\leq \sum_{j=1}^{B} p_j \sqrt{\frac{(b_j - b_{j-1})^2}{p_j n} \log \frac{4B}{\delta}} \\
&= \sum_{j=1}^{B} \sqrt{\frac{p_j (b_j - b_{j-1})^2}{n} \log \frac{4B}{\delta}} \\
&\leq \sum_{j=1}^{B} \sqrt{\frac{2(b_j - b_{j-1})^2}{Bn} \log \frac{4B}{\delta}} \\
&\leq \sqrt{\frac{2}{Bn} \log \frac{4B}{\delta}} \sum_{j=1}^{B} (b_j - b_{j-1}) \\
&\leq c_1 \frac{1}{\sqrt{Bn}} \sqrt{\log \frac{B}{\delta}}
\end{aligned}$$

By union bound, these hold with probability at least $1 - \delta$, which completes the proof. □

Step 2: Recall that we chose our bins so that each bin had an equal proportion of points in the recalibration set. In our proofs we required that this property (approximately) holds in the population as well. The following lemma shows this.

Restatement of Lemma 4.3. *For universal constant c, if $n \geq c\big(B \log \frac{B}{\delta}\big)$, with probability at least $1 - \delta$, the binning scheme $\mathcal{B}$ we chose is 2-well-balanced.*

Proof. Suppose we are given a bin construction set of size n, $T_n = \{(z_1, y_1), \ldots, (z_n, y_n)\}$. For any interval I, let $\hat{P}(I)$ be the empirical estimate of $P(I) = \mathbb{P}(g(Z) \in I)$ given by:

$$\hat{P}(I) = \frac{|\{(z_i, y_i) \in T_n \mid g(z_i) \in I\}|}{n}$$

We constructed the bins so that each interval I_j contains $\frac{n}{B}$ points, or in other words, $\hat{P}(I_j) = \frac{1}{B}$. We want to show that $\frac{1}{2B} \leq \mathbb{P}(g(Z) \in I_j) \leq \frac{2}{B}$. Since the intervals are chosen from data, we want a uniform concentration result that holds for all such intervals I_j.

We will use a discretization argument. The idea is that we will cover $[0, 1]$ with $10B$ disjoint small intervals such that for each of these intervals I'_j, $P(g(Z) \in I'_j) = \frac{1}{10B}$. We will then use Bernstein and union bound to get that with probability at least $1 - \delta$, for all I'_j, $|P(I'_j) - \hat{P}_j(I'_j)| \leq \frac{1}{100B}$. Given an arbitrary interval I, we can write it as an approximate union of these small intervals, which will allow us to concentrate $|P(I) - \hat{P}(I)|$.

Concentrating the small intervals: Fix some interval I'_j. Let $w_i = \mathbb{I}(g(z_i) \in I'_j)$ for $i = 1, \ldots, n$. Then $w_i \sim$ Bernoulli($\frac{1}{10B}$). $\hat{P}(I'_j)$ is simply the empirical average of n such values and as such with probability at least $1 - \frac{\delta}{10B}$:

$$|P(I'_j) - \hat{P}_j(I'_j)| \leq \sqrt{\frac{2}{10Bn} \log \frac{10B}{\delta}} + \frac{2}{3n} \log \frac{10B}{\delta}$$

If $n = cB \log \frac{B}{\delta}$ for a large enough constant c, we get:

$$|P(I'_j) - \hat{P}_j(I'_j)| \leq \frac{1}{100B}$$

And this was with probability at least $1 - \frac{\delta}{10B}$. So by union bound we get that with probability at least $1 - \delta$ this holds for all I'_j.

Concentrating arbitrary intervals: Now consider arbitrary $I \subseteq [0, 1]$. We can approximately write I as a union of the small I'_j intervals. More concretely, we can form a lower bound for $\hat{P}(I)$ by considering all I'_j contained in I:

$$S_L = \{I'_j \mid I'_j \subseteq I\}$$

Similarly we can form an upper bound for $\hat{P}(I)$ by considering all I'_j that have non-empty intersection with I:

$$S_U = \{I'_j \mid I'_j \cap I \neq \emptyset\}$$

We can then show:

$$\frac{9}{10}P(I) - \frac{1}{5B} \leq \hat{P}(I) \leq \frac{11}{10}P(I) + \frac{1}{5B}$$

Since in our case for all j, $\hat{P}(I_j) = \frac{1}{B}$, this gives us:

$$\frac{1}{2B} \leq P(I_j) \leq \frac{2}{B}$$

□

Step 1: Recall that step 1 essentially applies a scaling method—we fit a small number of parameters to the recalibration data. We show that if $\mathcal{G}$ contains $g^* \in \mathcal{G}$ with low calibration error, then the empirical risk minimizer $g \in \mathcal{G}$ of the mean-squared loss will also quickly converge to low calibration error. Intuitively, methods like Platt scaling fit a single parameter to the data so standard results in asymptotic statistics tell us they will converge quickly to their optimal error, at least in mean-squared error. We can combine this with a decomposition of the mean-squared error into calibration and refinement, and the injectivity of $g \in \mathcal{G}$, to show they also converge quickly in calibration error.

Lemma D.3 (Convergence of scaling)**.** *Given δ, there exists a constant c, such that for all n, L^2-CE$(g) \leq \min_{g' \in \mathcal{G}} L^2$-CE$(g') + \frac{c}{n}$, with probability at least $1 - \delta$.*

Proof. **From calibration error to mean-squared error**: We use the classic decomposition of the mean-squared error into calibration error (also known as reliability) and refinement[2]. For any $g \in \mathcal{G}$ we have:

$$\text{MSE}(g) = \underbrace{L^2\text{-CE}(g)}_{\text{calibration}} + \underbrace{\mathbb{E}[(\mathbb{E}[Y \mid g(Z)] - Y)^2]}_{\text{refinement}}$$

Note that the refinement term is constant for all injective $g \in \mathcal{G}$, since for injective g:

$$\mathbb{E}[(\mathbb{E}[Y \mid g(Z)] - Y)^2] = \mathbb{E}[(\mathbb{E}[Y \mid Z] - Y)^2]$$

This means that the difference in calibration error between any g and g' is precisely the difference in the mean-squared error. So it suffices to upper bound the generalization gap MSE(g) − MSE(g^*) for the mean-squared error. Our analysis is fairly standard: we will show asymptotic convergence in the parameter space, and then use a Taylor expansion to show convergence in the MSE loss.

Parameter convergence: Recall that $\hat{\theta}$ denotes the parameters estimated by optimizing the empirical mean-squared error objective on n samples in step 1 of our algorithm, and θ^* denotes the optimal parameters that minimize the mean-squared error objective on the population. From Theorem 5.23 of [43], on the asymptotic parameter convergence of M-estimators, we have as $n \to \infty$:

$$\sqrt{n}(\hat{\theta} - \theta^*) \to_D N(0, \Sigma)$$

Then for each $1 \leq i \leq d$, we have:

$$\sqrt{n}(\hat{\theta}_i - \theta^*_i) \to_D N(0, \sigma_i^2)$$

We will show that there exists c_i such that for each i and for all n, with probability at least $1 - \frac{\delta}{d}$:

$$|\hat{\theta}_i - \theta^*_i| \leq \frac{c_i}{n}$$

[2]Note that the refinement term can be further decomposed into resolution (also known as sharpness) and irreducible uncertainty.

To see this, we begin with the definition of convergence in distribution, which says that the CDFs converge pointwise at every point where the CDF is continuous, which for a Gaussian is every point. That is, letting z_i be a sample from $N(0, \sigma_i^2)$, we have for all c:

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}(\hat{\theta}_i - \theta_i^*) \geq c) = \mathbb{P}(z_i \geq c)$$

By considering the CDF at each point and its negative, we can show the same result for the absolute value:

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}|\hat{\theta}_i - \theta_i^*| \geq c) = \mathbb{P}(|z_i| \geq c)$$

The tails of a normal distribution are bounded, so we can choose c_i' such that:

$$\mathbb{P}(|z_i| \geq c_i') \leq \frac{\delta}{2d}$$

By definition of limit, this means that we can choose N_i such that for all $n \geq N_i$, we have:

$$\mathbb{P}(\sqrt{n}|\hat{\theta}_i - \theta_i^*| \geq c_i') \leq \frac{\delta}{d}$$

In other words, for all $n \geq N_i$, with probability at least $1 - \frac{\delta}{d}$:

$$|\hat{\theta}_i - \theta_i^*| \leq \frac{c_i'}{\sqrt{n}}$$

Since this only does not hold for finitely many values $1, \cdots, N_i - 1$, we can 'absorb' these cases into the constant. That is, for each $n \in \{1, \cdots, N_i - 1\}$, there exists r_n such that if we use n samples, then except with probability $\frac{\delta}{d}$, $|\hat{\theta}_i - \theta_i^*| \leq r_n$. So then we can choose c_i such that for all n:

$$|\hat{\theta}_i - \theta_i^*| \leq \frac{c_i' + \max_{1\leq m<N_i} r_m\sqrt{m}}{\sqrt{n}} \leq \frac{c_i}{\sqrt{n}}$$

We apply union bound over the indices i, and can then bound the ℓ_2-norm of the difference between the estimated and optimal parameters, so that we can choose k such that for all n, with probability at least $1 - \delta$:

$$||\hat{\theta} - \theta^*||_2^2 \leq \frac{k}{n}$$

Loss convergence: We denote the loss by L, defined as:

$$L(\theta) = \text{MSE}(g_\theta) = \mathbb{E}[(Y - g_\theta(X))^2]$$

We approximate the loss L by the first few terms of its Taylor expansion, which we denote by $\widetilde{L}$:

$$\widetilde{L}(\theta) = L(\theta^*) + \nabla L(\theta^*)^T(\hat{\theta} - \theta^*) + (\hat{\theta} - \theta^*)^T\nabla^2 L(\theta^*)(\hat{\theta} - \theta^*)$$

We assumed that L was twice differentiable with continuous second derivative, and that θ^* minimized the loss in an open set, so $\nabla L(\theta^*) = 0$, and we also have (see e.g. Theorem 3.3.18 in [44]):

$$\lim_{||\hat{\theta}-\theta^*||_2\to 0} \frac{L(\hat{\theta}) - \widetilde{L}(\hat{\theta})}{||\hat{\theta} - \theta^*||_2^2} = 0$$

By the definition of a limit if we fix $\epsilon > 0$, there exists $R > 0$ such that if $||\hat{\theta} - \theta^*||_2 \leq R$ then $L(\hat{\theta}) - \widetilde{L}(\hat{\theta}) \leq \epsilon||\hat{\theta} - \theta^*||_2^2$. For some large enough N_0, if $n \geq N_0$, then with probability at least $1 - \delta$, $||\hat{\theta} - \theta^*||_2 \leq R$. As before, since this only does not hold for finitely many N, we can fold these cases into the constant so that there exists ϵ' such that for all n, $L(\hat{\theta}) - \widetilde{L}(\hat{\theta}) \leq \epsilon'||\hat{\theta} - \theta^*||_2^2$ with probability at least $1 - \delta$. Plugging in $\widetilde{L}(\hat{\theta})$, we have:

$$L(\hat{\theta}) - L(\theta^*) \leq (\hat{\theta} - \theta^*)^T\nabla^2 L(\theta^*)(\hat{\theta} - \theta^*) + \epsilon'||\hat{\theta} - \theta^*||_2^2$$

We can bound this by the operator norm of the Hessian, and then use the parameter convergence result:

$$L(\hat{\theta}) - L(\theta^*) \leq (||\nabla^2 L(\theta^*)||_{op} + \epsilon')||\hat{\theta} - \theta^*||_2^2 \leq \frac{c}{n}$$

which holds with probability at least $1 - \delta$, as desired.

□

Finally, we have the tools to prove the main theorem:

Proof of Theorem 4.1. The proof pieces together Lemmas D.3, D.2, 4.3 and Proposition 3.3.

For any fixed $c_1 > 0$, there exists c_1' such that if $n \geq c_1'\big(\frac{1}{\epsilon^2}\big)$, from Lemma D.3, step 1 of our algorithm gives us g with L^2-CE$(g) \leq \min_{g' \in \mathcal{G}} L^2$-CE$(g') + c_1\epsilon^2$, with probability at least $1 - \frac{\delta}{3}$.

Next, for universal constant c_2, if $n \geq c_2(B \log \frac{B}{\delta})$, from Lemma 4.3, step 2 chooses a 2-well-balanced binning scheme $\mathcal{B}$ with probability at least $1 - \frac{\delta}{3}$.

From Proposition 3.3, L^2-CE$(g_{\mathcal{B}}) \leq L^2$-CE$(g) \leq \min_{g' \in \mathcal{G}} L^2$-CE$(g') + c_1\epsilon^2$. Then from Lemma D.2, for any $c_3 > 0$, there exists c_3' such that if $n \geq c_3'\big(\frac{1}{\epsilon^2} \log \frac{B}{\delta}\big)$, step 3 gives us $\hat{g_{\mathcal{B}}}$ with $||\hat{g_{\mathcal{B}}} - g_{\mathcal{B}}||_2 \leq c_3\epsilon$ with probability at least $1 - \frac{\delta}{3}$. We want to say that since $\hat{g_{\mathcal{B}}}$ is close to $g_{\mathcal{B}}$ and $g_{\mathcal{B}}$ has low calibration error, this must mean that $\hat{g_{\mathcal{B}}}$ has low calibration error.

To do this we represent the ℓ_2 calibration error of any g as the distance between g and a perfectly recalibrated version of g. That is, we define the perfectly recalibrated version of g as:

$$\omega(g)(z) = \mathbb{E}[Y \mid g(Z) = z]$$

Then for any g, we can write ℓ_2-CE$(g) = ||g - \omega(g)||_2$. By triangle inequality on the ℓ_2 norm on functions, we have:

$$||\hat{g_{\mathcal{B}}} - \omega(g_{\mathcal{B}})||_2 \leq ||\hat{g_{\mathcal{B}}} - g_{\mathcal{B}}||_2 + ||g_{\mathcal{B}} - \omega(g_{\mathcal{B}})||_2 \leq c_3\epsilon + \sqrt{\min_{g' \in \mathcal{G}} L^2\text{-CE}(g') + c_1\epsilon^2}$$

Now the LHS is not quite the ℓ_2 calibration error of $\hat{g_{\mathcal{B}}}$, which is $||\hat{g_{\mathcal{B}}} - \omega(\hat{g_{\mathcal{B}}})||_2$ [3]. However, since g is injective, $g_{\mathcal{B}}$ takes on a different value for each interval $I_j \in \mathcal{B}$. If $\hat{g_{\mathcal{B}}}$ also takes on a different value for each interval $I_j \in \mathcal{B}$, then we can see that $\omega(g_{\mathcal{B}}) = \omega(\hat{g_{\mathcal{B}}})$. If not, $\omega(\hat{g_{\mathcal{B}}})$ can only merge some of the intervals of $\omega(g_{\mathcal{B}})$, and by Jensen's we can show:

$$||\hat{g_{\mathcal{B}}} - \omega(\hat{g_{\mathcal{B}}})||_2 \leq ||\hat{g_{\mathcal{B}}} - \omega(g_{\mathcal{B}})||_2 \leq c_3\epsilon + \sqrt{\min_{g' \in \mathcal{G}} L^2\text{-CE}(g') + c_1\epsilon^2}$$

An alternative way to see this is to add infinitesimal noise to $\hat{g_{\mathcal{B}}}$ for each interval I_j, in which case we get $\omega(g_{\mathcal{B}}) = \omega(\hat{g_{\mathcal{B}}})$. Finally we convert back from the ℓ_2-CE to L^2-CE:

$$L^2\text{-CE}(\hat{g_{\mathcal{B}}}) = ||\hat{g_{\mathcal{B}}} - \omega(\hat{g_{\mathcal{B}}})||_2^2 \leq \min_{g' \in \mathcal{G}} L^2\text{-CE}(g') + (c_3^2 + c_1)\epsilon^2 + 2\sqrt{(c_3^2\epsilon^2)\big(\min_{g' \in \mathcal{G}} L^2\text{-CE}(g') + c_1\epsilon^2\big)}$$

By the AM-GM inequality, we have:

$$2\sqrt{(c_3^2\epsilon^2)\big(\min_{g' \in \mathcal{G}} L^2\text{-CE}(g') + c_1\epsilon^2\big)} \leq (c_3^2 + c_1)\epsilon^2 + \min_{g' \in \mathcal{G}} L^2\text{-CE}(g')$$

Combining these, we get:

$$L^2\text{-CE}(\hat{g_{\mathcal{B}}}) \leq 2 \min_{g' \in \mathcal{G}} L^2\text{-CE}(g') + 2(c_3^2 + c_1)\epsilon^2$$

By e.g. choosing $c_1 = 0.1$ and $c_3 = 0.1$, we have $2(c_3^2 + c_1) \leq 1$, which gives us the desired result. By union bound over each step, we have this with probability at least $1 - \delta$.

□

D.2 Bounding the mean-squared error

We also show that if we use lots of bins, discretization has little impact on model quality as measured by the mean-squared error. Note that recalibration itself typically *reduces/improves* the mean-squared error. However, in our method after fitting a recalibration function like Platt scaling does, we discretize the function outputs. This reduces the calibration error and allows us to measure the calibration error, but it does increase the mean-squared error by a small amount. Here we upper bound the increase in mean-squared error. In other words, our method allows for the calibration error of the final model to be measured, and has little impact on the mean-squared error.

[3] This is a very technical point, so at a first pass the reader may skip the following discussion.

Proposition D.4 (MSE Bound)**.** *If $\mathcal{B}$ is a 2-well-balanced binning scheme of size B and $B = \widetilde{\Omega}(n)$, where $\widetilde{\Omega}$ hides* $\log$ *factors, then* $\text{MSE}(\hat{g}_{\mathcal{B}}) \leq \text{MSE}(g) + O(\frac{1}{B})$.

To show this we begin with a lemma showing that if f and g are close in ℓ_1 norm, then their mean-squared errors are close:

Lemma D.5. *For $f, g : \mathcal{Z} \to [0, 1]$,* $\text{MSE}(f) \leq \text{MSE}(g) + 2||f - g||_1$.

Proof.

$$\begin{aligned}\mathbb{E}[(f(Z) - Y)^2 - (g(Z) - Y)^2] &= \mathbb{E}[(f(Z) - g(Z))(f(Z) + g(Z) - 2Y)] \\ &\leq \mathbb{E}[|f(Z) - g(Z)||f(Z) + g(Z) - 2Y|] \\ &\leq \mathbb{E}[2|f(Z) - g(Z)|] \\ &= 2||f - g||_1\end{aligned}$$

Where the third line followed because $-2 \leq f(Z) + g(Z) - 2Y \leq 2$. □

Next, we show that in the limit of infinite data, if we bin with a well-balanced binning scheme then the MSE cannot increase by much.

Lemma D.6. *Let $\mathcal{B}$ be an α-well-balanced binning scheme of size B. Then* $\text{MSE}(g_{\mathcal{B}}) \leq \text{MSE}(g) + \frac{2\alpha}{B}$.

Proof. We bound $||g_{\mathcal{B}} - g||_1$ and then use Lemma D.5. We use the law of total expectation, conditioning on $\beta(g(Z))$, the bin that $g(Z)$ falls into.

$$\begin{aligned}||g_{\mathcal{B}} - g||_1 &= \mathbb{E}[|g_{\mathcal{B}}(Z) - g(Z)|] \\ &\leq \mathop{\mathbb{E}}_{\beta(g(Z))}\Big[\mathop{\mathbb{E}}_{Z|\beta(g(Z))}[|g_{\mathcal{B}}(Z) - g(Z)|]\Big] \\ &\leq \mathop{\mathbb{E}}_{\beta(g(Z))}\Big[b_{\beta(g(Z))} - b_{\beta(g(Z))-1}\Big]\end{aligned}$$

We now use the fact that $\mathcal{B}$ is α-well-balanced.

$$\begin{aligned}\mathop{\mathbb{E}}_{\beta(g(Z))}\Big[(b_{\beta(g(Z))} - b_{\beta(g(Z))-1})\Big] &= \sum_{i=1}^{B} \mathbb{P}\big(g(Z) \in [b_{\beta(g(Z))-1}, b_{\beta(g(Z))}]\big)(b_{\beta(g(Z))} - b_{\beta(g(Z))-1}) \\ &\leq \sum_{i=1}^{B} \frac{\alpha}{B}(b_{\beta(g(Z))} - b_{\beta(g(Z))-1}) \\ &\leq \frac{\alpha}{B}\end{aligned}$$

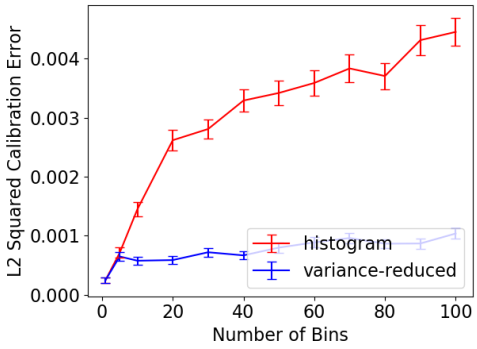
Finally, from Lemma D.5, we get that $\text{MSE}(g_{\mathcal{B}}) \leq \text{MSE}(g) + \frac{2\alpha}{B}$. □

The above lemma bounds the increase in MSE due to binning in the infinite sample case – next we deal with the finite sample case and prove proposition D.4:

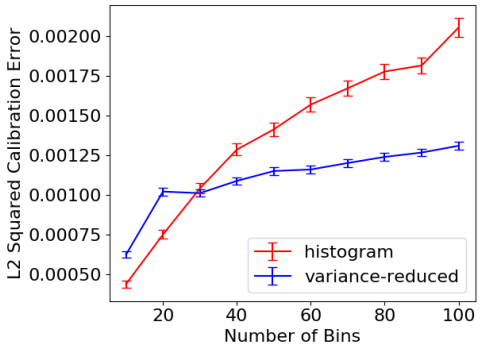
Proof of Proposition D.4: Ignoring all $\log$ factors, from Theorem D.2 if $n = \widetilde{\Omega}(B)$, we have $||\hat{g}_{\mathcal{B}} - g_{\mathcal{B}}||_1 = O(\frac{1}{\sqrt{nB}})$. Then from Lemma D.5, $\text{MSE}(\hat{g}_{\mathcal{B}}) \leq \text{MSE}(g_{\mathcal{B}}) + O(\frac{1}{\sqrt{Bn}}) \leq \text{MSE}(g_{\mathcal{B}}) + O(\frac{1}{B})$. From Theorem D.6, since $\mathcal{B}$ is 2-well-balanced, we have $\text{MSE}(g_{\mathcal{B}}) \leq \text{MSE}(g) + O(\frac{1}{B})$. This gives us $\text{MSE}(\hat{g}_{\mathcal{B}}) \leq \text{MSE}(g) + O(\frac{1}{B})$. □

D.3 Alternative binning schemes

We note that there are alternative binning schemes in the literature. For example, the B bins can be chosen as $I_1 = [0, \frac{1}{B}], I_2 = (\frac{1}{B}, \frac{2}{B}], \ldots, I_B = (\frac{B-1}{B}, 1]$. The main problem with this binning scheme is that we may not be able to measure the calibration error efficiently, which is critical. However, if we choose the bins like this, and are lucky that the binning scheme happens to be 2-well-balanced, we can improve the bounds on the MSE that we proved above. This motivates



(a) Effect of number of bins B on top calibration error L^2-TCE on ImageNet.



(b) Effect of number of bins B on top calibration error L^2-TCE on CIFAR-10.

Figure 7: Recalibrating using 1,000 data points on ImageNet and CIFAR-10, the scaling-binning calibrator typically achieves lower L^2 calibration error than histogram binning, especially when the number of bins B is large. The difference is very significant on ImageNet, where our method does better when $B \geq 10$, and gets a nearly 5 times lower calibration error when $B = 100$. For CIFAR-10 our method does better when $B > 30$, which supports the theory, which predicts that our method does better when B is large. However, when B is small, practitioners should try both histogram binning and the scaling-binning calibrator.

alternative hybrid binning schemes, where we try to keep the width of the bins as close to $1/B$ as possible, while ensuring that each bin contains lots of points as well. We think analyzing what binning schemes lead to the best bounds, and seeing if this can improve the calibration method, is a good direction for future research.

E Experimental details and ablations for section 4

We give more experimental details for our CIFAR-10 experiment, show experimental results for top-label calibration in ImageNet and CIFAR-10, and give details and results for our synthetic experiments. Note that the code is available in the supplementary folder for completeness.

Experimental details: We detail our experimental protocol for CIFAR-10 first. The CIFAR-10 validation set has 10,000 data points. We sampled, with replacement, a recalibration set of 1,000 points. In our theoretical approach and analysis, we split up these sets into multiple parts. For example, we used the first part for training a function, second part for bin construction, third part for binning. In practice, using the same set for all three steps worked out better, for both histogram binning and the scaling-binning calibrator. We believe that there may be theoretical justification for merging these sets, although we leave that for future work. For the marginal calibration experiment we ran either the scaling-binning calibrator (we fit a sigmoid in the function fitting step) or histogram binning. We calibrated each of the K classes seprately as described in Section 2, and measured the marginal calibration error on the entire set of 10K points. We repeated this entire procedure 100 times, and computed mean and 90% confidence intervals.

In this experiment, we are checking a very precise hypothesis—assuming that the empirical distribution on the 10,000 validation points is the true data distribution, how do these methods perform? This is similar to the experimental protocol used in e.g. [18]. An alternative experimental protocol would have been to first split the CIFAR-10 data into two sets of size $(1000, 9000)$. We could have then used the first set to recalibrate the model using either the scaling-binning calibrator or histogram binning, and then used the remaining 9,000 examples to estimate the calibration error on the ground truth distribution, using Bootstrap to compute confidence intervals. However, when we ran this experiment, we noticed that the results were very sensitive to which set of 1,000 points we used to recalibrate. Multiple runs of this experiment led to very different results. The point is that there are two sources of randomness—the randomness in the data the recalibration method operates on, and the randomness in the data used to evaluate and compare the recalibrators. In our protocol we account for both of these sources of randomness.

Top-label calibration experiments: We also ran experiments on top-label calibration, for both ImageNet and CIFAR-10. The protocol is exactly as described above, except instead of calibrating each of the K classes, we calibrated the top probability prediction of the model. More concretely, for each input x_i, the uncalibrated model outputs a probability p_i corresponding to its top prediction k_i, where the true label is y_i. We create a new dataset $\{(p_1, \mathbb{I}(k_1 = y_1)), \ldots, (p_n, \mathbb{I}(k_n = y_n))\}$ and run the scaling-binning calibrator (fitting a sigmoid in the function fitting step, as in Platt scaling) or histogram binning on this dataset, using B bins. This calibrates the probability corresponding to the top prediction of the model. We evaluate the recalibrated models on the top-label calibration error metric (L^2-TCE) described in Section 2. For both CIFAR-10 and ImageNet we sampled, with replacement, a recalibration set of 1,000 points for the recalibration data, and we measured the calibration error on the entire set (10,000 points for CIFAR-10, and 50,000 points for ImageNet) as above. We show 90% confidence intervals for all plots.

Figure 7a shows that on ImageNet the scaling-binning calibrator gets significantly lower calibration errors than histogram binning when $B \geq 10$, and nearly a 5 times lower calibration error when $B = 100$. Both methods get similar calibration errors when $B = 1$ or $B = 5$. Figure 7b shows that on CIFAR-10 when B is high, the scaling-binning calibrator gets lower calibration errors than histogram binning, but when B is low histogram binning gets lower calibration errors. We believe that the difference might be because the CIFAR-10 model is highly accurate at top-label prediction to begin with, getting an accuracy of over 93%, so there is not much scope for re-calibration. In any case, this ablation tells us that practitioners should try multiple methods when recalibrating their models and evaluate their calibration error.

(A) Synthetic experiments to validate bounds: Suppose the model, before recalibration, outputs values in $[0, 1]$, that is, $\mathcal{Z} = [0, 1]$. We first describe the scaling family we use, which is Platt scaling after applying a log-transform [12], otherwise known as beta calibration [45]. Let σ be the standard sigmoid function given by:

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

Then, our recalibration family $\mathcal{G}$ consists of g parameterized by a, c, given by:

$$g(z; a, c) = \sigma\Big(a \log \frac{z}{1 - z} + c\Big)$$

In this set of synthetic experiments, we assume well-specification, that is $P(Y = 1 \mid Z = z) = g(z; a, c)$ for some a, c. We set $P(Z) = \text{Uniform}[0, 1]$. Since we know $P(Y = 1 \mid Z)$, we can approximate the true L^2-CE in this case, even for scaling methods. To do this, we sample $m = 10000$ points $z_1, \ldots, z_m$ independently from $P(Z)$. An *unbiased* estimate of the L^2-CE then is:

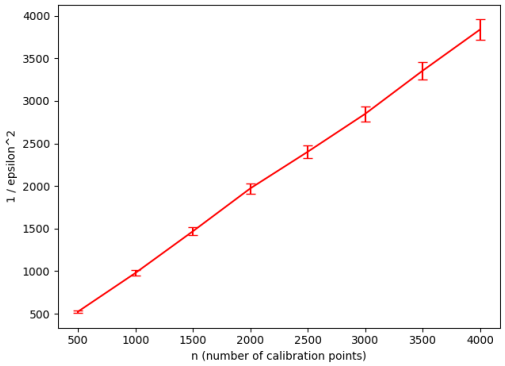
$$L^2\text{-CE}(g) \approx \frac{1}{m} \sum_{i=1}^{m} \big[P(Y \mid Z = z_i) - g(z_i)\big]^2$$

For each n (number of recalibration samples) and B (number of bins), we run either histogram binning or the scaling-binning calibrator with scaling family $\mathcal{G}$ and evaluate its calibration error as described above. We repeat this 1000 times, and compute 90% confidence intervals. We fix $a = 2$ and $c = 1$.

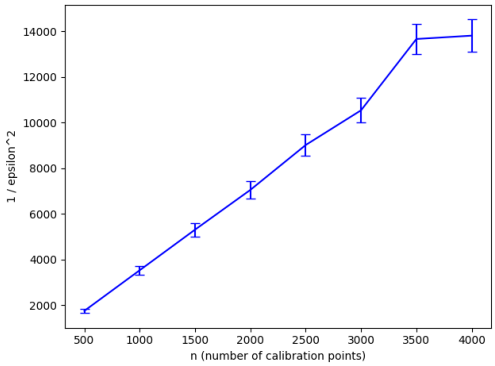
In the first sub-experiment we fix $B = 10$ and vary n, plotting $1/\epsilon^2$ in Figure 8 (recall that ϵ^2 is the L^2 calibration error). We plot the calibration errors for each method in a different plot because of the difference in scales, the scaling-binning calibrator achieves a much lower calibration error than histogram binning. As the theory predicts, we see that $1/\epsilon^2$ is approximately linear in n for both calibrators. For example, when $B = 10$ if we increase from $n = 1000$ to $n = 2000$ the L^2 calibration error of histogram binning decreases by 2.00 ± 0.06 times, and the L^2 calibration error of our method decreases by 1.98 ± 0.09 times.

In the second sub-experiment we fix $n = 2000$ and vary B, plotting $1/\epsilon^2$ in Figure 11. For the scaling-binning calibrator $1/\epsilon^2$ is nearly constant (within the margin of error), but for histogram binning $1/\epsilon^2$ scales close to $1/B$. When $n = 2000$ and we increase from 5 to 20 bins, our method's L^2-CE decreases by $2\% \pm 7\%$ but for histogram binning it increases by 3.71 ± 0.15 *times*. For reference, we plot $P(Y \mid Z = z)$ in Figure 10a.

(B) Synthetic experiments to compare the scaling-binning calibrator and the scaling method: We run an illustrative toy experiment to show that there are some cases where the scaling-binning

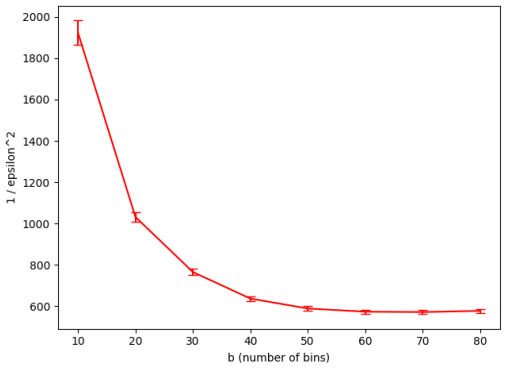


(a) Histogram binning.

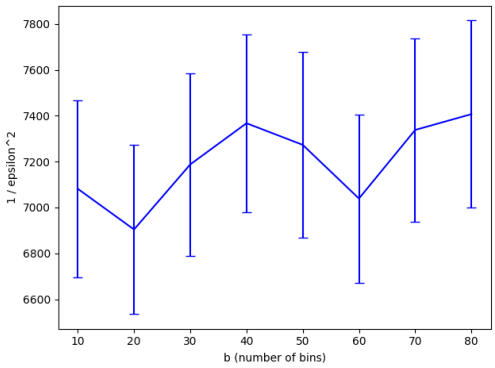


(b) The scaling-binning calibrator.

Figure 8: Plots of $1/\epsilon^2$ against n (recall that ϵ^2 is the L^2 calibration error). We see that for both methods $1/\epsilon^2$ increases approximately linearly with n, which match the theoretical bounds.

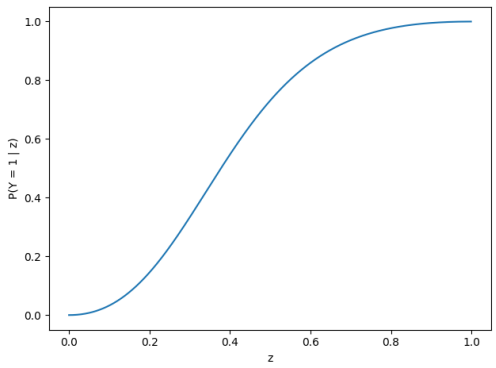


(a) Histogram binning.

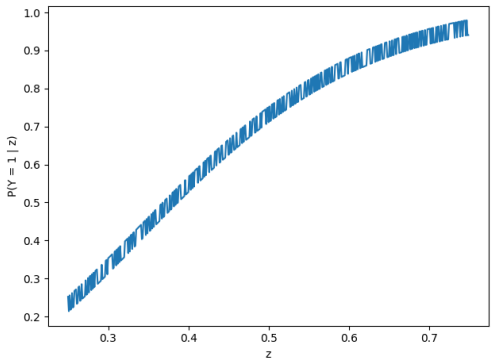


(b) The scaling-binning calibrator.

Figure 9: Plots of $1/\epsilon^2$ against b (recall that ϵ^2 is the L^2 calibration error). Note that the Y axis for the scaling-binning calibrator is clipped to 6600 and 7800 to show the relevant region. We see that for histogram binning $1/\epsilon^2$ scales close to $1/B$, in other words the calibration error increases with the number of bins (important note: the plot decreases because we plot the inverse $1/\epsilon^2$). For the scaling-binning calibrator $1/\epsilon^2$ is relatively constant, within the margin of estimation error, as predicted by the theory.



(a) $P(Y \mid Z = z)$ for Experiment (A)



(b) $P(Y \mid Z = z)$ for Experiment (B)

Figure 10: Plots of $P(Y \mid Z = z)$ against z for both synthetic experiments.

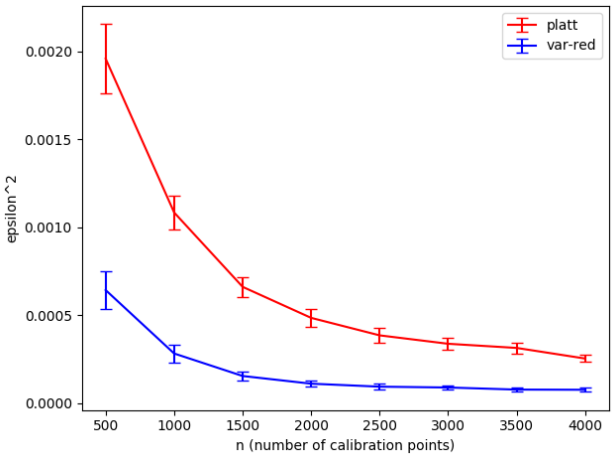


Figure 11: Plot of ϵ^2 (L^2 calibration error) against number of samples n used to recalibrate. We can see in this case the scaling-binning calibrator consistently gets lower calibration error.

calibrator does better than the underlying scaling method—there are other cases where the underlying scaling method does better. the scaling-binning calibrator can do better because if we have infinite data, Proposition 3.3 showed that the binned version $g_{\mathcal{B}}$ has lower calibration error than g. On the other hand, in step 3 of the scaling-binning calibrator algorithm we empirically bin the outputs of the scaling method which incurs an estimation error, and could mean the scaling-binning calibrator has higher calibration error than the underlying scaling method. Our key advantage is that unlike scaling methods our method has measurable calibration error so if we are not calibrated we can get more data or use a different scaling family.

Building on the previous synthetic experiments, in this experiment, we set the ground truth $P(Y = 1 \mid Z = z) = g(z; a, c) + h(z)$ where for each z, $h(z) \in \{-0.02, 0.02\}$ with equal probability. In this case we set $P(Z) = \text{Uniform}[0.25, 0.75]$ so that $P(Y = 1 \mid Z = z) \in [0, 1]$. We fix $B = 10$ and vary n, plotting the L^2 calibration error ϵ^2 in Figure 11. With $B = 10$ bins, $n = 3000$ the L^2 calibration error is 5.2 ± 1.1 times lower for the scaling-binning calibrator than the underlying scaling method using a sigmoid recalibrator. For reference, we plot $P(Y \mid Z = z)$ in Figure 10b.

F Proofs for section 5

In this section we prove the finite sample bounds for the plugin and debiased estimators. We follow a very similar structure for both the plugin estimator and the debiased estimators.

We first give a proof for the plugin estimator. At a high level, we decompose the plugin estimator into three terms (Lemma F.2), and then bound each of these terms. Most of these terms simply involve algebraic manipulation and standard concentration results, except Lemma F.4 which requires some tricky conditioning.

The debiased estimator decomposes into three terms as well, two of these terms are the same as those in the plugin estimator. Bounding the third term (Lemma F.8) is the key to the improved sample complexity of the plugin estimator. The debiased estimator is not completely unbiased. However, with high probability if we condition on the x_is in the evaluation set, each of these error terms is unbiased. We can then use Hoeffding's to concentrate each term near 0. The errors in each bin are then independent which leads to some cancelations of the error terms when we sum them up.

We use the following notation simplification to simplify the theorem statements and proofs:

$$p_i = P(f(X) = s_i)$$
$$y_i^* = \mathbb{E}[Y \mid f(X) = s_i]$$
$$e_i = (s_i - y_i^*)$$

Then, if we let E^{*2} denote the actual L^2 calibration error, we have:

$$E^{*2} = \sum_{i=1}^{b} p_i e_i^2$$

We begin by noting that $\hat{p_i}$ is close to p_i for all i. This is a standard application of either Bernstein's inequality or the multiplicative Chernoff bound.

Lemma F.1. *Suppose $p_i > \frac{12}{n} \log \frac{2B}{\delta}$ for all i. Then we can define $c(n) < 0.5$ such that except with probability δ for all i we have:*

$$|\hat{p_i} - p_i| < c(n) p_i := \sqrt{\frac{3}{n \min p_i} \log \frac{2B}{\delta}} p_i$$

F.1 Analysis of plugin estimator (proof of Theorem 5.3)

The following lemma is crucial – we decompose the plugin estimator into three terms that we can bound separately.

Lemma F.2 (Plugin decomposition)**.** *The plugin estimator satisfies the following decomposition:*

$$\hat{E}_{\text{pl}}^2 = \underbrace{\sum_{i=1}^{b} \hat{p_i} e_i^2}_{(P1)} \underbrace{- 2 \sum_{i=1}^{b} \hat{p_i} e_i (\hat{y_i} - y_i^*)}_{(P2)} + \underbrace{\sum_{i=1}^{b} \hat{p_i} (\hat{y_i} - y_i^*)^2}_{(P3)}$$

Proof. The proof is by algebra:

$$\begin{aligned}
\hat{E}_{\text{pl}}^2 &= \sum_{i=1}^{b} \hat{p_i} (s_i - \hat{y_i})^2 \\
&= \sum_{i=1}^{b} \hat{p_i} [e_i - (\hat{y_i} - y_i^*)]^2 \\
&= \underbrace{\sum_{i=1}^{b} \hat{p_i} e_i^2}_{(P1)} \underbrace{- 2 \sum_{i=1}^{b} \hat{p_i} e_i (\hat{y_i} - y_i^*)}_{(P2)} + \underbrace{\sum_{i=1}^{b} \hat{p_i} (\hat{y_i} - y_i^*)^2}_{(P3)}
\end{aligned}$$

□

We now bound each of these three terms with the following three lemmas. We condition on $|\hat{p_i} - p_i| < c(n)p_i < 0.5p_i$ for all i, which holds with high probability from Lemma F.1.

Lemma F.3. *Let $(P1)$ be as defined in Lemma F.2. Suppose $|\hat{p_i} - p_i| < c(n)p_i$ for all i. Then*

$$|(P1) - {E^*}^2| \leq c(n){E^*}^2$$

Proof. The proof is by algebra.

$$\begin{aligned}
|(P1) - {E^*}^2| &= |\sum_{i=1}^{b} \hat{p_i} e_i^2 - \sum_{i=1}^{b} p_i e_i^2| \\
&= |\sum_{i=1}^{b} (\hat{p_i} - p_i) e_i^2| \\
&\leq \sum_{i=1}^{b} |(\hat{p_i} - p_i)| e_i^2 \\
&\leq \sum_{i=1}^{b} c(n) p_i e_i^2 \\
&\leq c(n) \sum_{i=1}^{b} p_i e_i^2 \\
&\leq c(n) {E^*}^2
\end{aligned}$$

□

Lemma F.4. *Let $(P2)$ be as defined in Lemma F.2. Suppose $|\hat{p_i} - p_i| < c(n)p_i < 0.5p_i$ for all i. Then with probability $\geq 1 - \delta$:*

$$|(P2)| \leq \sqrt{\frac{2(1 + c(n)){E^*}^2}{n} \log \frac{2}{\delta}}$$

Proof. Recall that we evaluated our estimators on an independent and identically distributed evaluation set $T_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Also, note that since $|\hat{p_i} - p_i| < p_i$, $\hat{p_i} > 0$. Let $Z = (f(x_1), \cdots, f(x_n))$ be a random variable.

$\hat{y_i}$ simply takes the empirical average of the label values, and is therefore an unbiased estimator of y_i^* even if we condition on Z:

$$\mathbb{E}[\hat{y_i} - y_i^* \mid Z] = 0$$

Next we look at the distribution of $\hat{y_i} - y_i^* \mid Z$. For all $(x_j, y_j) \in T_n$, $y_j \in \{0, 1\}$. Additionally, $\{y_j \mid (x_j, y_j) \in T_n\} \mid Z$ is also independently (but not identically) distributed. So by Hoeffding's lemma, $\hat{y_i} - y_i^* \mid Z$ is sub-Gaussian with parameter $\frac{1}{4\hat{p_i}n}$.

Here, we note that $\hat{p_i}$ is a constant given Z. Then, we get that $\hat{p_i} e_i (\hat{y_i} - y_i^*) \mid Z$ has expected value 0 and is sub-Gaussian with parameter:

$$\sigma_i^2 = {\hat{p_i}}^2 e_i^2 \frac{1}{4\hat{p_i}n} = \frac{\hat{p_i} e_i^2}{4n}$$

This means that the sum, $(P2)$ has expected value 0 and is sub-Gaussian with parameter:

$$\sigma^2 = 2^2 \sum_{i=1}^{B} \sigma_i^2 = 4 \sum_{i=1}^{B} \frac{\hat{p_i} e_i^2}{4n} \leq \frac{(1 + c(n)){E^*}^2}{n}$$

By applying the sub-Gaussian tail inequality, we get that with probability at least $1 - \delta$,

$$|(P2)| \leq \sqrt{\frac{2(1 + c(n)){E^*}^2}{n} \log \frac{2}{\delta}}$$

Since this was true for all Z, this is true if we marginalize over Z as well, which completes the proof. □

Lemma F.5. *Let $(P3)$ be as defined in Lemma F.2. Suppose $|\hat{p_i} - p_i| < c(n)p_i < 0.5p_i$ for all i. Then with probability $\geq 1 - \delta$:*

$$|(P3)| \leq \frac{B}{2n} \log \frac{2B}{\delta}$$

Proof. Fix arbitrary $\hat{p_i}$s satisfying $|\hat{p_i} - p_i| < c(n)p_i < 0.5p_i$. Note that this gives us $\hat{p_i} > 0$.

By Hoeffding's bound, for any fixed i, with probability at least $1 - \frac{\delta}{b}$:

$$|\hat{y_i} - y_i^*| \leq \sqrt{\frac{1}{2\hat{p_i}n} \log \frac{2B}{\delta}}$$

Applying union bound over $i = 1, \cdots, B$, we get that the above holds *for all* i with probability at least $1 - \delta$. Then with probability at least $1 - \delta$, for all i:

$$|\hat{p_i}(\hat{y_i} - y_i^*)^2| \leq \frac{1}{2n} \log \frac{2B}{\delta}$$

Summing over the bins $i = 1, \cdots, B$, we get:

$$0 \leq (P3) \leq \frac{B}{2n} \log \frac{2B}{\delta}$$

□

Bounding the error of the plugin estimator simply involves combining the bounds for each of the terms, $P1$, $P2$, $P3$.

Theorem F.6. *Let $p_i = P(f(X) = s_i)$ and suppose $p_i > \frac{12}{n} \log \frac{2B}{\delta}$ for all i. Let $c(n)$ be defined as:*

$$c(n) = \sqrt{\frac{3}{n \min p_i} \log \frac{2B}{\delta}}$$

Then for the plugin estimator, with probability at least $1 - 3\delta$,

$$|\hat{E_{pl}}^2 - {E^*}^2| \leq c(n){E^*}^2 + \sqrt{\frac{2(1 + c(n)){E^*}^2}{n} \log \frac{2}{\delta}} + \frac{B}{2n} \log \frac{2B}{\delta}$$

Proof. We have:

$$|\hat{E}_{\text{pl}}^2 - {E^*}^2| \leq |P1 - {E^*}^2| + |P2| + |P3|$$

From Lemma F.1 we have $|\hat{p_i} - p_i| < c(n)p_i < 0.5p_i$ with probability $\geq 1 - \delta$. Conditioning on this, we combine Lemmas F.3, F.4, F.5 with union bound to get the desired result. □

We then prove the final bound for the plugin estimator, which we recall below.

Restatement of Theorem 5.3. *Suppose we have a binned model with L^2 calibration error ${E^*}^2 = \epsilon^2$, where the binning scheme is 2-well-balanced, that is for all $s \in S$, $\mathbb{P}(f(X) = s) \geq \frac{1}{2B}$.* [4] *If $n \geq c\frac{B}{\epsilon^2} \log \frac{B}{\delta}$ for some universal constant c then for the plugin estimator: $\frac{1}{2}{E^*}^2 \leq \hat{E}_{\text{pl}}^2 \leq \frac{3}{2}{E^*}^2$ with probability at least $1 - \delta$.*

Proof. This is now a simple corollary of Theorem F.6. For large enough constant c, where c is a constant independent of all the other variables, we choose $n = c\frac{B}{\epsilon^2} \log \frac{B}{\epsilon^2}$. Plugging it into the bound of Theorem F.6, we get the desired result, that $|\hat{E}^2 - {E^*}^2| \leq \frac{1}{2}{E^*}^2$. Notice that the dominating term is term $(P3)$ in Theorem F.6—we will see that the debiased estimator improves on this.

In fact, we can also show that in the worst case the plugin estimator will need at least $O(\frac{B}{\epsilon^2})$ samples to estimate the calibration error. To see this, first note that the bias of the plugin estimator, which comes from term $(P3)$ is at least $\frac{B}{n}$. Furthermore, in the analysis of the debiased estimator we show that the variance of this term is on the order of $O(\frac{\sqrt{B}}{n})$. So if $n < 0.1\frac{B}{\epsilon^2}$ we can consider very large B, and use Chebyshev to show that that with high probability the estimation error is larger than ϵ^2.

□

[4] We do not need the upper bound of the 2-well-balanced property.

F.2 Analysis of debiased estimator (proof of Theorem 5.4)

Next, we bound the error of the debiased estimator. The proof follows along the lines of the plugin estimator. We begin with a decomposition (Lemma F.7), similar to the decomposition of the plugin estimator. However, one of the terms in the decomposition, $C3$, is different. Lemma F.8 bounds this term $C3$. The rest of the proof is the same as for the plugin estimator, so we omit the other proofs.

As with the plugin estimator, we have a decomposition for the debiased estimator.

Lemma F.7 (Debiased decomposition)**.** *The debiased estimator satisfies the following decomposition:*

$$\hat{E}^2 = \underbrace{\sum_{i=1}^{b} \hat{p}_i e_i^2}_{(C1)} - \underbrace{2\sum_{i=1}^{b} \hat{p}_i e_i(\hat{y}_i - y_i^*)}_{(C2)} + \underbrace{\sum_{i=1}^{b} \hat{p}_i \Big[(\hat{y}_i - y_i^*)^2 - \frac{\hat{y}_i(1-\hat{y}_i)}{\hat{p}_i n - 1}\Big]}_{(C3)}$$

As with the plugin estimator, we bound each of the three terms. Notice that $C1$ and $C2$ are the same as terms $P1$ and $P2$ in the plugin estimator decomposition, so the bounds for those carry over. The next lemma bounds the error in $C3$.

Lemma F.8. *Let $(C3)$ be as defined in Lemma F.7. Suppose $|\hat{p}_i - p_i| < c(n)p_i < 0.5p_i$ for all i. Then with probability $\geq 1 - \delta$:*

$$|(C3)| \leq \frac{3\sqrt{B}}{n} \log \frac{n}{\delta} + \frac{\delta}{n}$$

Proof. Let $Z = (f(x_1), \cdots, f(x_n))$ be a random variable. We note that for all i, $\hat{p}_i$ is a deterministic function of Z. For convenience, define t_i as follows:

$$t_i = (\hat{y}_i - y_i^*)^2 - \frac{\hat{y}_i(1-\hat{y}_i)}{\hat{p}_i n - 1}$$

Computing the expectation: The debiased estimator debiases the plugin estimator. In particular, we briefly explain why $\mathbb{E}[C3 \mid Z] = 0$. Since $\hat{y}_i$ is the mean of $n\hat{p}_i$ draws of a Bernoulli with parameter y_i^*, we have:

$$\mathbb{E}[(\hat{y}_i - y_i^*)^2 \mid Z] = \frac{y_i^*(1-y_i^*)}{n\hat{p}_i}$$

The term we subtracted is the unbiased estimate of the standard deviation of the samples, so from elementary statistics:

$$\mathbb{E}\Big[\frac{\hat{y}_i(1-\hat{y}_i)}{\hat{p}_i n - 1} \mid Z\Big] = \frac{y_i^*(1-y_i^*)}{n\hat{p}_i}$$

Which implies that $\mathbb{E}[C3 \mid Z] = 0$.

Bounding each term: By Hoeffding's bound, for any fixed i, we get that with probability at least $1 - \frac{\delta}{n}$:

$$|\hat{y}_i - y_i^*| \leq \sqrt{\frac{1}{2\hat{p}_i n} \log \frac{2n}{\delta}}$$

Let E_i be the event that this is indeed the case. Condition on E_i holding for all i – by union bound this happens with probability at least $1 - \delta$. With some algebra, we then get:

$$|\hat{p}_i t_i| = \Big|\hat{p}_i \Big[(\hat{y}_i - y_i^*)^2 - \frac{\hat{y}_i(1-\hat{y}_i)}{\hat{p}_i n - 1}\Big]\Big| \leq \frac{3}{2n} \log \frac{B}{\delta}$$

Concentration: Next, we analyze the concentration of $T = \big[(C3) \mid Z, \forall i.E_i\big]$ around its mean μ. $|\hat{p}_i t_i|$ is bounded so is sub-Gaussian with parameter:

$$\sigma_i^2 = \frac{9}{4n^2} \log \frac{B}{\delta}$$

Each term $\hat{p_i}t_i$ in the sum is independent, even when conditioned on Z. So T is sub-Gaussian with parameter:

$$\sigma^2 = \sum_{i=1}^{B} \sigma_i^2 = \frac{9B}{4n^2} \log \frac{B}{\delta}$$

So by the sub-Gaussian tail bound, we have:

$$|T - \mu| \leq \sqrt{2\sigma^2 \log \frac{1}{\delta}} \leq \frac{3\sqrt{2}}{2} \frac{\sqrt{B}}{n} \sqrt{\log \frac{n}{\delta} \log \frac{1}{\delta}}$$

This can be simplified to:

$$|T - \mu| \leq \frac{3\sqrt{B}}{n} \log \frac{n}{\delta}$$

Bounding the bias: Although $\mathbb{E}[C3 \mid Z] = 0$, conditioning on E_i introduces some bias. However, we can show this bias is small. First, notice that $|t_i| \leq 1$. The event E_i holds with probability at least $1 - \frac{\delta}{n}$. Then by the law of total expectation, conditioning on E_i shifts the mean by at most $\frac{\delta}{n}$ – in other words $|\mathbb{E}[t_i \mid E_i, Z]| \leq \frac{\delta}{n}$. Summing over t_is, we get:

$$|\mathbb{E}[(C3) \mid Z, \forall i.E_i]| \leq \sum_{i=1}^{B} \hat{p_i} |\mathbb{E}[t_i \mid E_i, Z]| \leq \frac{\delta}{n}$$

Finishing up: Combining the bias and concentration, we get that with probability at least $1 - 2\delta$:

$$|(C3)| \leq \frac{3\sqrt{B}}{n} \log \frac{n}{\delta} + \frac{\delta}{n}$$

□

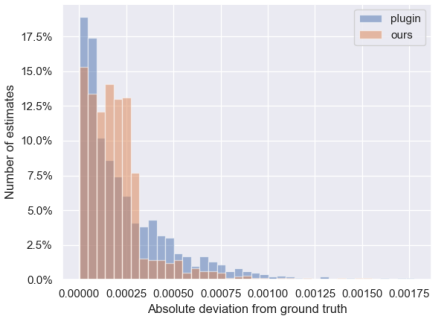
We combine the bounds for $(C1)$, $(C2)$, $(C3)$, as in Theorem F.6, to bound the estimation error of the debiased estimator.

Theorem F.9. *In the same setting as Theorem F.6, for the debiased estimator, with probability at least* $1 - 4\delta$*,*

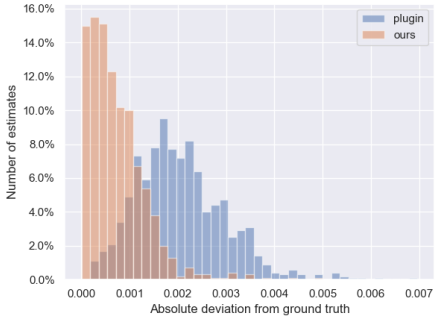
$$|\hat{E}^2 - {E^*}^2| \leq c(n){E^*}^2 + \sqrt{\frac{2(1 + c(n)){E^*}^2}{n} \log \frac{2}{\delta}} + \frac{3\sqrt{B}}{n} \log \frac{n}{\delta} + \frac{\delta}{n}$$

We interpret the bound in Theorem F.9 in two regimes. In the first regime, we fix the problem parameters $p_i, {E^*}^2$, and look at what happens as we send n to infinity. In that case, the second term dominates, and we see that the error is approximately proportional to $\frac{1}{\sqrt{n}}$, which is the same as for the plugin estimator. However, in general we do not need to estimate the calibration error extremely finely, and may be satisfied as long as we estimate the calibration error within a constant multiplicative factor. That is, we might only need n to be large enough so that our estimate $\hat{E}^2$ is on the right order, e.g. between $0.5{E^*}^2$ and $1.5{E^*}^2$ (where 0.5 and 1.5 can be replaced by other constants). In that regime, the third term dominates and the error is approximately proportional to $\frac{\sqrt{B}}{n}$, which is better than for the plugin estimator where it is proportional to $\frac{B}{n}$ (see Theorem F.6). This is captured in the final bound, where the proof closely parallels that of Theorem 5.3.

Restatement of Theorem 5.4. *Suppose we have a binned model with* L^2 *calibration error* ${E^*}^2 = \epsilon^2$ *and for all* $s \in S$, $\mathbb{P}(f(X) = s) \geq \frac{1}{2B}$*. If* $n \geq c\frac{\sqrt{B}}{\epsilon^2} \log \frac{B}{\delta}$ *for some universal constant* c *then for the debiased estimator:* $\frac{1}{2}{E^*}^2 \leq \hat{E}^2 \leq \frac{3}{2}{E^*}^2$ *with probability at least* $1 - \delta$*.*



(a) B = 10 bins



(b) $B = 100$ bins

Figure 12: Histograms of the absolute value of the difference between estimated and ground truth L^2 calibration errors (0 on the x-axis). For $B = 10$ bins, the results are mixed but we avoid very bad estimates. For $B = 100$ our estimates are much closer to ground truth.

G Additional experiments for section 5

In Section 5 we ran an experiment on CIFAR-10 to show that the debiased estimator gives estimates closer to the true calubration error than the plugin estimator. To give more insight into this, Figure 12 shows a histogram of the absolute difference between the estimates and ground truth for the plugin and debiased estimator, over the 1,000 resamples, when we use $B = 10$ or $B = 100$ bins. For $B = 10$ bins it is not completely clear which estimator is doing better but the debiased estimator avoids very bad estimates. However, when $B = 100$, the debiased estimator produces estimates much closer to the ground truth (0 on the x-axis).

We also ran a multiclass calibration experiment on CIFAR-10 to show that our estimator allows us to select models with a lower mean-squared error subject to a given calibration constraint. In this case we split the validation set into S_C and S_E of size 6000 and 4000 respectively, and recalibrated a trained model on S_C. On S_E, we estimate the calibration error using the plugin and debiased estimators and use 100 Bootstrap resamples to compute a 90% upper confidence bound on the estimate (from the variance of the Bootstrap samples). We compute the mean-squared error and the upper bounds on the calibration error for $B = 10, 15, \ldots, 100$ and show the Pareto curve in Figure 13. Figure 13 shows that for any desired calibration error, the debiased estimator enables us to pick out models with a better mean-squared error. For example, if we want a model with ℓ_2 calibration error less than 1.5%, the debiased estimator tells us we can confidently use 100 bins, while relying on the plugin estimator only lets us use 15 bins and incurs a 13% higher mean-squared error.

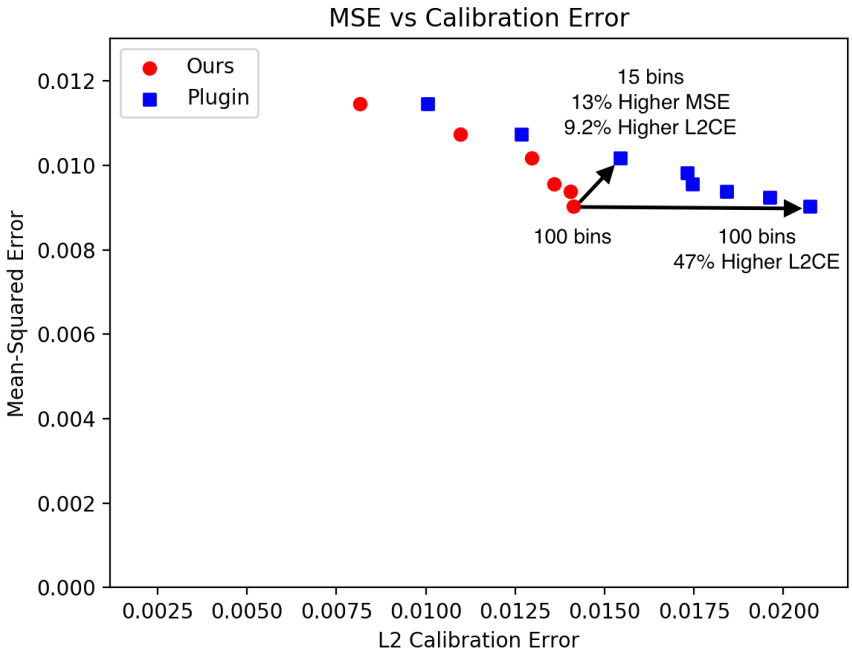


Figure 13: Plot of mean-squared error against 90% upper bounds on the calibration error computed by the debiased estimator and the plugin estimator, when we vary the number of bins B. For a given calibration error, our estimator enables us to choose models with a better mean-squared error. If we want a model with ℓ_2 calibration error less than 0.015, the debiased estimator tells us we can confidently use 100 bins, while relying on the plugin estimator only lets us use 15 bins and incurs a 13% higher mean-squared error.