Better Uncertainty Calibration via Proper Scores for Classification and Beyond

Sebastian Gruber
German Cancer Research Center (DKFZ), Germany
German Cancer Consortium (DKTK), Frankfurt, Germany
Goethe University Frankfurt, Germany
sebastian.gruber@dkfz.de

Florian Buettner
German Cancer Research Center (DKFZ), Germany
German Cancer Consortium (DKTK), Frankfurt, Germany
Frankfurt Cancer Institute, Germany
Goethe University Frankfurt, Germany
florian.buettner@dkfz.de

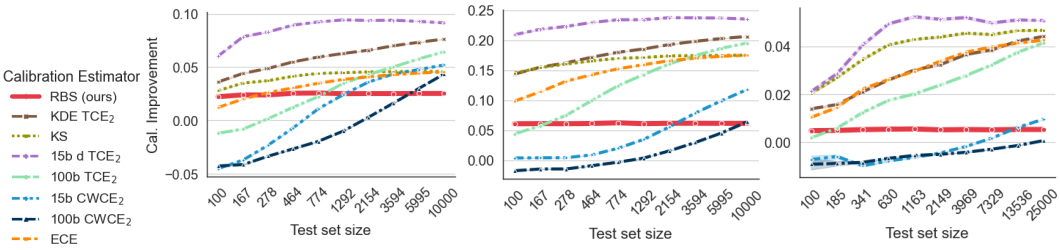
Abstract

With model trustworthiness being crucial for sensitive real-world applications, practitioners are putting more and more focus on improving the uncertainty calibration of deep neural networks. Calibration errors are designed to quantify the reliability of probabilistic predictions but their estimators are usually biased and inconsistent. In this work, we introduce the framework of *proper calibration errors*, which relates every calibration error to a proper score and provides a respective upper bound with optimal estimation properties. This relationship can be used to reliably quantify the model calibration improvement. We theoretically and empirically demonstrate the shortcomings of commonly used estimators compared to our approach. Due to the wide applicability of proper scores, this gives a natural extension of recalibration beyond classification.

1 Introduction

Deep learning became a dominant cornerstone of machine learning research in the last decade and deep neural networks can surpass human-level predictive performance on a wide range of tasks [19, 46, 17]. However, Guo et al. [14] have shown that for modern neural networks, better classification accuracy can come at the cost of systematic overconfidence in their predictions. Practitioners in sensitive forecasting domains, such as cancer diagnostics [17], genotype-based disease prediction [25] or climate prediction [57], require for models to not only have high predictive power but also to reliably communicate uncertainty. This raises the need to quantify and improve the quality of predictive uncertainty, ideally via a dedicated metric. An uncertainty-aware model should give probabilistic predictions which represent the true likelihood of events depending on the very prediction. To quantify the extend to which this condition is violated, calibration errors have been introduced. In general, their estimators are usually biased [45] and inconsistent [52]. This, in turn, is highly problematic since we cannot quantify how reliable a model is if we do not know how reliable the metric is. Especially the medical field is a domain that requires high model trustworthiness, but with low expert availability and/or disease frequency we often encounter a small data regime. Resampling strategies can be viable options for optimization on small datasets but also reduce the evaluation set

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) TS of DenseNet 40 on CIFAR10

(b) ETS of Wide ResNet 32 on CIFAR100

(c) DIAG of DenseNet 161 on ImageNet

Figure 1: Estimated calibration improvement for various settings. The calibration error is estimated before and after a recalibration method (TS / ETS / DIAG) is applied and the difference (i.e. calibration improvement) is shown for increasing test set size. All common calibration errors are sensitive with respect to the test set size and can substantially over- or underestimate the effect of performing recalibration. Only RBS robustly estimates the improvement in calibration error for all test set sizes.

size even more. We will discover that little data exacerbates the estimation bias and propose reliable alternatives for improving uncertainty calibration.

Since deep neural networks often yield uncalibrated confidence scores [36], a variety of different post-hoc recalibration approaches have been proposed [7, 31]. These methods use the validation set to transform predictions returned by a trained neural network such that they become better calibrated. A key desired property of recalibration methods is to not reduce the accuracy after the transformation. Therefore, most modern approaches are restricted to accuracy preserving transformations of the model outputs [42, 14, 61, 44]. When recalibrating a model, it is crucial to have a reliable estimate of how much the chosen method improves the underlying model. However, when using current estimators for calibration errors, their biased nature results in estimates that are highly sensitive to the number of samples in the test set that are used to compute the calibration error before and after recalibration (Fig. 1; c.f. section 5).

Our **contributions** for better uncertainty calibration are summarized in the following. We...

- ... give an overview of current calibration error literature, place the errors into a taxonomy, and show which are insufficient for calibration quantification. This also includes several theoretical results, which highlight the shortcomings of current approaches.
- ... introduce the framework of **proper calibration errors**, which gives important guarantees and relates every element to a proper score. We can reliably estimate the improvement of an injective recalibration method w.r.t. a proper calibration error via its related proper score - even in non-classification settings.
- ... show that common calibration estimators are highly sensitive w.r.t. the test set size. We demonstrate that for commonly used estimators, the estimated improvement of recalibration methods is heavily biased and becomes monotonically worse with fewer test data.

2 Related Work

In this section, we give an extensive overview of published work regarding quantifying model calibration and model recalibration. Important definitions will be directly given, while others are placed in Appendix C. These will be the basis for our theoretical findings. Further, we will motivate the definition of proper calibration errors, which are directly related to proper scores. Consequently, we will also present important aspects of the framework around proper scores in later parts of this section.

2.1 Calibration errors

For clarity, we introduce shortly the notation used throughout this work. Assume we have random variables X and Y corresponding to feature and target variable, and feature and target space $\mathscr{X}$ and

$\mathscr{Y}$. We have $\mathbb{P}_Y, \mathbb{P}_{Y|X} \in \mathscr{P}$, where $\mathbb{P}_Y$ refers to the distribution of Y, $\mathbb{P}_{Y|X}$ to the conditional distribution given X, and $\mathscr{P}$ a set of distributions on $\mathscr{Y}$. Even though some approaches explore calibration for regression tasks [47, 56, 6], it is most dominantly considered for classification. To distinguish between the general case and n-class classification, we refer to $\mathscr{P}_n$ as the n-dimensional simplex of corresponding categorical distributions.

A popular task is the calibration of the predicted top-label $C = \arg\max_k f_k(X)$ of a model $f\colon \mathscr{X} \to \mathscr{P}_n$ [14, 30, 34, 41, 52, 50, 44]. Here, the top-label confidence should represent the accuracy of this very prediction. Formally, we try to reach the condition $f_C(X) = \mathbb{P}(Y = C \mid f_C(X))$. However, the condition is weaker as one might expect, and referring to a model fulfilling this condition as (perfectly) calibrated can give a false sense of security [52, 55]. This holds especially in forecasting domains, where low probability estimates can still be highly relevant. For example, assigning probability mass to an aggressive type of cancer can still trigger action even if it is not predicted as the most likely outcome. Further, we might also be interested in predictive uncertainty for non-classification tasks. Consequently, we use the stricter and more general condition that the complete prediction $f(X)$ should be equal to the complete conditional distribution $\mathbb{P}_{Y|f(X)}$ of the target given this very prediction as introduced by Widmann et al. [56]. More formally, we state:

Definition 2.1. A model $f\colon \mathscr{X} \to \mathscr{P}$ is **calibrated** if and only if $f(X) = \mathbb{P}_{Y|f(X)}$.

Note that $\mathscr{P}$ can be a set of arbitrary distributions, which incorporates $\mathscr{P}_n$ (classification) as a special case.

One of the first metrics for assessing model calibration that is still widely used in recent literature is the Brier score (BS) [38, 42, 16, 44]. For a model $f\colon \mathscr{X} \to \mathscr{P}_n$ the **Brier score** [3] is defined as

$$\mathrm{BS}(f) = \mathbb{E}\left[\|f(X) - Y'\|_2^2\right], \tag{1}$$

where Y' is one-hot-encoded Y. The estimator of the BS is equivalent to the mean squared error, illustrating that it does not purely capture model calibration. Rather, the BS can be interpreted as a comprehensive measure of model performance, simultaneously capturing model fit and calibration. This becomes more obvious via the canonical decomposition of the BS into a calibration and sharpness term [38]. Based on this decomposition, we can derive the following error. For $1 \leq p \in \mathbb{R}$, the ℓ_p **calibration error** (CE_p) of model $f\colon \mathscr{X} \to \mathscr{P}_n$ is defined as

$$\mathrm{CE}_p(f) = \mathbb{E}\left[\left\|f(X) - \mathbb{P}_{Y|f(X)}\right\|_p^p\right]. \tag{2}$$

The BS decomposition only supports the case $p = 2$, but a general ℓ_p formulation became more common in recent years [52, 55, 61]. The CE_p is intractable to reliably estimate in the general case. This is mostly due to the term $\mathbb{P}_{Y|f(X)}$ since we never have samples of every possible prediction for continuous models. This is in contrast to the original work of Murphy [38], where only models with a finite prediction space are considered so that computing the CE_2 does not require a special procedure. To assess the calibration of a continuous binary model Platt [42] used histogram estimation, transforming the infinite prediction space to a finite one. This is also referred to as equal width binning. Similarly, Nguyen & O'Connor [40] introduced an equal mass binning scheme for continuous binary models. Both, equal width and equal mass binning schemes, suffer from the requirement of setting a hyperparameter. This can significantly influence the estimated value [31] and there is no optimal default since every setting has a different bias-variance tradeoff [41]. The first calibration estimator for a continuous one-vs-all multi-class model was given by Naeini et al. [39] and is still the most commonly used measure to quantify calibration. It is referred to as the **expected calibration error** (ECE) of model $f\colon \mathscr{X} \to \mathscr{P}_n$ and defined as

$$\mathrm{ECE}(f) = \sum_{i=1}^{m} p_i \left|\mathrm{conf}_i - \mathrm{acc}_i\right| \tag{3}$$

with the bin frequency $p_i = \mathbb{P}(f_C(X) \in B_i)$, the bin-wise mean confidence $\mathrm{conf}_i = \mathbb{E}[f_C(X) \mid f_C(X) \in B_i]$, and the bin-wise accuracy $\mathrm{acc}_i = \mathbb{P}(Y = C \mid f_C(X) \in B_i)$, for $m \in \mathbb{N}$ bins $B_i \coloneqq \left(\frac{i-1}{m}, \frac{i}{m}\right]$. We can reliably estimate p_i, conf_i, and acc_i via the respective means. These are then 'plugged-in' to estimate the ECE.

Kull et al. [29] and Nixon et al. [41] independently introduced another calibration estimator, which also captures the extend to which the condition $\mathbb{P}(Y = k \mid f_k(X)) = f_k(X)$ is violated for each

class $k \in \mathscr{Y}$. They respectively use equal width and equal mass binning. Similar to these estimators, Kumar et al. [31] introduced the **class-wise calibration error** (CWCE_p) and, similar to the ECE, the **top-label calibration error** (TCE_p). They only formulated the squared case $p = 2$ but suggested the extension of their definitions to general p-norms, which we provide in Appendix C.

Furthermore, Kumar et al. [31] and Vaicenavicius et al. [52] proved independently that using a fixed binning scheme for estimation leads to a lower bound of the expected error. Zhang et al. [61] circumvent binning schemes by using kernel density estimation to estimate the TCE_p.

Orthogonal ways to quantify model miscalibration have been proposed to not depend on binning or kernel density estimation schemes. Gupta et al. [16] introduced the **Kolmogorov-Smirnov calibration error** (KS) (c.f. Appendix C), which is based on the Kolmogorov-Smirnov test between empirical cumulative distribution functions. Its estimator does not require setting a hyperparameter but the authors did not provide further theoretical aspects.

Estimators of the TCE_p and CWCE_p are in general not differentiable. Consequently, Kumar et al. [30] proposed the **Maximum mean calibration error** (MMCE) (c.f. Appendix C), which has a differentiable estimator. It is a kernel-based error, comparing the top-label confidence and the conditional accuracy, similar to the ECE.

Widmann et al. [55] argued that the MMCE is insufficient for quantifying calibration of a model, similar as the ECE. They further proposed the **Kernel calibration error** (KCE) (c.f. Appendix C). It is based on matrix-valued kernels and unlike the MMCE, which only uses the top-label prediction, includes the whole model prediction. The squared KCE has an unbiased estimator based on a U-statistic with quadratic runtime complexity with respect to the data size. However, even though the KCE is positive, the U-statistic estimator can give negative values. To this end, the authors use the estimated value as a test statistic w.r.t. the null hypothesis that the model is calibrated.

Furthermore, Widmann et al. [55] and Widmann et al. [56] proposed to unify different definitions of calibration errors in a theoretical framework, which also includes variance regression calibration. However, the framework allows calibration errors, which are zero even if the model is not calibrated at all.

2.2 Recalibration

A plethora of recalibration methods have been proposed to improve model calibration after training by transforming the model output probabilities [18, 42, 58, 59, 39, 40, 28, 14, 29, 31, 61, 16, 44]. These methods are optimized on a specific calibration set, which is usually the validation set. Key desiderata of these methods include for the algorithms to be accuracy-preserving and data-efficient [61], reflecting that typical use-cases include settings in sensitive domains where accuracy should remain unchanged and often little data is available to train and evaluate the models. Such accuracy-preserving methods only adjust the probability estimate in such a way that the predicted top-label remains the same. The most commonly used accuracy-preserving recalibration method is temperature scaling (TS) [14], where the model logits are divided by a single parameter $T \in \mathbb{R}_{>0}$ before computing the predictions via softmax. A more expressive extension of TS is ensemble temperature scaling (ETS) [61], where a weighted ensemble of TS output, model output, and label smoothing is computed. Rahimi et al. [44] proposed different classes of order-preserving transformations. A specifically interesting one is the class of diagonal intra order-preserving functions (DIAG). Here, the model logits are transformed elementwise with a scalar, monotonic, and continuous function, which is represented by neural networks of unconstrained monotonic functions [54].

2.3 Proper scores

Gneiting & Raftery [13] give an extensive overview of proper scores. Unfortunately, their presented definitions assume maximization as the model training objective. To stay in line with recent machine learning literature, we flip the sign when it is required in the following definitions, similar as in [4]. We specifically do not constrain ourselves to classification, which is a special case. Assume we give a prediction in $\mathscr{P}$ for an event in $\mathscr{Y}$ and we want to score how good the prediction was. A function $S\colon \mathscr{P} \times \mathscr{Y} \to \overline{\mathbb{R}}$ with $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{-\infty, \infty\}$ is called **scoring rule** or just **score**. Examples are the Brier score and the log score for classification, or the Dawid-Sebastiani score (DSS), which extends the MSE for variance regression [8, 12]. It is defined as $S(P, y) = \frac{(\mu_P - y)^2}{\sigma_P^2} + \log \sigma_P^2$. To compare

distributions, we use the expected score $s_S \colon \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ defined as $s_S(P, Q) = \mathbb{E}_{Y \sim Q}[S(P, Y)]$. A scoring rule S is defined to be **proper** if and only if $s_S(P, Q) \geq s_S(Q, Q)$ holds for all $P, Q \in \mathscr{P}$, and **strictly proper** if and only if $P \neq Q \implies s_S(P, Q) > s_S(Q, Q)$. In other words, a score is proper if predicting the target distribution gives the best expected value and strictly proper if no other prediction can achieve this value. Given a proper score S and $P, Q \in \mathscr{P}$, the associated **divergence** $d_S \colon \mathscr{P} \times \mathscr{P} \to \mathbb{R}_{\geq 0}$ is defined as $d_S(P, Q) = s_S(P, Q) - s_S(Q, Q)$ and the associated **generalized entropy** $g_S \colon \mathscr{P} \to \mathbb{R}$ as $g_S(Q) = s_S(Q, Q)$. For strictly proper S, d_S is only zero if $P = Q$; for (strictly) proper S, g_S is (strictly) concave. An example of a divergence-entropy combination is the Kullback-Leibler divergence and the Shannon entropy associated to the log score.

Associated entropies and divergences are used in the calibration-sharpness decomposition introduced by Bröcker [4] for proper scores of categorical distributions. In Lemma 4.1 we will prove that this result holds for proper scores of arbitrary distributions, as long as additional conditions are met. Further, associated divergences will be the backbone for our definition of *proper calibration errors* in Section 4.

3 Theoretical analysis of calibration errors

In this section, we present theoretical results regarding calibration errors used in current literature. First, we place these calibration errors into a taxonomy, which we introduce in Theorem 3.1. Next, we give an example of how much errors lower in the hierarchy can differ from CE_2 in Proposition 3.2. We illustrate the practical relevance by introducing a perfect calibration method according to most errors in Proposition D.2, which fails to be calibrated. Last, we analyse the ECE estimate with respect to the data size in Proposition 3.3. This proposition is the basis for explaining the empirical (miss-)behaviour of the ECE observed in Section 5. All proofs are presented in Appendix D.

To the best of the authors' knowledge, other publications provided relations between at most two distinct calibration errors or none at all while introducing a new one. The taxonomy in the following theorem is a novel contribution, improving overview of a whole body of work regarding quantifying calibration.

Theorem 3.1. *Given a model $f : \mathscr{X} \to \mathscr{P}_n$ and $1 \leq p \in \mathbb{R}$, we have*

$$BS(f) = 0 \implies \begin{Bmatrix} CE_p(f) = 0 \\ KCE(f) = 0 \\ f \text{ calibrated} \end{Bmatrix} \implies CWCE_p(f) = 0 \implies \begin{Bmatrix} TCE_p(f) = 0 \\ MMCE(f) = 0 \\ KS(f) = 0 \end{Bmatrix} \implies ECE(f) = 0$$

where statements inside curly brackets $\{\ldots\}$ are equivalent. Further, we have

$$2^{\frac{1}{p}} \geq [BS(f)]^{\frac{1}{p}} \overset{*}{\geq} [CE_p(f)]^{\frac{1}{p}} \geq CWCE_p(f) \geq TCE_p(f) \geq TCE_1(f) \geq \begin{cases} KS(f) \\ ECE(f) \\ c \cdot MMCE(f) \end{cases} \geq 0$$

*where * BS is only included for $p = 2$. The kernel dependent constant $c \in \mathbb{R}$ is given in Appendix D.2 according to Kumar et al. [30].*

From this theorem follows that it is ambiguous to refer to *perfect calibration* just because there exists a calibration error which is zero for a model. Now, we further demonstrate how large the gap between calibration errors can be.

Proposition 3.2. *Assume the model $f \colon \mathscr{X} \to \mathscr{P}_n$ is surjective. There exists a joint distribution $\mathbb{P}_{X,Y}$ such that for all $E \in \{MMCE, KS, ECE, TCE_p, CWCE_p \mid 1 \leq p \in \mathbb{R}\}$:*

$$E(f) = 0 \quad \text{and} \quad CE_2(f) \geq 0.99 - \frac{1}{n}.$$

Recall that $\mathrm{CE}_2(f) = 0$ if and only if f is calibrated. In other words, most used calibration errors can be zero, but the model is still far from calibrated. An example of a model transformation with perfect ECE but uncalibrated outputs is given in Proposition D.2.

Among all calibration errors, the ECE is still the most commonly used [24, 26, 44, 50, 36, 49, 23, 35, 37, 15, 53, 10], even though its estimator is knowingly biased [31, 45]. Let $\hat{\mathrm{ECE}}_{(n)}$ denote the ECE

estimator for n data instances as originally defined in Guo et al. [14]. The following gives an analysis how this estimate behaves approximately with respect to n and how this is further impacted by the ground truth ECE value. For simplicity, we omit the fixed model from the notation.

Proposition 3.3. *For $\mu_{(n)}$ defined in Appendix D.5, we have*

$$\mathbb{E}\left[\hat{ECE}_{(n)}\right] \approx \mu_{(n)} \geq ECE \quad , \quad \frac{\mathrm{d}\mu_{(n)}}{\mathrm{d}n} < 0 \quad , \quad \frac{\mathrm{d}^2\mu_{(n)}}{(\mathrm{d}n)^2} > 0 \quad \textit{and} \quad \frac{\mathrm{d}^2\mu_{(n)}}{\mathrm{d}n\,\mathrm{d}ECE} > 0.$$

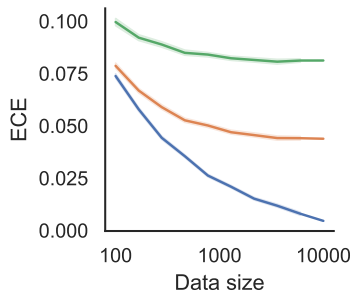


Figure 2: Estimated ECE of simulated models with perfect calibration (**blue**), mediocre calibration (**orange**), and bad calibration (**green**). Better calibration exacerbates ECE bias with respect to the data size, leading to unreliably calibration improvement quantification.

In words, the ECE estimator can be approximated by a function, which is strictly convex and monotonically decreasing in the data size. The smaller the data size, the larger the change of the function. Further, this change is also influenced by the true ECE value, such that, for low ground truth ECE, the function changes even more rapidly with the data size. Analogous results can be found for CWCE_1, CWCE_2, and TCE_2 with binning estimators as used in Kull et al. [29], Nixon et al. [41] and Kumar et al. [31].

To confirm the goodness of the approximation, we evaluated the ECE estimator on simulated models in Figure 2. The models are designed such that their true level of calibration is known. Including the results of Figure 1, the empirical curves are consistent with Proposition 3.3. Further details on the simulation are given in Appendix G.

The results in this section motivate a formal framework of proper calibration errors which are zero if and only if the model is calibrated and with robust estimators.

4 Proper calibration errors

In this section, we introduce the definition of *proper calibration errors*. We provide an easy-to-estimate upper bound and investigate some properties. As a preliminary step, we generalize a proper score decomposition. Again, all proofs are presented in Appendix D.

Bröcker [4] introduced a calibration-sharpness decomposition of proper scores w.r.t. categorical distributions. We extend this decomposition to proper scores of arbitrary distributions.

Lemma 4.1. *Let $\mathscr{P}$ be a set of arbitrary distributions for which exists a proper score S with some mild conditions. For random variables Q and Y with $Q, \mathbb{P}_Y, \mathbb{P}_{Y|Q} \in \mathscr{P}$, we have the decomposition*

$$\mathbb{E}\left[S\left(Q, Y\right)\right] = \underbrace{g_S\left(\mathbb{P}_Y\right)}_{\textit{generalized entropy}} - \underbrace{\mathbb{E}\left[d_S\left(\mathbb{P}_Y, \mathbb{P}_{Y|Q}\right)\right]}_{\textit{sharpness}} + \underbrace{\mathbb{E}\left[d_S\left(Q, \mathbb{P}_{Y|Q}\right)\right]}_{\textit{calibration}}.$$

Substituting Q with $f\left(X\right)$ and S with the Brier score, the calibration term equals the previously defined CE_2 of a model f. Lemma 4.1 motivates the following definition, which we introduce:

Definition 4.2. Given a model $f\colon \mathscr{X} \to \mathscr{P}$, we say

$$\mathrm{CE}_S\left(f\right) := \mathbb{E}\left[d_S\left(f\left(X\right), \mathbb{P}_{Y|f(X)}\right)\right]$$

is a **(strictly) proper calibration error** if and only if d_S is a divergence associated with a (strictly) proper score S.

This gives $\mathrm{CE}_{\mathrm{BS}} \equiv \mathrm{CE}_2$ as an example of a strictly proper calibration error for classification since the Brier score is a strictly proper score on $\mathscr{P}_n$. Strictly proper calibration errors have the highly desired property: $\mathrm{CE}_S\left(f\right) = 0 \overset{a.s.}{\iff} f$ is calibrated. Since proper scores are not restricted to classification, the above definition gives a natural extension of calibration errors beyond classification.

Additionally, by generalizing the definition of proper scores, we can show that the squared KCE is a strictly proper calibration error (Appendix F). But, in general, there does not exist an unbiased estimator of a proper calibration error, since we cannot estimate $\mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]$ in an unbiased manner. Because we do not want lower bounds for errors used in sensitive applications, we introduce the following theorem about how to construct an upper bound.

Theorem 4.3. *For all proper calibration errors with* $\inf_{P\in\mathscr{P}} g_S(P) \in \mathbb{R}$*, there exists an associated* ***calibration upper bound***

$$\mathscr{U}_S(f) \geq \mathit{CE}_S(f)$$

defined as $\mathscr{U}_S(f) = \mathbb{E}\left[S\left(f(X), Y\right)\right] - \inf_{P\in\mathscr{P}} g_S(P)$*. Under a classification setting and further mild conditions, we have* $\lim_{\mathit{ACC}(f)\to 1} \mathscr{U}_S(f) - \mathit{CE}_S(f) = 0$.

In other words, we can always construct a non-trivial upper bound of a proper calibration error as long as the generalized entropy function has a finite infimum. The calibration upper bound approaches the true calibration error for models with high accuracy. Our proposed calibration upper bounds are provably reliable to use since they all have a minimum-variance unbiased estimator. In the following example, we derive the calibration upper bound $\mathscr{U}_{\text{BS}}$ of the Brier Score.

Example 4.4. The scoring rule induced by the Brier score is defined as $S_{\text{BS}}(f(X), Y) = \|f(X) - Y'\|^2$, where Y' is the one-hot encoding of Y. Using the definition of the associated entropy gives us $g_{\text{BS}}(Q) = \mathbb{E}_{Y\sim Q}\left[S_{\text{BS}}(Q, Y)\right] = \mathbb{E}_{Y\sim Q}\left[\|Q - Y'\|^2\right]$.

To find its infimum, note that $\|.\|^2 \geq 0$ and $g_{\text{BS}}\left((1, 0, \ldots, 0)^\intercal\right) = 0$. Thus, $\inf_{P\in\mathscr{P}} g_{\text{BS}}(P) = 0$, which gives $\mathscr{U}_{\text{BS}}(f) = \mathbb{E}\left[\|f(X) - Y'\|^2\right] = \text{BS}(f)$. This makes the Brier score itself an upper bound of its induced calibration error.

By using Theorem 3.1 we receive $\sqrt{\mathscr{U}_{\text{BS}}(f)} = \sqrt{\text{BS}(f)} \geq \text{ECE}(f)$. Consequently, we might also be interested in transforming the estimated upper bound in a non-linear way for interpretability purposes. In general, any unbiased estimator $\hat{\theta}$ becomes biased after a non-linear transformation t, since $\mathbb{E}\left[t\left(\hat{\theta}\right)\right] \neq t\left(\mathbb{E}\left[\hat{\theta}\right]\right)$. But, if t is continuous, our estimator is still asymptotically unbiased and consistent [48].[1] We will further investigate the empirical robustness w.r.t. data size in Section 5 with t as the square root and $\sqrt{\mathscr{U}_{\text{BS}}(f)}$ as the root calibration upper bound (RBS). Given a dataset $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and a model f, we will estimate this quantity via $\sqrt{\mathscr{U}_{\text{BS}}(f)} \approx \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|f(X_i) - Y_i'\|^2}$.

Furthermore, $\mathscr{U}_S$ has the following properties, which are helpful for the application of recalibration method optimization and selection.

Proposition 4.5. *Given injective functions* $h, h' : \mathscr{P} \to \mathscr{P}$ *we have*

$$\mathscr{U}_S(h\circ f) - \mathscr{U}_S(f) = \mathit{CE}_S(h\circ f) - \mathit{CE}_S(f) \quad ,$$

$$\mathscr{U}_S(h\circ f) > \mathscr{U}_S(h'\circ f) \iff \mathit{CE}_S(h\circ f) > \mathit{CE}_S(h'\circ f)$$

and (assuming S *is differentiable)*

$$\frac{\mathrm{d}\mathscr{U}_S(h\circ f)}{\mathrm{d}h} = \frac{\mathrm{d}\mathit{CE}_S(h\circ f)}{\mathrm{d}h}.$$

This is a generalization of Proposition 4.2 presented in [61]. It tells us that we can reliably estimate the improvement of any injective recalibration method via the upper bound. Furthermore, we get access to the calibration gradient and can compare different transformations. At first, injectivity seems like a significant restriction. But, we argue in the following that injectivity - rather than being accuracy-preserving - is a desired property of general recalibration methods. For example, we can construct a recalibration method, which is calibrated and accuracy-preserving, but only predicts a finite set of distinct values (see Appendix E). Specifically, we would only predict two distinct values for any input in binary classification. To exclude such naive solutions which substantially

[1] follows from Continuous Mapping Theorem and Theorem 3.2.6 of Takeshi [48]

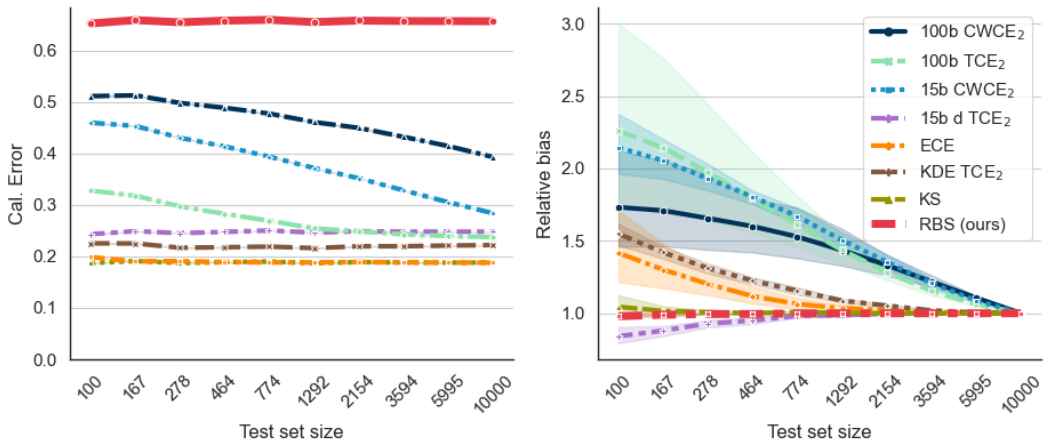


Figure 3: **Left:** Different calibration error estimates versus the test set size of ResNet Wide 32 and CIFAR100. The red line corresponds to the square root of the Brier score which is an upper bound of the true $\sqrt{\text{CE}_2}$. The other errors are lower bounds. **Right:** Relative change versus data size with respect to error at full size. Averaging across a multitude of models shows a systematic trend. An unbiased estimator would give a flat line.

reduce model sharpness, we restrict ourselves to injective transformations of $\mathscr{P}_n \to \mathscr{P}_n$. These do provably not impact the model sharpness and preserve, at least partly, the continuity of the output space. Examples of injective transformations are TS, ETS, and DIAG. These state-of-the-art methods show very competitive performances even when compared to non-injective recalibration methods [61, 44]. Further, Proposition 4.5 also holds when replacing $\mathscr{U}_S$ with the expected score s_S without further conditions. This is useful when $\mathscr{U}_S$ does not exist, but we still want to perform recalibration as in Section 5 in the case of the DSS.

5 Experiments

In the following, we investigate the behavior of calibration error estimators in three settings.
First, we use varying test set sizes for the estimators and compare their values. This will show how well the inequalities in Theorem 3.1 hold in practical settings and how robust the estimators are. Second, we explore what the estimated improvements of several recalibration methods are. This is done after the recalibration methods are already optimized on a given validation set; we only vary the size of the test set and compute calibration errors on these test sets before and after recalibration. In both settings, the straighter a line is, the more robust and, consequently, trustworthy is the estimator for practical applications. Third, we investigate how our framework can be used to improve calibration for tasks beyond classification by performing probabilistic regression with subsequent recalibration.

In all experiments we evaluate the following estimators: CWCE$_2$ with 15 equal width bins ('15b CWCE$_2$'), CWCE$_2$ with 100 equal width bins ('100b CWCE$_2$'), ECE with 15 equal width bins ('ECE'), TCE$_2$ with 100 equal width bins ('100b TCE$_2$'), TCE$_2$ with 15 equal mass bins and debias term ('15b d TCE$_2$'), TCE$_2$ with kernel density estimation ('KDE TCE$_2$'), KS ('KS') and the root calibration upper bound $\sqrt{\mathscr{U}_{\text{BS}}}$ ('RBS (ours)').The bin amounts are chosen based on past literature [31, 36]. The experiments are conducted across several model-dataset combinations, for which logit sets are openly accessible [2] [29, 44]. This includes the models Wide ResNet 32 [60], DenseNet 40, and DenseNet 161 [22] and the datasets CIFAR10, CIFAR100 [27], and ImageNet [9]. We did not conduct model training ourselves and refer to [29] and [44] for further details. We include TS, ETS, and DIAG as injective recalibration methods. Further details and results on additional models and datasets are reported in the Appendix G.

[2] https://github.com/markus93/NN_calibration/ and https://github.com/AmirooR/IntraOrderPreservingCalibration

Table 1: Comparing the mean squared error (MSE) with the average predicted variance (Avg Var) before and after recalibration for various training iterations. Recalibration gives a better average match between prediction and real error.

Iteration	500	750	1000	1250	1500	1750	2000
MSE	10.48	5.87	4.3	3.51	3.12	2.74	2.57
Avg Var (Calibrated)	11.04	6.89	5.4	4.55	4.11	3.63	3.32
Avg Var (Uncalibrated)	0.83	0.83	0.83	0.83	0.83	0.82	0.82

Robustness of calibration errors to test set size We illustrate the estimated values of our introduced upper bound and the other errors, which are lower bounds of the unknown CE_2 on the left of Figure 3. On the right, we aggregate across several models to show the systematic drop-off according to Proposition 3.3. The relative bias is computed by $Error(n)/Error(10000)$ and allows an aggregation of models with different calibration levels. Included models are DenseNet 40, Wide ResNet 32, ResNet 110 SD, ResNet 110, and LeNet 5, all trained on CIFAR10. All values represent the calibration of the given model without recalibration transformation. Only our proposed upper bound and KS are stable, and Appendix G shows this holds across a wide range of different settings. The theoretically highest lower bound ($CWCE_2$ with 100 bins) is also constantly the highest estimated lower bound, but it is sensitive to the test set size. Results for further settings presented in Appendix G show similar results.

Quantifying recalibration improvement Next, we assessed how well all estimators were able to quantify the improvement in calibration error after applying different injective recalibration methods (Fig. 1). Only our proposed upper bound estimator RBS is again robust throughout all settings. According to Proposition 4.5 and since RBS is asymptotically unbiased and consistent, it can be regarded as a reliable approximation of the real improvement of the presented recalibration methods. For all other estimators, there is a general trend to estimate recalibration improvement higher for large test set sizes. In other settings, especially for small test set sizes, calibration improvement is underestimated to the extent that negative improvements (poorer calibration than before) are suggested. Results on other settings presented in Appendix G show similar results. Taken together, these experiments demonstrate the unreliability of existing calibration estimators, in particular, when used to evaluate recalibration methods. In contrast, our proposed upper bound estimator is stable across different settings.

Variance regression calibration We consider variance regression to demonstrate the usefulness of proper calibration errors outside the classification setting. To this end, we predict sales prices with an uncertainty estimate in the UCI dataset *Residential Building*, which consists of a high feature (107) to data instances (372) ratio [43]. Our model of choice is a fully-connected mixture density network predicting mean and variance [2]. Similar to classification, we are interested in recalibration of the predicted variance to adjust possible under- or overconfidence. We use our proposed framework to derive a proper calibration error induced by the proper score DSS for recalibration. Further, we compare DSS [12] to squared KCE (SKCE) [56] and analyze the average predicted variance throughout model training. We use Platt scaling ($x \mapsto wx + b$ with parameters $w, b \in \mathbb{R}$ [42]) on the predicted variance in each training iteration to show how recalibration benefits uncertainty awareness. We expect high uncertainty awareness at the start of the model training with a drop-off at later iterations. As can be seen in Figure 4, recalibration is able to adjust the uncertainty estimate of a model as desired. Further, the DSS estimate, which captures predicted mean and variance correctness, directly communicates the improved variance fit. Contrary, the SKCE estimate appears more erratic between iteration steps and seemingly ignores the changes in variance and the recalibration improvement.

One might also be interested at how the predicted variance corresponds to the mean squared error throughout model training. As we can see in Table 1, only after calibration using a proper calibration error, the average predicted variance (Avg Var) corresponds to the mean squared error.

Next, to assess how well the predicted variance corresponds to the instance-level error, we compute the ratio between the squared error of the predictive mean μ and the predicted variance σ^2 (SE Var Ratio) for each individual sample via $\frac{1}{n}\sum_{i=1}^{n}\frac{(\mu_i - y_i)^2}{\sigma_i^2}$.An SE Var Ratio of '1' for a given

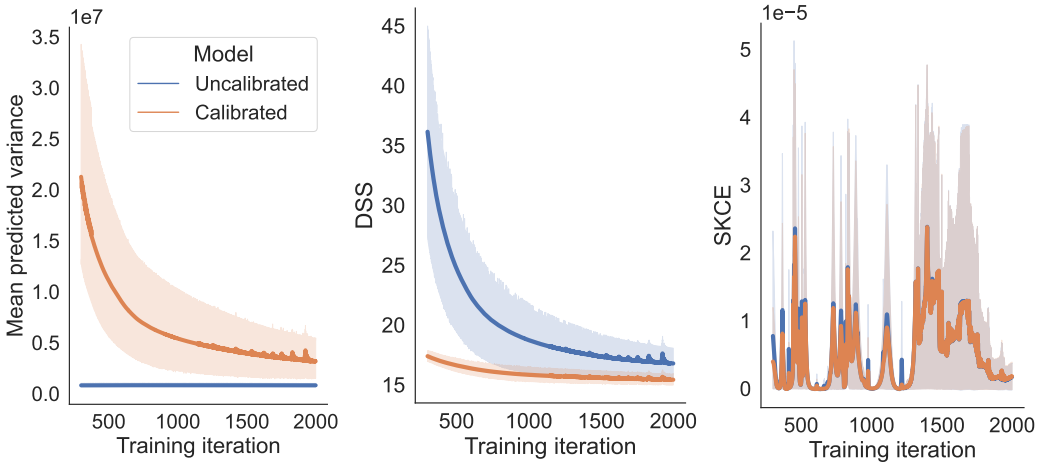


Figure 4: **Left:** Average predicted variance throughout model training before and after recalibration. Initially, due to a bad fit, recalibration adjusts the variance accordingly for better communicated uncertainty. Once the model fit improves, the predicted variance requires less adjustment due to less uncertainty in each prediction. **Middle:** DSS communicates reasonably changes in the variance due to recalibration. **Right:** SKCE fails to capture the variance trend and behaves erratically.

Table 2: Instance level ratio between the squared error and the predicted variance before and after recalibration for various training iterations. Recalibration improves the instance level prediction of the squared error.

Iteration	500	1000	1500	2000
SE Var Ratio (Calibrated)	0.82 ± 2.17	0.79 ± 2.43	0.79 ± 2.56	0.79 ± 2.59
SE Var Ratio (Uncalibrated)	11.33 ± 30.72	5.36 ± 15.09	4.07 ± 12.13	3.51 ± 11.22

instance means that the predictive uncertainty (variance) exactly matches the squared error, an SE Var Ratio of '10' means that the model is overconfident and the squared error is ten times as large as the predicted variance. As we can see in Table 2, recalibration through our framework gives consistently conservative estimates on the squared error, whereas the uncalibrated uncertainties are highly overconfident (with errors more than ten times larger than the prediction).

We perform further variance regression experiments in Appendix G.

6 Conclusion

In this work, we address the problem of reliably quantifying the effect of recalibration on predictive uncertainty for classification and other probabilistic tasks. This is critical for adjusting under- or overconfidence via recalibration. To this end, we first provide a taxonomy of existing calibration errors. We discover that most errors are lower bounds of a proper calibration error and fail to assess if a model is calibrated. This motivates our definition of *proper calibration errors*, which provides a general class of errors for arbitrary probabilistic predictions. Since proper calibration errors cannot be estimated in the general case, we introduce upper bounds, which directly measure the calibration change for injective transformations. This allows us to reliably adjust model uncertainty via recalibration. We demonstrate theoretically and empirically that the estimated calibration improvement can be highly misleading for commonly used estimators, including the ECE. In stark contrast, our upper bound is robust to changes in data size and estimates robustly the improvement via injective recalibration. We further show in additional experiments that our approach can be applied successfully to variance regression.

References

[1] Aitchison, J. and Shen, S. M. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2): 261–272, 1980. ISSN 00063444. URL http://www.jstor.org/stable/2335470.

[2] Bishop, C. M. Mixture density networks. 1994.

[3] Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1 – 3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

[4] Bröcker, J. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, Jul 2009. ISSN 1477-870X. doi: 10.1002/qj.456. URL http://dx.doi.org/10.1002/qj.456.

[5] Capiński, M. and Kopp, P. E. *Measure, integral and probability*, volume 14. Springer, 2004.

[6] Chung, Y., Neiswanger, W., Char, I., and Schneider, J. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification, 2021.

[7] Dawid, A. P. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379): 605–610, 1982. doi: 10.1080/01621459.1982.10477856. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477856.

[8] Dawid, A. P. and Sebastiani, P. Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, pp. 65–81, 1999.

[9] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[10] Fan, H., Ferianc, M., Que, Z., Niu, X., Rodrigues, M. L., and Luk, W. Accelerating bayesian neural networks via algorithmic and hardware optimizations. *IEEE Transactions on Parallel and Distributed Systems*, 2022.

[11] Friedman, J. H. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

[12] Gneiting, T. and Katzfuss, M. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1: 125–151, 2014.

[13] Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

[14] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

[15] Gupta, A., Kvernadze, G., and Srikumar, V. Bert & family eat word salad: Experiments with text understanding. *ArXiv*, abs/2101.03453, 2021.

[16] Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., and Hartley, R. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2020.

[17] Haggenmüller, S., Maron, R. C., Hekler, A., Utikal, J. S., Barata, C., Barnhill, R. L., Beltraminelli, H., Berking, C., Betz-Stablein, B., Blum, A., Braun, S. A., Carr, R., Combalia, M., Fernandez-Figueras, M.-T., Ferrara, G., Fraitag, S., French, L. E., Gellrich, F. F., Ghoreschi, K., Goebeler, M., Guitera, P., Haenssle, H. A., Haferkamp, S., Heinzerling, L., Heppt, M. V., Hilke, F. J., Hobelsberger, S., Krahl, D., Kutzner, H., Lallas, A., Liopyris, K., Llamas-Velasco, M., Malvehy, J., Meier, F., Müller, C. S., Navarini, A. A., Navarrete-Dechent, C., Perasole, A., Poch, G., Podlipnik, S., Requena, L., Rotemberg, V. M., Saggini, A., Sangueza, O. P., Santonja, C., Schadendorf, D., Schilling, B., Schlaak, M., Schlager, J. G., Sergon, M., Sondermann, W., Soyer, H. P., Starz, H., Stolz, W., Vale, E., Weyers, W., Zink, A., Krieghoff-Henning, E., Kather, J. N., von Kalle, C., Lipka, D. B., Fröhling, S., Hauschild, A., Kittler, H., and Brinker, T. J. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European Journal of Cancer*, 156:202–216, 2021. ISSN 0959-8049. doi: https://doi.org/10.1016/j.ejca.2021.06.049. URL https://www.sciencedirect.com/science/article/pii/S0959804921004445.

[18] Hastie, T. and Tibshirani, R. Classification by pairwise coupling. *The annals of statistics*, 26(2):451–471, 1998.

[19] He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[20] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[21] Hoeffding, W. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. doi: 10.1214/aoms/1177730196. URL `https://doi.org/10.1214/aoms/1177730196`.

[22] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[23] Islam, A., Chen, C.-F., Panda, R., Karlinsky, L., Radke, R. J., and Feris, R. S. A broad study on the transferability of visual representations with contrastive learning. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8825–8835, 2021.

[24] Joo, T., Chung, U., and Seo, M. Being bayesian about categorical probability. In *ICML*, 2020.

[25] Katsaouni, N., Tashkandi, A., Wiese, L., and Schulz, M. H. Machine learning based disease prediction from genotype data. *Biological Chemistry*, 402(8):871–885, 2021. doi: doi:10.1515/hsz-2021-0109. URL `https://doi.org/10.1515/hsz-2021-0109`.

[26] Kristiadi, A., Hein, M., and Hennig, P. Being bayesian, even just a bit, fixes overconfidence in relu networks. In *ICML*, 2020.

[27] Krizhevsky, A. et al. Learning multiple layers of features from tiny images. 2009.

[28] Kull, M., Filho, T. S., and Flach, P. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 623–631. PMLR, 20–22 Apr 2017. URL `https://proceedings.mlr.press/v54/kull17a.html`.

[29] Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in Neural Information Processing Systems*, 32:12316–12326, 2019.

[30] Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2805–2814. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/kumar18a.html`.

[31] Kumar, A., Liang, P., and Ma, T. Verified uncertainty calibration. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 3792–3803, 2019.

[32] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[33] Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.-J., Fei-Fei, L., Yuille, A., Huang, J., and Murphy, K. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 19–34, 2018.

[34] Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32:13153–13164, 2019.

[35] Menon, A. K., Rawat, A. S., Reddi, S. J., Kim, S., and Kumar, S. A statistical perspective on distillation. In *ICML*, 2021.

[36] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[37] Morales-Álvarez, P., Hernández-Lobato, D., Molina, R., and Hernández-Lobato, J. M. Activation-level uncertainty in deep neural networks. In *ICLR*, 2021.

[38] Murphy, A. H. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595 – 600, 1973. doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml.

[39] Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2901–2907. AAAI Press, 2015. ISBN 0262511290.

[40] Nguyen, K. and O'Connor, B. Posterior calibration and exploratory analysis for natural language processing models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1587–1598, 2015.

[41] Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

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

[43] Rafiei, M. H. and Adeli, H. A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management*, 142(2):04015066, 2016.

[44] Rahimi, A., Shaban, A., Cheng, C.-A., Hartley, R., and Boots, B. Intra order-preserving functions for calibration of multi-class neural networks. *Advances in Neural Information Processing Systems*, 33: 13456–13467, 2020.

[45] Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. Mitigating bias in calibration error estimation, 2021.

[46] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[47] Song, H., Diethe, T., Kull, M., and Flach, P. Distribution calibration for regression. In *International Conference on Machine Learning*, pp. 5897–5906. PMLR, 2019.

[48] Takeshi, A. *Advanced econometrics*. Harvard university press, 1985.

[49] Tian, J., Yung, D., Hsu, Y.-C., and Kira, Z. A geometric perspective towards neural calibration via sensitivity decomposition. In *NeurIPS*, 2021.

[50] Tomani, C., Gruber, S., Erdem, M. E., Cremers, D., and Buettner, F. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10124–10132, June 2021.

[51] Tsagris, M., Beneki, C., and Hassani, H. On the folded normal distribution. *Mathematics*, 2(1):12–28, feb 2014. doi: 10.3390/math2010012. URL https://doi.org/10.3390%2Fmath2010012.

[52] Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3459–3467. PMLR, 2019.

[53] Wang, X., Liu, H., Shi, C., and Yang, C. Be confident! towards trustworthy graph neural networks via confidence calibration. In *NeurIPS*, 2021.

[54] Wehenkel, A. and Louppe, G. Unconstrained monotonic neural networks. *Advances in Neural Information Processing Systems*, 32:1545–1555, 2019.

[55] Widmann, D., Lindsten, F., and Zachariah, D. Calibration tests in multi-class classification: A unifying framework. *Advances in Neural Information Processing Systems*, 32:12257–12267, 2019.

[56] Widmann, D., Lindsten, F., and Zachariah, D. Calibration tests beyond classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=-bxf89v3Nx.

[57] Yen, M.-H., Liu, D.-W., Hsin, Y.-C., Lin, C.-E., and Chen, C.-C. Application of the deep learning for the prediction of rainfall in southern taiwan. *Scientific reports*, 9(1):1–9, 2019.

[58] Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *ICML*, 1, 05 2001.

[59] Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pp. 694–699, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 158113567X. doi: 10.1145/775047.775151. URL https://doi.org/10.1145/775047.775151.

[60] Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.

[61] Zhang, J., Kailkhura, B., and Han, T. Y.-J. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, pp. 11117–11128. PMLR, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We mention required conditions for the proofs either directly or refer to Appendix D.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We see positive societal impacts from improved uncertainty awareness via recalibration.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Major conditions are directly stated; minor conditions in Appendix D as referred.
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix D.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Full code is located in the supplementary material and further experimental details are in Appendix G.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Relevant training details are located in Appendix G. A lot of models are not trained by ourselves. Instead, we loaded their results from publicly accessible URLs. We refer to each URL and the original work, where the training was conducted.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We repeated experiments until the error bars are barely visible or not visible anymore. Only exception is Figure 4, where aggregations hinder visibility due to high variances in training different seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix G.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We used assets from [29] and Rahimi et al. [44] located in https://github.com/markus93/NN_calibration/ and https://github.com/AmirooR/IntraOrderPreservingCalibration .
   (b) Did you mention the license of the assets? [No] Each lincense can be viewed in the respective github repository.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include the code for reproducing the experiments and figures in the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The data is publicly accessible and we cited each respective source.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The data is often abstract in nature (we used pretrained logits provided by a publicly accessible source) or downloaded the dataset from the UCI dataset database.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] No crowdsourcing or human subjects

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No crowdsourcing or human subjects

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] No crowdsourcing or human subjects

A Overview

In this appendix we

- Introduce some notation in section B that we will use throughout the appendix.
- Give rigorous definitions of calibration errors omitted in the main paper in Section C
- Provide proofs for all claims that we make in the main text in Section D.
- Provide details for specific recalibration transformation that illustrate the shortcomings of existing approaches (Section E).
- Give a detailed overview of proper U-scores that can be used to further generalize our proposed framework of proper calibration errors (Section F).
- Give more experimental details and report results from additional experiments (Section G).

B Notation

The following is implied throughout the appendix. We will use

- The underlying probability space $(\Omega, \mathscr{F}, \mathbb{P})$, $\mathscr{X}$ the feature space, and $\mathscr{Y}$ the target space.
- Random variables $X \colon \Omega \to \mathscr{X}$ and $Y \colon \Omega \to \mathscr{Y}$.
- $\mathbb{P}_{Y|X=x}(y) \coloneqq \frac{\mathbb{P}(\{\omega \in \Omega | X(\omega)=x \wedge Y(\omega)=y\})}{\mathbb{P}(\{\omega \in \Omega | X(\omega)=x\})}$ and $\mathbb{P}_Y(y) \coloneqq \mathbb{P}(\{\omega \in \Omega \mid Y(\omega) = y\})$ for $x \in \mathscr{X}$ and $y \in \mathscr{Y}$.
- $\mathbb{P}_Y, \mathbb{P}_{Y|X=x} \in \mathscr{P}_n$ with $\mathscr{P}_n = \{p \in [0,1]^n \mid \sum_k p_k = 1\}$, and $\mathscr{Y} = \{1, \ldots, n\}$ for categorical Y with $n \in \mathbb{N}$ classes.
- The index '$-k$' on a finite vector to denote the removal of index k.
- The random variable $C \colon \Omega \to \mathscr{Y}$ defined as $C \coloneqq \arg\max_k f_k(X)$ for $f \colon \mathscr{X} \to \mathscr{P}_n$. It can be regarded as the top-label prediction of f.

The notation regarding the (conditional) probability measures will be used for arbitrary random variables.

C Definitions

A systematic overview of the multitude of calibration errors proposed in the recent literature requires a common notation that can be used to harmonize definitions. For the sake of clarity, we use formulations close to the notation introduced in Kumar et al. [31] and adjust the other errors accordingly, while retaining the notation of the original work whenever possible.

We follow Kumar et al. [31] and define top-label and class-wise calibration errors in expectation:

Definition C.1. The **top-label calibration error** of model $f \colon \mathscr{X} \to \mathscr{P}_n$ is defined as

$$\text{TCE}_p(f) = \left(\mathbb{E}\left[|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))|^p\right]\right)^{\frac{1}{p}}$$

with $C \coloneqq \arg\max_k f_k(X)$ and the **class-wise calibration error** is defined as

$$\text{CWCE}_p(f) = \left(\sum_{k \in \mathscr{Y}} \mathbb{E}\left[|f_k(X) - \mathbb{P}(Y = k \mid f_k(X))|^p\right]\right)^{\frac{1}{p}}$$

for $1 \leq p \in \mathbb{R}$.

Note that we removed the weighting factors from the original definition in Kumar et al. [31] for easier comparison with the other errors and a fixed upper limit (we will show that $\text{CWCE}_p \leq 2^{\frac{1}{p}}$).

Definition C.2. The **Kolmogorov-Smirnov calibration error** [16] of model $f \colon \mathscr{X} \to \mathscr{P}_n$ is given by

$$\text{KS}(f) = \mathbb{E}\left[\text{KS}(f, C)\right],$$

where $C = \arg\max_k f_k(X)$ and $\text{KS}(f, k) = \max_{\sigma \in [0,1]} \left| \int_{[0,\sigma]} z - \mathbb{P}(Y = k \mid f_k(X) = z) \, \mathrm{d}\mathbb{P}_{f_k(X)}(z) \right|$.

Definition C.3. Given a reproducing kernel Hilbert space $\mathscr{H}$ with kernel $k\colon [0,1]\times[0,1]\to\mathbb{R}$ the **maximum mean calibration error** [30] of model $f\colon \mathscr{X}\to\mathscr{P}_n$ is

$$\begin{aligned}&\mathrm{MMCE}\left(f\right)=\\&\quad\left\|\mathbb{E}\left[\left(f_C\left(X\right)-\mathbb{P}\left(Y=C\mid f_C\left(X\right)\right)\right)k\left(f_C\left(X\right),.\right)\right]\right\|_{\mathscr{H}}.\end{aligned}$$

Definition C.4. Given a reproducing kernel Hilbert space $\mathscr{H}$ with kernel $k\colon \mathscr{P}_n\times\mathscr{P}_n\to\mathbb{R}^n\times\mathbb{R}^n$ the **kernel calibration error** [55] of model $f\colon \mathscr{X}\to\mathscr{P}_n$ is

$$\mathrm{KCE}\left(f\right)=\left\|\mathbb{E}\left[\left(f\left(X\right)-\mathbb{P}_{Y|f(X)}\right)k\left(f\left(X\right),.\right)\right]\right\|_{\mathscr{H}}.$$

We also need the following score related definitions in the proofs. These are simply a repetition from the main paper.

Definition C.5. Given a proper score S and $P,Q\in\mathscr{P}$, the expected score $s_S\colon \mathscr{P}\times\mathscr{P}\to\mathscr{R}$ is defined as $s_S\left(P,Q\right)=\mathbb{E}_{Y\sim Q}\left[S\left(P,Y\right)\right]=\int_{\mathscr{Y}}S\left(P,y\right)\mathrm{d}Q\left(y\right)$.

Definition C.6. Given a proper score S and $P,Q\in\mathscr{P}$, the associated **divergence** $d_S\colon \mathscr{P}\times\mathscr{P}\to\mathbb{R}_{\geq 0}$ is defined as $d_S\left(P,Q\right)=s_S\left(P,Q\right)-s_S\left(Q,Q\right)$ and the associated **generalized entropy** $g_S\colon \mathscr{P}\to\mathbb{R}$ as $g_S\left(Q\right)=s_S\left(Q,Q\right)$.

D Proofs

D.1 Helpers

The following will be of use in several proofs.

Lemma D.1. *Assume that S is a proper score for which CE_S exists, then we have*

$$CE_S\left(f\right)=\mathbb{E}\left[S\left(f\left(X\right),Y\right)\right]-\mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right].$$

Proof.

$$\begin{aligned}\mathrm{CE}_S\left(f\right)&\overset{\text{def 4.2}}{=}\mathbb{E}\left[d_S\left(f\left(X\right),\mathbb{P}_{Y|f(X)}\right)\right]\\&\overset{\text{def }C.6}{=}\mathbb{E}\left[s_S\left(f\left(X\right),\mathbb{P}_{Y|f(X)}\right)-s_S\left(\mathbb{P}_{Y|f(X)},\mathbb{P}_{Y|f(X)}\right)\right]\\&\overset{\text{def }C.6}{=}\mathbb{E}\left[s_S\left(f\left(X\right),\mathbb{P}_{Y|f(X)}\right)\right]-\mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\&=\int s_S\left(z,\mathbb{P}_{Y|f(X)=z}\right)\mathrm{d}\mathbb{P}_{f(X)}\left(z\right)-\mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\&\overset{\text{def }C.5}{=}\int\int S\left(z,y\right)\mathrm{d}\mathbb{P}_{Y|f(X)=z}\left(y\right)\mathrm{d}\mathbb{P}_{f(X)}\left(z\right)-\mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\&=\int S\left(z,y\right)\mathrm{d}\mathbb{P}_{Y,f(X)}\left(y,z\right)-\mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\&=\mathbb{E}\left[S\left(f\left(X\right),Y\right)\right]-\mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\end{aligned}\tag{4}$$

□

D.2 Theorem 3.1

Given a model $f : \mathscr{X} \to \mathscr{P}_n$ and the above defined errors, we have

$$\begin{aligned} &\text{BS}(f) = 0 \\ \implies &\text{CE}_p(f) = 0 \\ \iff &\text{KCE}(f) = 0 \\ \iff &f \text{ is calibrated} \\ \implies &\text{CWCE}_p(f) = 0 \\ \implies &\text{TCE}_p(f) = 0 \\ \iff &\text{MMCE}(f) = 0 \\ \iff &\text{KS}(f) = 0 \\ \implies &\text{ECE}(f) = 0 \end{aligned} \tag{5}$$

and

$$\begin{aligned} 2^{\frac{1}{p}} &\geq (\text{BS}(f))^{\frac{1}{p}} \\ &\overset{*}{\geq} (\text{CE}_p(f))^{\frac{1}{p}} \\ &\geq \text{CWCE}_p(f) \\ &\geq \text{TCE}_p(f) \\ &\geq \text{TCE}_1(f) \\ &\geq \begin{cases} \text{KS}(f) \\ \text{ECE}(f) \\ c \cdot \text{MMCE}(f) \end{cases} \geq 0 \end{aligned} \tag{6}$$

for $1 \leq p \in \mathbb{R}$. * BS is only included for $p = 2$. We define $c = (\max_r k(r,r))^{-\frac{1}{2}}$ as given in Theorem 3 of Kumar et al. [30].

Proof. **Regarding** $\text{BS}(f) = 0 \implies \text{CE}_p(f) = 0 \iff \text{KCE}(f) = 0 \iff f$ is calibrated:

$$\begin{aligned} \text{BS}(f) = 0 &\iff \mathbb{P}_{Y|X} \overset{\text{a.s.}}{=} f(X) \\ &\implies \mathbb{P}_{Y|f(X)} \overset{\text{a.s.}}{=} f(X) \\ &\iff \begin{cases} \text{CE}_p(f) = 0 \\ \text{KCE}(f) = 0 \\ f \text{ is calibrated} \end{cases} \end{aligned} \tag{7}$$

The last equivalence follows from Definition 2.1 and 2, and according to Widmann et al. [55]. Since the equivalence in the last line holds for each, it follows $\text{CE}_p(f) = 0 \iff \text{KCE}(f) = 0 \iff f$ is calibrated. Example sketch for $\text{BS}(f) = 0 \not\Leftarrow \text{CE}_p(f) = 0$: Set $f(.) = \mathbb{P}_Y$, then $f(X) = \mathbb{P}_Y = \mathbb{P}_{Y|\mathbb{P}_Y} = \mathbb{P}_{Y|f(X)}$, but $\text{BS}(f) > 0$.

Regarding $\text{CE}_p(f) = 0 \implies \text{CWCE}_p(f) = 0$:

$$\begin{aligned} \text{CE}_p(f) = 0 &\iff \mathbb{P}_{Y|f(X)} \overset{\text{a.s.}}{=} f(X) \\ &\iff \mathbb{P}(Y = k \mid f(X)) \overset{\text{a.s.}}{=} f_k(X) \quad \forall k \\ &\implies \mathbb{E}_{f_{-k}(X)}[\mathbb{P}(Y = k \mid f(X)) \mid f_k(X)] \overset{\text{a.s.}}{=} \mathbb{E}_{f_{-k}}[f_k(X) \mid f_k(X)] \quad \forall k \\ &\iff \mathbb{P}(Y = k \mid f_k(X)) \mid f_k(X) \overset{\text{a.s.}}{=} f_k(X) \quad \forall k \\ &\iff \sum_{k \in \mathscr{Y}} \mathbb{E}[(\mathbb{P}(Y = k \mid f_k(X)) - f_k(X))^p] = 0 \\ &\iff \text{CWCE}_p(f) = 0 \end{aligned} \tag{8}$$

An example for $\text{CE}_p(f) = 0 \not\Longleftarrow \text{CWCE}_p(f) = 0$ is given in Proposition 3.2.

Regarding $\text{CWCE}_p(f) = 0 \implies \text{TCE}_p(f) = 0 \iff \text{MMCE}(f) = 0$:

$$\begin{aligned}
\text{CWCE}_p(f) = 0 &\iff \mathbb{P}(Y = k \mid f_k(X)) \overset{\text{a.s.}}{=} f_k(X) \quad \forall k \\
&\implies \mathbb{P}(Y = C \mid f_C(X)) \overset{\text{a.s.}}{=} f_C(X) \\
&\iff \mathbb{P}\left(Y = \arg\max_k f_k(X) \mid \max_k f_k(X)\right) \overset{\text{a.s.}}{=} \max_k f_k(X) \\
&\iff \begin{cases} \text{TCE}_p(f) = 0 \\ \text{MMCE}(f) = 0 \end{cases}
\end{aligned} \tag{9}$$

See Theorem 1 in Kumar et al. [30] regarding MMCE. Note that we could not verify their claim that MMCE is a proper score, which is even contradictive to our findings. A sketch for an example where $\text{CWCE}_p(f) = 0 \not\Longleftarrow \text{TCE}_p(f) = 0$ is if $\mathbb{P}(Y = C \mid f_C(X)) \overset{\text{a.s.}}{=} f_C(X)$ and $\mathbb{P}(Y = \arg\min_k f_k(X) \mid \min_k f_k(X)) \neq \min_k f_k(X)$.

Regarding $\text{TCE}_p(f) = 0 \iff \text{KS}(f) = 0$:

$$\begin{aligned}
&\text{TCE}_p(f) = 0 \\
&\iff \mathbb{P}\left(Y = \arg\max_k f_k(X) \mid \max_l f_l(X)\right) \overset{\text{a.s.}}{=} \max_m f_m(X) \\
&\iff \mathbb{P}(Y = C \mid f_C(X)) \overset{\text{a.s.}}{=} f_C(X) \\
&\overset{(i)}{\iff} \int_{\sigma'} \mathbb{P}(Y = C \mid f_C(X) = z)\, \mathrm{d}\mathbb{P}_{f_C(X)|C}(z) \overset{\text{a.s.}}{=} \int_{\sigma'} z \mathrm{d}\mathbb{P}_{f_C(X)|C}(z), \quad \forall \sigma' \subset [0,1] \\
&\iff \int_{[0,\sigma]} \mathbb{P}(Y = C \mid f_C(X) = z)\, \mathrm{d}\mathbb{P}_{f_C(X)|C}(z) \overset{\text{a.s.}}{=} \int_{[0,\sigma]} z \mathrm{d}\mathbb{P}_{f_C(X)|C}(z), \quad \forall \sigma \in [0,1] \\
&\iff \mathbb{E}\left[\max_{\sigma \in [0,1]} \left| \int_{[0,\sigma]} z - \mathbb{P}(Y = C \mid f_C(X) = z)\, \mathrm{d}\mathbb{P}_{f_C(X)|C}(z) \right| \right] = 0 \\
&\iff \mathbb{E}[\text{KS}(f, C)] = 0 \\
&\iff \text{KS}(f) = 0
\end{aligned} \tag{10}$$

(i) according to Theorem 4.22 of Capiński & Kopp [5].

Regarding $\text{TCE}_p(f) = 0 \implies \text{ECE}(f) = 0$:

$$\begin{aligned}
&\text{TCE}_p(f) = 0 \\
&\iff \mathbb{P}(Y = C \mid f_C(X)) \overset{\text{a.s.}}{=} f_C(X) \\
&\overset{\text{(i)}}{\implies} \forall i = 1, \ldots, m\colon\ \mathbb{P}(Y = C \mid f_C(X) \in B_i) \overset{\text{a.s.}}{=} \mathbb{E}[f_C(X) \mid f_C(X) \in B_i] \\
&\overset{\text{def 3}}{\iff} \text{ECE}(f) = 0
\end{aligned} \tag{11}$$

(i) with B_i defined as in definition 3; follows since $\mathbb{P}(Y = C \mid f_C(X) \in B_i) = \int_{B_i} \mathbb{P}(Y = C \mid f_C(X) = z)\, \mathrm{d}\mathbb{P}_{f_C(X)}(z) \overset{\text{a.s.}}{=} \int_{B_i} f_C(X)\, \mathrm{d}\mathbb{P}_{f_C(X)}(z) = \mathbb{E}[f_C(X) \mid f_C(X) \in B_i]$.

An intuition of why $\text{TCE}_1(f) = 0 \not\Longleftarrow \text{ECE}(f) = 0$ is given in example 3.2 of Kumar et al. [31].

Regarding $2 \geq \text{BS}(f) \geq \text{CE}_2(f)$:

$$
\begin{aligned}
2 &= \|e_1 - e_2\|_2^2 \\
&\geq \mathbb{E}\left[\|f(X) - e_Y\|_2^2\right] \\
&\overset{\text{def 1}}{=} \text{BS}(f) \\
&\overset{\text{(i)}}{\geq} \text{BS}(f) - \mathbb{E}\left[g_{\text{BS}}\left(\mathbb{P}_{Y|f(X)}\right)\right] \\
&\overset{\text{le } D.1}{=} \text{CE}_{\text{BS}}(f) \\
&\overset{\text{(ii)}}{=} \text{CE}_2(f)
\end{aligned} \tag{12}
$$

(i) g_{BS} non-negative, follows from definition C.6.
(ii) compare definition 2 with the squared calibration term in [38].

Regarding $2 \geq \text{CE}_p(f)$:

$$
\begin{aligned}
2 &= \|e_1 - e_2\|_p^p \\
&\geq \mathbb{E}\left[\left\|f(X) - \mathbb{P}_{Y|f(X)}\right\|_p^p\right] \\
&\overset{\text{def 2}}{=} \text{CE}_p(f)
\end{aligned} \tag{13}
$$

Regarding $(\text{CE}_p(f))^{\frac{1}{p}} \geq \text{CWCE}_p(f)$:

In the following, we will use Tonelli's theorem to split the expectation into two and the Jensen's inequality for the convex function $|.|^p$.

$$
\begin{aligned}
\text{CE}_p(f) &= \mathbb{E}\left[\left\|f(X) - \mathbb{P}_{Y|f(X)}\right\|_p^p\right] \\
&= \sum_{k \in \mathscr{Y}} \mathbb{E}\left[|f_k(X) - \mathbb{P}(Y = k \mid f(X))|^p\right] \\
&\overset{\text{Tonelli}}{=} \sum_{k \in \mathscr{Y}} \mathbb{E}_{f_k(X)}\left[\mathbb{E}_{f_{-k}(X)}\left[|f_k(X) - \mathbb{P}(Y = k \mid f(X))|^p \mid f_k(X)\right]\right] \\
&\overset{\text{Jensen}}{\geq} \sum_{k \in \mathscr{Y}} \mathbb{E}_{f_k(X)}\left[\left|\mathbb{E}_{f_{-k}(X)}\left[f_k(X) - \mathbb{P}(Y = k \mid f(X)) \mid f_k(X)\right]\right|^p\right] \\
&= \sum_{k \in \mathscr{Y}} \mathbb{E}_{f_k(X)}\left[|f_k(X) - \mathbb{P}(Y = k \mid f_k(X))|^p\right] \\
&\overset{\text{def } C.1}{=} (\text{CWCE}_p(f))^p
\end{aligned} \tag{14}
$$

Regarding $\text{CWCE}_p(f) \geq \text{TCE}_p(f)$:

We will use $F \coloneqq f(X)$ for shorter notation.

$$
\begin{aligned}
(\text{CWCE}_p(f))^p &\overset{\text{def } C.1}{=} \sum_{k \in \mathscr{Y}} \mathbb{E}_{f_k(X)} \left[|f_k(X) - \mathbb{P}(Y = k \mid f_k(X))|^p \right] \\
&= \sum_{k \in \mathscr{Y}} \mathbb{E}_{F_k} \left[|F_k - \mathbb{P}(Y = k \mid F_k)|^p \right] \\
&= \sum_{k \in \mathscr{Y}} \mathbb{E}_{F} \left[|F_k - \mathbb{P}(Y = k \mid F_k)|^p \right] \\
&= \mathbb{E}_F \left[\sum_{k \in \mathscr{Y}} |F_k - \mathbb{P}(Y = k \mid F_k)|^p \right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_F \left[\sum_{k \in \mathscr{Y}} \left| F_{(k)_F}(X) - \mathbb{P}\left(Y = (k)_F \mid F_{(k)_F}\right) \right|^p \right] \\
&\geq \mathbb{E}_F \left[\left| F_{(1)_F} - \mathbb{P}\left(Y = (1)_F \mid F_{(1)_F}\right) \right|^p \right] \\
&\overset{\text{(ii)}}{=} \mathbb{E}_F \left[|F_C - \mathbb{P}(Y = C \mid F_C)|^p \right] \\
&= \mathbb{E}_{f(X)} \left[|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))|^p \right] \\
&\overset{\text{def } C.1}{=} (\text{TCE}_p(f))^p
\end{aligned}
\tag{15}
$$

(i) Order all summands by F. We use notation of order statistics to refer to $(k)_F$ the index with the kth highest rank according to F.
(ii) From (i) follows $(1)_F = (1)_{f(X)} = \arg\max_k f_k(X) = C$.

Regarding $\text{TCE}_p(f) \geq \text{TCE}_1(f)$:

Let $p \geq q \geq 1$. This makes $(.)^{\frac{p}{q}}$ a convex function for positive arguments. We will show the more general $\text{TCE}_p(f) \geq \text{TCE}_q(f)$. From this directly follows $\text{TCE}_p(f) \geq \text{TCE}_1(f)$.

$$
\begin{aligned}
&\text{TCE}_p(f) \\
&= \left(\mathbb{E}\left[|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))|^p \right] \right)^{\frac{1}{p}} \\
&= \left(\mathbb{E}\left[|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))|^{q \frac{p}{q}} \right] \right)^{\frac{1}{p}} \\
&\overset{\text{Jensen}}{\geq} \left(\mathbb{E}\left[|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))|^q \right] \right)^{\frac{1}{p}\frac{p}{q}} \\
&= \left(\mathbb{E}\left[|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))|^q \right] \right)^{\frac{1}{q}} \\
&= \text{TCE}_q(f)
\end{aligned}
\tag{16}
$$

Regarding $\text{TCE}_1(f) \geq \text{KS}(f)$:

We will show the more general $\text{TCE}_p(f) \geq \text{KS}(f)$, from which $\text{TCE}_1(f) \geq \text{KS}(f)$ follows.

We will make use of the indicator function for a set A defined as $\mathbb{1}_A(a) = \begin{cases} 1, & a \in A \\ 0, & \text{else.} \end{cases}$

$$
\begin{aligned}
(\mathrm{TCE}_p(f))^p &= \mathbb{E}\left[\left|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right|^p\right] \\
&\overset{\text{Tonelli}}{=} \mathbb{E}_C\left[\mathbb{E}_{f_C(X)}\left[\left|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right|^p \mid C\right]\right] \\
&= \mathbb{E}_C\left[\mathbb{E}_{f_C(X)}\left[\mathbb{1}_{[0,1]}(f_C(X))\left|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right|^p \mid C\right]\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_C\left[\max_{\sigma\in[0,1]} \mathbb{E}_{f_C(X)}\left[\mathbb{1}_{[0,\sigma]}(f_C(X))\left|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right|^p \mid C\right]\right] \\
&= \mathbb{E}_C\left[\max_{\sigma\in[0,1]} \mathbb{E}_{f_C(X)}\left[\left|\mathbb{1}_{[0,\sigma]}(f_C(X))\left(f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right)\right|^p \mid C\right]\right] \\
&\overset{\text{Jensen}}{\geq} \mathbb{E}_C\left[\max_{\sigma\in[0,1]} \left|\mathbb{E}_{f_C(X)}\left[\mathbb{1}_{[0,\sigma]}(f_C(X))\left(f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right) \mid C\right]\right|^p\right] \\
&= \mathbb{E}_C\left[\max_{\sigma\in[0,1]} \left|\int_{[0,1]} \mathbb{1}_{[0,\sigma]}(z)\left(z - \mathbb{P}(Y = C \mid f_C(X) = z)\right) \mathrm{d}\mathbb{P}_{f_C(X)|C}(z)\right|^p\right] \\
&= \mathbb{E}_C\left[\max_{\sigma\in[0,1]} \left|\int_{[0,\sigma]} z - \mathbb{P}(Y = C \mid f_C(X) = z) \mathrm{d}\mathbb{P}_{f_C(X)|C}(z)\right|^p\right] \\
&\overset{\text{Jensen}}{\geq} \left(\mathbb{E}_C\left[\max_{\sigma\in[0,1]} \left|\int_{[0,\sigma]} z - \mathbb{P}(Y = C \mid f_C(X) = z) \mathrm{d}\mathbb{P}_{f_C(X)|C}(z)\right|\right]\right)^p \\
&\overset{\text{def } C.2}{=} \left(\mathbb{E}_C\left[\mathrm{KS}(f, C)\right]\right)^p \\
&\overset{\text{def } C.2}{=} (\mathrm{KS}(f))^p
\end{aligned}
\tag{17}
$$

(i) $\sigma \geq \sigma' \implies \mathbb{1}_{[0,\sigma]}(f_C(X))\left|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right|^p \geq \mathbb{1}_{[0,\sigma']}(f_C(X))\left|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right|^p \geq 0$.

Regarding $\mathrm{TCE}_1(f) \geq c \cdot \mathrm{MMCE}(f)$:

This is given in the proof of Theorem 3 of Kumar et al. [30]. Note that Kumar et al. [30] used ECE in their theorem, but their proof is actually given for TCE_1. Since $\mathrm{ECE}(f) = 0 \not\Rightarrow \mathrm{MMCE}(f) = 0$, we have $\mathrm{ECE}(f) \not\geq c \cdot \mathrm{MMCE}(f)$.

Regarding $\mathrm{TCE}_1(f) \geq \mathrm{ECE}(f)$:

A similar statement for binary models is given in Proposition 3.3 of Kumar et al. [31] or for general models in Theorem 2 of Vaicenavicius et al. [52]. Since our formulations differ, we provide an independent proof.

Let $\mathscr{B} \coloneqq \sigma\left(\left\{\left(0, \frac{1}{m}\right], \ldots, \left(\frac{m-1}{m}, 1\right]\right\}\right)$ be the σ-algebra generated by the binning scheme of size $m \in \mathbb{N}$ used for the ECE.

$$
\begin{aligned}
\mathrm{TCE}_1(f) &= \mathbb{E}\left[\left|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right|\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left|f_C(X) - \mathbb{P}(Y = C \mid f_C(X))\right| \mid \mathscr{B}\right]\right] \\
&\overset{\text{(i)}}{\geq} \mathbb{E}\left[\left|\mathbb{E}\left[f_C(X) - \mathbb{P}(Y = C \mid f_C(X)) \mid \mathscr{B}\right]\right|\right] \\
&= \mathbb{E}\left[\left|\mathbb{E}\left[f_C(X) \mid \mathscr{B}\right] - \mathbb{P}(Y = C \mid \mathbb{E}\left[f_C(X) \mid \mathscr{B}\right])\right|\right] \\
&= \sum_{i=1}^{m} \mathbb{P}\left(f(X) \in \left(\frac{i-1}{m}, \frac{i}{m}\right]\right)\left|\mathbb{E}\left[f_C(X) \mid f(X) \in \left(\frac{i-1}{m}, \frac{i}{m}\right]\right] - \mathbb{P}\left(Y = C \mid f(X) \in \left(\frac{i-1}{m}, \frac{i}{m}\right]\right)\right| \\
&\overset{\text{def } 3}{=} \mathrm{ECE}(f)
\end{aligned}
\tag{18}
$$

(i) We use conditional Jensen's inequality [5].

□

D.3 Proposition 3.2

For all $\epsilon > 0$ and surjective $f\colon \mathscr{X} \to \mathscr{P}_n$ there exists a joint distribution $\mathbb{P}_{X,Y}$ such that for all $E \in \{\text{MMCE}, \text{KS}, \text{ECE}, \text{TCE}_p, \text{CWCE}_p \mid 1 \leq p \in \mathbb{R}\}$:

$$E(f) = 0 \quad \wedge \quad \text{CE}_2(f) \geq 1 - \frac{1}{n} - \epsilon.$$

Proof. Assume arbitrary $\epsilon > 0$ and surjective $f\colon \mathscr{X} \to \mathbb{P}_n$. Choose $\mathbb{P}_{X,Y}$ such that $\mathbb{E}\left[\|f(X)\|_2^2\right] \leq \frac{1}{n} + \epsilon$ and

$$\mathbb{P}(Y = k \mid f(X)) = \begin{cases} 1 & \text{, with probability } f_k(X) \\ 0 & \text{, else.} \end{cases}$$

This is possible, since $\|.\|_2^2\colon \mathscr{P}_n \to \left[\frac{1}{n}, 1\right]$ and f are surjective, from which folows $\forall \epsilon > 0\ \exists x \in \mathscr{X}\colon \quad \frac{1}{n} + \epsilon \geq \|f(x)\|_2^2$.

Write $F \coloneqq f(X)$ and $\mathbf{Y} \coloneqq e_Y$ (one-hot encoded Y).

Then we have $\mathbb{P}(Y = k \mid F_k) = \mathbb{E}[\mathbf{Y}_k \mid F_k] = \mathbb{E}_{F_{-k}}[\mathbb{E}[\mathbf{Y}_k \mid F] \mid F_k] = F_k$ and consequently $\text{CWCE}_p(f) = 0$. The other errors follow from Theorem 3.1. But we also have

$$\begin{aligned}
\text{CE}_2(f) &= \mathbb{E}\left[\left\|\mathbb{P}_{Y|f(X)} - f(X)\right\|_2^2\right] \\
&= \mathbb{E}\left[\|\mathbb{E}[\mathbf{Y} \mid F] - F\|_2^2\right] \\
&= \sum_{k \in \mathscr{Y}} \mathbb{E}\left[(\mathbb{E}[\mathbf{Y}_k \mid F] - F_k)^2\right] \\
&= \sum_{k \in \mathscr{Y}} \mathbb{E}\left[(\mathbb{E}[\mathbf{Y}_k \mid F])^2\right] - 2\mathbb{E}[\mathbb{E}[\mathbf{Y}_k \mid F] F_k] + \mathbb{E}\left[F_k^2\right] \\
&= \sum_{k \in \mathscr{Y}} \mathbb{E}[\mathbb{E}[\mathbf{Y}_k \mid F]] - 2\mathbb{E}[\mathbb{E}[\mathbf{Y}_k \mid F] F_k] + \mathbb{E}\left[F_k^2\right] \\
&= 1 - 2\sum_{k \in \mathscr{Y}} \mathbb{E}[\mathbb{E}[\mathbf{Y}_k \mid F] F_k] + \sum_{k \in \mathscr{Y}} \mathbb{E}\left[F_k^2\right] \\
&= 1 - 2\sum_{k \in \mathscr{Y}} \mathbb{E}[\mathbb{E}[\mathbf{Y}_k \mid F_k] F_k] + \sum_{k \in \mathscr{Y}} \mathbb{E}\left[F_k^2\right] \\
&= 1 - 2\sum_{k \in \mathscr{Y}} \mathbb{E}\left[F_k^2\right] + \sum_{k \in \mathscr{Y}} \mathbb{E}\left[F_k^2\right] \\
&= 1 - \sum_{k \in \mathscr{Y}} \mathbb{E}\left[F_k^2\right] \\
&= 1 - \mathbb{E}\left[\|F\|_2^2\right] \\
&\geq 1 - \frac{1}{n} - \epsilon
\end{aligned} \tag{19}$$

□

D.4 Proposition D.2

Proposition 3.2 tells us about the existence of settings such that common errors are insufficient to capture miscalibration. We might still wonder how likely it is to encounter such a situation in practice. Indeed, we can come up with a recalibration transformation that is *perfect* according to these errors and accuracy-preserving but not calibrated. For this, assume that $f\colon \mathscr{X} \to \mathscr{P}_n$ is a trained model. Define $t^f\colon \mathscr{P}_n \to \mathscr{P}_n$ to replace the largest entry in its input with the accuracy of model f. The other entries are set such that the output is a unit vector. A more formal definition is provided in the proof.

Proposition D.2. *For all models $f\colon \mathscr{X} \to \mathscr{P}_n$ and $E \in \{MMCE, KS, ECE, TCE_p \mid 1 \leq p \in \mathbb{R}\}$ we have*

$$E\left(t^f \circ f\right) = 0 \quad \textit{and} \quad ACC\left(t^f \circ f\right) = ACC\left(f\right).$$

But, $t^f \circ f$ is not calibrated in general.

Proof. Assume we are given a model $f\colon \mathscr{X} \to \mathscr{P}_n$.

Define $\sigma\colon \mathscr{P}_n \times \mathscr{P}_n \to \mathscr{P}_n$ to order the entries of its second input according to the values given in the first input. Let $\sigma^{-1}\colon \mathscr{P}_n \times \mathscr{P}_n \to \mathscr{P}_n$ revert the ordering in the second input according to the entries of its first input. For easier notation, we will write $\sigma_u\left(v\right) \coloneqq \sigma\left(u, v\right)$ and $\sigma_u^{-1}\left(v\right) \coloneqq \sigma^{-1}\left(u, v\right)$, which gives $\forall u, v \in \mathscr{P}\colon\ \sigma_u^{-1} \circ \sigma_u\left(v\right) = v$. I.e. σ_u^{-1} is the inverse of σ_u given u.

Define $c_f \coloneqq \left(\text{ACC}\left(f\right), \frac{1-\text{ACC}(f)}{n-1}, \ldots, \frac{1-\text{ACC}(f)}{n-1}\right)^\intercal \in \mathscr{P}_n$.

Now, we can give a formal definition of t^f, which is defined as $t^f\left(p\right) = \sigma_p^{-1}\left(c_f\right)$.

Regarding accuracy:

We will use $[.]_k$ to denote entry with index k of the expression inside the brackets. Since we can assume $\text{ACC}\left(f\right) > \frac{1-\text{ACC}(f)}{n-1}$ in every practical setting, we have

$$\begin{aligned}
&\arg\max_k t_k^f \circ f\left(X\right) \\
&= \arg\max_k \left[\sigma_{f(X)}^{-1}\left(c_f\right)\right]_k \\
&\overset{(i)}{=} \arg\max_k \left[\sigma_{f(X)}^{-1} \circ \sigma_{f(X)}\left(f\left(X\right)\right)\right]_k \\
&= \arg\max_k \left[f\left(X\right)\right]_k \\
&= \arg\max_k f_k\left(X\right).
\end{aligned} \tag{20}$$

(i) c_f and $\sigma_{f(X)}\left(f\left(X\right)\right)$ have their largest entry at index $k = 1$.

This states that t^f is accuracy-preserving.

Regarding zero TCE:

Note that $\text{ACC}(f) = \mathbb{P}\left(Y = \arg\max_k f_k\left(X\right)\right)$. Using this, we have $\mathbb{P}\left(Y = \arg\max_k t_k^f \circ f\left(X\right) \mid \max_k t_k^f \circ f\left(X\right)\right) = \mathbb{P}\left(Y = \arg\max_k f_k\left(X\right) \mid \text{ACC}\left(f\right)\right) = \mathbb{P}\left(Y = \arg\max_k f_k\left(X\right)\right) = \text{ACC}\left(f\right) = \max_k t_k^f \circ f\left(X\right)$. It follows $\text{TCE}_p\left(t^f \circ f\right) = 0$.

Proof for the other errors follows from Theorem 3.1.

□

Even though t^f is the perfect transformation according to ECE and accuracy, it is not the correct choice if the whole model prediction is relevant and supposed to be calibrated.

D.5 Proposition 3.3

We will write $\hat{Y} = \arg\max_k f_k\left(X\right)$ for the top-label prediction of classifier f.

Define

$$\mu_{(n)} = \sum_{i=1}^{m} p_i \left\{\sqrt{\frac{2}{\pi}}\sigma_i \exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right) + \mu_i\left[1 - 2\Phi\left(-\frac{\mu_i}{\sigma_i}\right)\right]\right\} \tag{21}$$

with

$$\mu_i = \underbrace{\mathbb{E}\left[f_C\left(X\right) \mid f_C\left(X\right) \in B_i\right]}_{=\text{conf}_i} - \underbrace{\mathbb{P}\left(Y = \hat{Y} \mid f_C\left(X\right) \in B_i\right)}_{=\text{acc}_i} \tag{22}$$

as the true unknown difference between model confidence and model accuracy in bin i,

$$\sigma_i^2 = \frac{1}{n_i} \underbrace{\mathbb{V}\left[f_C\left(X\right) \mid f_C\left(X\right) \in B_i\right]}_{V_i^{\text{conf}} :=} + \frac{1}{n_i} \underbrace{\text{acc}_i \left(1 - \text{acc}_i\right)}_{V_i^{\text{acc}} :=} \tag{23}$$

as the combined model and accuracy sample variance in bin i, and Φ as the cumulative distribution function (cdf) of a standard normal distribution.

The ECE for data size n and m bins is estimated by

$$\hat{\text{ECE}}_{(n)} = \sum_{i=1}^{m} \hat{p}_b \left|\hat{\text{acc}}_b - \hat{\text{conf}}_b\right| \tag{24}$$

where $\hat{p}_i = \frac{n_i}{n}$ is the estimated bin frequency, $\hat{\text{acc}}_i = \frac{1}{n_i} \sum_j \mathbb{1}\left\{\hat{Y}_j = Y_j \wedge \hat{Y}_j \in B_i\right\}$ the estimated bin accuracy, $\hat{\text{conf}}_i = \frac{1}{n_i} \sum_j \hat{Y}_j \mathbb{1}\left\{\hat{Y}_j \in B_i\right\}$ the estimated bin confidence, and $n_i = \sum_j \mathbb{1}\left\{\hat{Y}_j \in B_i\right\}$ is the number of data instances in bin i. We assume equal width binning, i.e. $B_i = \left(\frac{i-m}{m}, \frac{i}{m}\right]$.

We have

$$\mathbb{E}\left[\hat{\text{ECE}}_{(n)}\right] \approx \mu_{(n)} \geq \text{ECE} \quad , \quad \frac{\mathrm{d}\mu_{(n)}}{\mathrm{d}n} < 0 \quad , \quad \frac{\mathrm{d}^2\mu_{(n)}}{\left(\mathrm{d}n\right)^2} > 0 \quad \text{and} \quad \frac{\mathrm{d}^2\mu_{(n)}}{\mathrm{d}n \; \mathrm{dECE}} > 0.$$

Proof. First,

$$\begin{aligned}
&\mathbb{E}\left[\hat{\text{ECE}}_{(n)}\right] \\
&\overset{\text{def}}{=} \mathbb{E}\left[\sum_{i=1}^{m} \hat{p}_i \left|\hat{\text{acc}}_i - \hat{\text{conf}}_i\right|\right] \\
&\overset{\text{def}}{=} \mathbb{E}\left[\sum_{i=1}^{m} \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\left\{\hat{Y}_j \in B_i\right\} \left|\hat{\text{acc}}_i - \hat{\text{conf}}_i\right|\right] \\
&= \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}\left\{\hat{Y}_j \in B_i\right\} \left|\hat{\text{acc}}_i - \hat{\text{conf}}_i\right|\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\sum_{i=1}^{m} \mathbb{1}\left\{\hat{Y}_j \in B_i\right\} \mathbb{E}\left[\left|\hat{\text{acc}}_i - \hat{\text{conf}}_i\right| \mid \hat{Y}_j\right]\right] \\
&\overset{\text{(i)}}{\approx} \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{m} \mathbb{E}\left[\mathbb{1}\left\{\hat{Y}_j \in B_i\right\}\right] \mathbb{E}\left[\left|\hat{\text{acc}}_i - \hat{\text{conf}}_i\right|\right] \\
&\overset{\text{iid}}{=} \sum_{i=1}^{m} \mathbb{P}\left(\hat{Y} \in B_i\right) \mathbb{E}\left[\left|\hat{\text{acc}}_i - \hat{\text{conf}}_i\right|\right]
\end{aligned} \tag{25}$$

(i) 'knowing' a single summand in a mean estimator does not change much.

As one can see, the ECE estimator approximately consists of several $\mathbb{E}\left[\left|\hat{\text{acc}}_i - \hat{\text{conf}}_i\right|\right]$. According to the central limit theorem (CLT), we have $\lim_{n_i \to \infty} \left(\frac{\hat{\text{acc}}_i - \text{acc}_i}{\sqrt{V_i^{\text{acc}}/n_i}}\right) \sim \mathscr{N}\left(0, 1\right)$ and

$\lim_{n_i \to \infty} \left(\frac{\hat{\text{conf}}_i - \text{conf}_i}{\sqrt{V_i^{\text{conf}}/n_i}} \right) \sim \mathscr{N}(0,1)$. We assume $\hat{\text{acc}}_i$ and $\hat{\text{conf}}_i$ approximately follow the normal distributions given by the CLT, i.e. $\hat{\text{acc}}_i \sim \mathscr{N}\left(\text{acc}_i, \frac{V_i^{\text{acc}}}{n_i}\right)$ and $\hat{\text{conf}}_i \sim \mathscr{N}\left(\text{conf}_i, \frac{V_i^{\text{conf}}}{n_i}\right)$. This gives $\hat{\text{acc}}_i - \hat{\text{conf}}_i \sim \mathscr{N}\left(\text{acc}_i - \text{conf}_i, \frac{V_i^{\text{conf}} + V_i^{\text{acc}}}{n_i}\right)$. [3] If $X \sim \mathscr{N}(\mu, \sigma^2)$, then $|X|$ is a folded normal distribution (FN) with $\mathbb{E}[|X|] = \sqrt{\frac{2}{\pi}}\sigma \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left[1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right]$ with Φ the cdf of a standard normal distribution [51]. We also have

$$\sqrt{\frac{2}{\pi}}\sigma \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left[1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right] = \mathbb{E}[|X|] \geq |\mathbb{E}[X]| = |\mu| \tag{26}$$

(by Jensen's inequality) and

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\sigma}\mathbb{E}[|X|] &= \frac{\mathrm{d}}{\mathrm{d}\sigma}\left(\sqrt{\frac{2}{\pi}}\sigma \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left[1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right]\right) \\
&= \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \sqrt{\frac{2}{\pi}}\sigma \exp\left(-\frac{\mu^2}{2\sigma^2}\right)\frac{\mu^2}{\sigma^3} - \mu 2\phi\left(-\frac{\mu}{\sigma}\right)\frac{\mu}{\sigma^2} \\
&= \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)\frac{\mu^2}{\sigma^2} - \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)\frac{\mu^2}{\sigma^2} \\
&= \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right).
\end{aligned} \tag{27}$$

Consequently, $\left|\hat{\text{acc}}_i - \hat{\text{conf}}_i\right|$ follows approximately a folded normal distribution with

$$\mathbb{E}\left[\left|\hat{\text{acc}}_i - \hat{\text{conf}}_i\right|\right] \approx \sqrt{\frac{2}{\pi}}\sigma_i \exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right) + \mu_i\left[1 - 2\Phi\left(-\frac{\mu_i}{\sigma_i}\right)\right] \tag{28}$$

where μ_i and σ_i are defined as above in equations (22) and (23).
Consequently, by combining equations (26), (28) and (25) we get the first result:

$$\begin{aligned}
\mathbb{E}\left[\hat{\text{ECE}}_{(n)}\right] &\approx \underbrace{\sum_{i=1}^{m} p_i \left\{\sqrt{\frac{2}{\pi}}\sigma_i \exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right) + \mu_i\left[1 - 2\Phi\left(-\frac{\mu_i}{\sigma_i}\right)\right]\right\}}_{=\mu_{(n)}} \\
&\geq \sum_{i=1}^{m} p_i |\mu_i| \\
&= \sum_{i=1}^{m} p_i |\text{acc}_i - \text{conf}_i| \\
&= \text{ECE}
\end{aligned} \tag{29}$$

As we can see, the average outcome depends on σ_i, which further depends on n_i, i.e. the data size influences our expected result. To get the next result, which shows the trend of this influence, we calculate the first derivative:

[3] http://www.stat.ucla.edu/~nchristo/introstatistics/introstats_normal_linear_combinations.pdf

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}n}\mu_{(n)} \\
&= \frac{\mathrm{d}}{\mathrm{d}n}\sum_{i=1}^{m} p_i \left\{ \sqrt{\frac{2}{\pi}}\sigma_i \exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right) + \mu_i \left[1 - 2\Phi\left(-\frac{\mu_i}{\sigma_i}\right)\right]\right\} \\
&= \sum_{j=1}^{m} \frac{\mathrm{d}n_j}{\mathrm{d}n}\frac{\mathrm{d}\sigma_j}{\mathrm{d}n_j}\frac{\mathrm{d}}{\mathrm{d}\sigma_j}\sum_{i=1}^{m} p_i \left\{ \sqrt{\frac{2}{\pi}}\sigma_i \exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right) + \mu_i \left[1 - 2\Phi\left(-\frac{\mu_i}{\sigma_i}\right)\right]\right\} \\
&= \sum_{j=1}^{m} \frac{\mathrm{d}n_j}{\mathrm{d}n}\frac{\mathrm{d}\sigma_j}{\mathrm{d}n_j}\frac{\mathrm{d}}{\mathrm{d}\sigma_j} p_j \left\{ \sqrt{\frac{2}{\pi}}\sigma_j \exp\left(-\frac{\mu_j^2}{2\sigma_j^2}\right) + \mu_j \left[1 - 2\Phi\left(-\frac{\mu_j}{\sigma_j}\right)\right]\right\} \\
&\overset{(27)}{=} \sum_{j=1}^{m} \frac{\mathrm{d}n_j}{\mathrm{d}n}\frac{\mathrm{d}\sigma_j}{\mathrm{d}n_j} p_j \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_j^2}{2\sigma_j^2}\right) \\
&= \sum_{j=1}^{m} \frac{\mathrm{d}n_j}{\mathrm{d}n}\frac{-\sqrt{V_j^{\text{conf}} + V_j^{\text{acc}}}}{2n_j^{3/2}} p_j \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_j^2}{2\sigma_j^2}\right) \\
&= \sum_{j=1}^{m} 1 \cdot \underbrace{\frac{-\sqrt{V_j^{\text{conf}} + V_j^{\text{acc}}}}{2n_j^{3/2}}}_{<0} \underbrace{p_j \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_j^2}{2\sigma_j^2}\right)}_{>0}.
\end{aligned}
\tag{30}
$$

This means $\mu_{(n)}$ is monotonically decreasing with growing data size.

Next, we have

$$
\begin{aligned}
&\frac{\mathrm{d}^2}{(\mathrm{d}n)^2}\mu_{(n)} \\
&\overset{(30)}{=} \frac{\mathrm{d}}{\mathrm{d}n}\sum_{j=1}^{m} \frac{-\sqrt{V_j^{\text{conf}} + V_j^{\text{acc}}}}{2n_j^{3/2}} p_j \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_j^2}{2\sigma_j^2}\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}n}\sum_{j=1}^{m} \frac{-\sqrt{V_j^{\text{conf}} + V_j^{\text{acc}}}}{2n_j^{3/2}} p_j \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_j^2 n_j}{2V_j^{\text{conf}} + 2V_j^{\text{acc}}}\right) \\
&= \sum_{i=1}^{m} \frac{\mathrm{d}n_i}{\mathrm{d}n}\frac{\mathrm{d}}{\mathrm{d}n_i}\sum_{j=1}^{m} \frac{-\sqrt{V_j^{\text{conf}} + V_j^{\text{acc}}}}{2n_j^{3/2}} p_j \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_j^2 n_j}{2V_j^{\text{conf}} + 2V_j^{\text{acc}}}\right) \\
&= \sum_{i=1}^{m} 1 \cdot \frac{\mathrm{d}}{\mathrm{d}n_i}\frac{-\sqrt{V_i^{\text{conf}} + V_i^{\text{acc}}}}{2n_i^{3/2}} p_i \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_i^2 n_i}{2V_i^{\text{conf}} + 2V_i^{\text{acc}}}\right) \\
&= \sum_{i=1}^{m} \underbrace{\frac{3\sqrt{V_i^{\text{conf}} + V_i^{\text{acc}}}}{4n_i^{5/2}} p_i \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_i^2 n_i}{2V_i^{\text{conf}} + 2V_i^{\text{acc}}}\right)}_{>0} \\
&\quad + \underbrace{\frac{\sqrt{V_i^{\text{conf}} + V_i^{\text{acc}}}}{2n_i^{3/2}} p_i \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\mu_i^2 n_i}{2V_i^{\text{conf}} + 2V_i^{\text{acc}}}\right) \frac{\mu_i^2}{2V_i^{\text{conf}} + 2V_i^{\text{acc}}}}_{>0}.
\end{aligned}
\tag{31}
$$

This means, in combination with the previous result, $\mu_{(n)}$ is a strictly convex and monotonically decreasing function of the data size n_b. The ECE estimate is increasingly sensitive to the data size for smaller data sizes, while for larger data sizes the sensitivity vanishes.

Next, we analyze how the goodness of calibration influences this behaviour. Recall that $\mu_i = \text{acc}_i - \text{conf}_i$, i.e. μ_i^2 is the ground truth squared calibration error of bin i. We have

$$
\begin{aligned}
&\frac{\mathrm{d}^2}{\mathrm{d}n\mathrm{d}\mu_i^2}\mu_{(n)} \\
&\overset{(30)}{=} \frac{\mathrm{d}}{\mathrm{d}\mu_i^2}\sum_{j=1}^{m}\frac{-\sqrt{V_j^{\text{conf}}+V_j^{\text{acc}}}}{2n_j^{3/2}}p_j\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu_j^2}{2\sigma_j^2}\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}\mu_i^2}\frac{-\sqrt{V_i^{\text{conf}}+V_i^{\text{acc}}}}{2n_i^{3/2}}p_i\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right) \\
&= \underbrace{\frac{\sqrt{V_i^{\text{conf}}+V_i^{\text{acc}}}}{4n_i^{3/2}\sigma_i^2}p_i\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu_i^2}{2\sigma_i^2}\right)}_{>0}
\end{aligned}
\tag{32}
$$

Consequently, if μ_i^2 increases (i.e. calibration gets worse), the gradient $\frac{\mathrm{d}}{\mathrm{d}n}\mathbb{E}\left[\hat{\text{ECE}}_{(n)}\right]$ monotonically approaches zero from beneath. Contrary, the gradient is the highest when $\mu_i = 0$. In other words, the sensitivity of the ECE estimate w.r.t. the data size monotonically depends on the goodness of calibration. With better calibration, the sensitivity gradually gets worse.

Further, we have

$$
\begin{aligned}
&\frac{\mathrm{d}\mu_i^2}{\mathrm{dECE}} \\
&= \frac{\mathrm{d}}{\mathrm{dECE}}\left(\text{acc}_i - \text{conf}_i\right)^2 \\
&= \frac{\mathrm{d}}{\mathrm{dECE}}\left|\text{acc}_i - \text{conf}_i\right|^2 \\
&= \frac{\mathrm{d}}{\mathrm{dECE}}\left(\frac{\text{ECE} - \sum_{j\neq i}p_j\left|\text{acc}_j - \text{conf}_j\right|}{p_i}\right)^2 \\
&= \underbrace{2\frac{\text{ECE} - \sum_{j\neq i}p_j\left|\text{acc}_j - \text{conf}_j\right|}{p_i^2}}_{>0}.
\end{aligned}
\tag{33}
$$

Combining equations (32) and (33) gives $\frac{\mathrm{d}^2}{\mathrm{d}n\mathrm{dECE}}\mu_{(n)} > 0$ as stated in the proposition.

□

D.6 Lemma 4.1

Let $\mathscr{P}$ be a set of arbitrary distributions for which exists a proper score S. Assume we have random variables Q and Y with $Q, \mathbb{P}_Y, \mathbb{P}_{Y|Q} \in \mathscr{P}$ for which $g_S\left(\mathbb{P}_Y\right), \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|Q}\right)\right], \mathbb{E}\left[\left|S\left(Q,Y\right)\right|\right], \mathbb{E}\left[\left|S\left(\mathbb{P}_Y,Y\right)\right|\right] < \infty$. The last two expectations are required for Fubini's theorem.

$$
\begin{aligned}
\mathbb{E}\left[S\left(Q,Y\right)\right] &= \int S\left(q,y\right)\mathrm{d}\mathbb{P}_{Y,Q}\left(y,q\right) \\
&\overset{\text{Fubini}}{=} \int\int S\left(q,y\right)\mathrm{d}\mathbb{P}_{Y|Q=q}\left(y\right)\mathrm{d}\mathbb{P}_{Q}\left(q\right) \\
&\overset{\text{def } C.5}{=} \int s_S\left(q,\mathbb{P}_{Y|Q=q}\right)\mathrm{d}\mathbb{P}_{Q}\left(q\right) \\
&= \mathbb{E}\left[s_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] \\
&= \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[s_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] - \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] \\
&\overset{\text{def } C.6}{=} \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[d_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] \\
&= s_S\left(\mathbb{P}_Y,\mathbb{P}_Y\right) - s_S\left(\mathbb{P}_Y,\mathbb{P}_Y\right) + \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[d_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] \\
&\overset{\text{def } C.5}{=} s_S\left(\mathbb{P}_Y,\mathbb{P}_Y\right) - \int S\left(\mathbb{P}_Y,y\right)\mathrm{d}\mathbb{P}_Y\left(y\right) + \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[d_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] \\
&= s_S\left(\mathbb{P}_Y,\mathbb{P}_Y\right) - \int S\left(\mathbb{P}_Y,y\right)\underbrace{\int \mathrm{d}\mathbb{P}_{Q|Y=y}\left(q\right)}_{=1}\mathrm{d}\mathbb{P}_Y\left(y\right) + \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[d_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] \\
&= s_S\left(\mathbb{P}_Y,\mathbb{P}_Y\right) - \int S\left(\mathbb{P}_Y,y\right)\mathrm{d}\mathbb{P}_{Y,Q}\left(y,q\right) + \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[d_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] \\
&\overset{\text{Fubini}}{=} s_S\left(\mathbb{P}_Y,\mathbb{P}_Y\right) - \int\int S\left(\mathbb{P}_Y,y\right)\mathrm{d}\mathbb{P}_{Y|Q=q}\left(y\right)\mathrm{d}\mathbb{P}_{Q}\left(q\right) + \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[d_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] \\
&\overset{\text{def } C.5}{=} s_S\left(\mathbb{P}_Y,\mathbb{P}_Y\right) - \int s_S\left(\mathbb{P}_Y,\mathbb{P}_{Y|Q=q}\right)\mathrm{d}\mathbb{P}_{Q}\left(q\right) + \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[d_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] \\
&= s_S\left(\mathbb{P}_Y,\mathbb{P}_Y\right) - \mathbb{E}\left[s_S\left(\mathbb{P}_Y,\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[s_S\left(\mathbb{P}_{Y|Q},\mathbb{P}_{Y|Q}\right)\right] + \mathbb{E}\left[d_S\left(Q,\mathbb{P}_{Y|Q}\right)\right] \\
&\overset{\text{def } C.6}{=} \underbrace{g_S\left(\mathbb{P}_Y\right)}_{\text{generalized entropy}} - \underbrace{\mathbb{E}\left[d_S\left(\mathbb{P}_Y,\mathbb{P}_{Y|Q}\right)\right]}_{\text{sharpness}} + \underbrace{\mathbb{E}\left[d_S\left(Q,\mathbb{P}_{Y|Q}\right)\right]}_{\text{calibration}}.
\end{aligned}
$$

D.7 Theorem 4.3

For all proper calibration errors with $\inf_{P\in\mathscr{P}} g_S\left(P\right) \in \mathbb{R}$, there exists an associated **calibration upper bound**

$$\mathscr{U}_S\left(f\right) \geq \mathrm{CE}_S\left(f\right)$$

defined as $\mathscr{U}_S\left(f\right) \coloneqq \mathbb{E}\left[S\left(f\left(X\right),Y\right)\right] - \inf_{P\in\mathscr{P}} g_S\left(P\right)$. Under a classification setting and further mild conditions, it is asymptotically equal to the CE_S with increasing model accuracy, i.e.

$$\lim_{\mathrm{ACC}(f)\to 1} \mathscr{U}_S\left(f\right) - \mathrm{CE}_S\left(f\right) = 0.$$

Proof. **Regarding existence of upper bound**

Assuming $\inf_{Q\in\mathscr{P}} g_S\left(Q\right) \in \mathbb{R}$.

$$
\begin{aligned}
\mathrm{CE}_S\left(f\right) &\overset{\text{le } D.1}{=} \mathbb{E}\left[S\left(f\left(X\right),Y\right)\right] - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right] \\
&\leq \mathbb{E}\left[S\left(f\left(X\right),Y\right)\right] - \mathbb{E}\left[\inf_{Q\in\mathscr{P}} g_S\left(Q\right)\right] \\
&= \mathbb{E}\left[S\left(f\left(X\right),Y\right)\right] - \inf_{Q\in\mathscr{P}} g_S\left(Q\right) \\
&\overset{\text{th } 4.3}{=} \mathscr{U}_S\left(f\right)
\end{aligned}
\tag{34}
$$

Regarding accuracy limes

Assuming mild conditions $g_S \colon \mathscr{P}_n \to \mathbb{R}$ is continuous and $g_S(e_1) = g_S(e_2) = \cdots = g_S(e_n)$. See Figure 2 in Gneiting & Raftery [13] for an example when this is violated. S does not have to be symmetric for this to hold.

$$
\begin{aligned}
&\lim_{\mathrm{ACC}(f)\to 1} \mathrm{CE}_S(f) - \mathscr{U}_S(f) \\
&\overset{\text{th } 4.3}{=} \lim_{\mathrm{ACC}(f)\to 1} \mathrm{CE}_S(f) - \mathbb{E}\left[S(f(X), Y)\right] + \inf_{Q\in\mathscr{P}_n} g_S(Q) \\
&\overset{\text{le } D.1}{=} \lim_{\mathrm{ACC}(f)\to 1} \mathbb{E}\left[S(f(X), Y)\right] - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right] - \mathbb{E}\left[S(f(X), Y)\right] + \inf_{Q\in\mathscr{P}_n} g_S(Q) \\
&= \lim_{\mathrm{ACC}(f)\to 1} \inf_{Q\in\mathscr{P}_n} g_S(Q) - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right] \\
&= \inf_{Q\in\mathscr{P}_n} g_S(Q) - \lim_{\mathrm{ACC}(f)\to 1} \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right] \\
&= \inf_{Q\in\mathscr{P}_n} g_S(Q) - \mathbb{E}\left[g_S\left(\lim_{\mathrm{ACC}(f)\to 1} \mathbb{P}_{Y|f(X)}\right)\right] \\
&\overset{\text{(i)}}{=} \inf_{Q\in\mathscr{P}_n} g_S(Q) - \mathbb{E}\left[g_S\left(e_{i(X)}\right)\right] \\
&\overset{\text{(ii)}}{=} \inf_{Q\in\mathscr{P}_n} g_S(Q) - \mathbb{E}\left[g_S(e_1)\right] \\
&= \inf_{Q\in\mathscr{P}_n} g_S(Q) - g_S(e_1) \\
&\overset{\text{(iii)}}{=} \inf_{Q\in\mathscr{P}_n} g_S(Q) - \inf_{Q\in\mathscr{P}_n} g_S(Q) \\
&= 0
\end{aligned}
\tag{35}
$$

(i) Perfect accuracy results in deterministic predictions, i.e. $\forall z \in \mathscr{P}_n \colon \lim_{\mathrm{ACC}(f)\to 1} \mathbb{P}_{Y|f(X)=z} \in \{e_i \mid n \geq i \in \mathbb{N}\}$. If we define $i \colon \mathscr{X} \to \mathbb{N}_{\leq n}$ as $i(X) \coloneqq \arg\max_k \lim_{\mathrm{ACC}(f)\to 1} \mathbb{P}(Y = k \mid f(X))$, then we have $e_{i(X)} = \lim_{\mathrm{ACC}(f)\to 1} \mathbb{P}_{Y|f(X)}$.
(ii) Follows from initial condition.
(iii) Since g_S is concave and by the definition of $\mathscr{P}_n$, we have

$$
\forall z \in \mathscr{P}_n \exists \lambda_1, \ldots, \lambda_n \geq 0, \sum_k \lambda_k = 1 \colon\ g_S(z) = g_S\left(\sum_k \lambda_k e_k\right) \geq \sum_k \lambda_k g_S(e_k) = \sum_k \lambda_k g_S(e_1) = g_S(e_1)\,. \tag{36}
$$

From this follows that $g_S(e_1) = \inf_{Q\in\mathscr{P}_n} g_S(Q)$.

□

D.8 Proposition 4.5

Given injective functions $h, h' : \mathscr{P} \to \mathscr{P}$ we have

$$
\mathscr{U}_S(h \circ f) - \mathscr{U}_S(f) = \mathrm{CE}_S(h \circ f) - \mathrm{CE}_S(f) \quad ,
$$

$$
\mathscr{U}_S(h \circ f) > \mathscr{U}_S(h' \circ f) \iff \mathrm{CE}_S(h \circ f) > \mathrm{CE}_S(h' \circ f)
$$

and (assuming S is differentiable)

$$
\frac{\mathrm{d}\mathscr{U}_S(h \circ f)}{\mathrm{d}h} = \frac{\mathrm{d}\mathrm{CE}_S(h \circ f)}{\mathrm{d}h}.
$$

Proof.

$$
\begin{aligned}
&\mathscr{U}_S\left(h \circ f\right) - \mathscr{U}_S\left(h' \circ f\right) \\
&\overset{\text{th } 4.3}{=} \mathbb{E}\left[S\left(h \circ f\left(X\right), Y\right)\right] - \inf_{Q \in \mathscr{P}_n} g_S\left(Q\right) - \mathbb{E}\left[S\left(h' \circ f\left(X\right), Y\right)\right] + \inf_{Q \in \mathscr{P}_n} g_S\left(Q\right) \\
&= \mathbb{E}\left[S\left(h \circ f\left(X\right), Y\right)\right] - \mathbb{E}\left[S\left(h' \circ f\left(X\right), Y\right)\right] \\
&= \mathbb{E}\left[S\left(h \circ f\left(X\right), Y\right)\right] - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right] - \mathbb{E}\left[S\left(h' \circ f\left(X\right), Y\right)\right] + \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}\left[S\left(h \circ f\left(X\right), Y\right)\right] - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|h \circ f(X)}\right)\right] - \mathbb{E}\left[S\left(h' \circ f\left(X\right), Y\right)\right] + \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|h' \circ f(X)}\right)\right] \\
&\overset{\text{le } D.1}{=} \mathrm{CE}_S\left(h \circ f\right) - \mathrm{CE}_S\left(h' \circ f\right)
\end{aligned}
\tag{37}
$$

from which follows $\mathscr{U}_S\left(h \circ f\right) - \mathscr{U}_S\left(f\right) = \mathrm{CE}_S\left(h \circ f\right) - \mathrm{CE}_S\left(f\right)$ and $\mathscr{U}_S\left(h \circ f\right) > \mathscr{U}_S\left(h' \circ f\right) \iff \mathrm{CE}_S\left(h \circ f\right) > \mathrm{CE}_S\left(h' \circ f\right)$. Further we have for differentiable S

$$
\begin{aligned}
&\frac{\mathrm{d}\mathscr{U}_S\left(h \circ f\right)}{\mathrm{d}h} \\
&\overset{\text{th } 4.3}{=} \frac{\mathrm{d}\mathbb{E}\left[S\left(h \circ f\left(X\right), Y\right)\right] - \inf_{Q \in \mathscr{P}_n} g_S\left(Q\right)}{\mathrm{d}h} \\
&= \frac{\mathrm{d}\mathbb{E}\left[S\left(h \circ f\left(X\right), Y\right)\right]}{\mathrm{d}h} \\
&= \frac{\mathrm{d}\mathbb{E}\left[S\left(h \circ f\left(X\right), Y\right)\right] - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]}{\mathrm{d}h} \\
&\overset{\text{(i)}}{=} \frac{\mathrm{d}\mathbb{E}\left[S\left(h \circ f\left(X\right), Y\right)\right] - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|h \circ f(X)}\right)\right]}{\mathrm{d}h} \\
&\overset{\text{le } D.1}{=} \frac{\mathrm{d}\mathrm{CE}_S\left(h \circ f\right)}{\mathrm{d}h}
\end{aligned}
\tag{38}
$$

(i) Since h is injective, we have $\forall z \in \mathscr{P}_n\colon \ \{x \in \mathscr{X} \mid f\left(x\right) = z\} = \{x \in \mathscr{X} \mid h \circ f\left(x\right) = h\left(z\right)\}$ and $\{(x, y) \in \mathscr{X} \times \mathscr{Y} \mid f\left(x\right) = z\} = \{(x, y) \in \mathscr{X} \times \mathscr{Y} \mid h \circ f\left(x\right) = h\left(z\right)\}$. Consequently $\mathbb{P}\left(Y \mid f\left(X\right) = z\right) = \frac{\mathbb{P}(Y, f(X) = z)}{\mathbb{P}(f(X) = z)} = \frac{\mathbb{P}(Y, h \circ f(X) = h(z))}{\mathbb{P}(h \circ f(X) = h(z))} = \mathbb{P}\left(Y \mid h \circ f\left(X\right) = h\left(z\right)\right)$. □

E Recalibration transformations

E.1 calibrated and accuracy-preserving

The binary case is directly given in the multi-class case, but if we only have a scalar output, which is common for binary classification, deriving the transformation is not that trivial. Consequently, we handle this case separately.

We will also make use of the following lemma.

Lemma E.1. *For random variables Y and X, we have*

$$\mathbb{P}\left(Y \mid \mathbb{P}\left(Y \mid X\right)\right) = \mathbb{P}\left(Y \mid X\right).$$

Proof. Proof directly follows from Proposition 1 in Vaicenavicius et al. [52] with $h \equiv \mathrm{id}$. □

E.1.1 Binary case (scalar output)

Assume we are given $f\colon \mathscr{X} \to [0, 1]$.

Define $t^f\colon \ [0, 1] \to [0, 1]$ as

$$
t^f\left(p\right) = \begin{cases} \mathbb{P}\left(Y = 1 \mid f\left(X\right) < 0.5\right) & \text{, if } p < 0.5 \\ \mathbb{P}\left(Y = 1 \mid f\left(X\right) \geq 0.5\right) & \text{, else} \end{cases}
\tag{39}
$$

The first line has as unbiased estimator the precision (or positive predictive value), the second the false omission rate.

This gives

$$\begin{aligned}\mathbb{P}\left(Y=1 \mid t^f \circ f(X)\right) &= \begin{cases}\mathbb{P}\left(Y=1 \mid \mathbb{P}\left(Y=1 \mid f(X)<0.5\right)\right) & \text{, if } f(X)<0.5 \\ \mathbb{P}\left(Y=1 \mid \mathbb{P}\left(Y=1 \mid f(X)\geq 0.5\right)\right) & \text{, else}\end{cases} \\ &= \begin{cases}\mathbb{P}\left(Y=1 \mid f(X)<0.5\right) & \text{, if } f(X)<0.5 \\ \mathbb{P}\left(Y=1 \mid f(X)\geq 0.5\right) & \text{, else}\end{cases} \\ &= t^f \circ f(X)\end{aligned} \tag{40}$$

i.e. $t^f \circ f$ is calibrated. Further, if $\mathbb{P}\left(Y=1 \mid f(X)<0.5\right) < \mathbb{P}\left(Y=1 \mid f(X)\geq 0.5\right)$, then $t^f \circ f$ has the same accuracy as f. This can be assumed as given for any meaningful classifier. The reduction in sharpness directly follows from the analog proof in the multi-class case.

E.1.2 Multi-class case (vector output)

Let $r\colon \mathscr{P}_n \to A$ with $A = \left\{a \in \{0,1\}^K \mid \sum_k a_k = 1\right\}$ be defined as $r(p) \coloneqq e_{\arg\max_k p_k}$. In words, r returns a vector of only zeros except a '1' at index $\arg\max_k p_k$ for input $p \in \mathscr{P}_n$.

Define $t^f\colon \mathscr{P}_n \to \mathscr{P}_n$ as

$$t^f(p) = \mathbb{P}\left(Y \mid r \circ f(X) = r(p)\right) \tag{41}$$

(For easier notation, we say $\mathbb{P}(Y) \in \mathscr{P}_n$)

Given a dataset $\{(X_1,Y_1),\ldots,(X_m,Y_m)\}$, an unbiased estimator of $\mathbb{P}\left(Y \mid r \circ f(X) = a\right) \forall a \in A$ is $P_a = \frac{1}{|I_a|}\sum_{i\in I_a} e_{Y_i}$ with $I_a = \{i \in \{1,\ldots,m\} \mid r \circ f(X_i) = a\}$. And since $|A| = n$, estimation is also feasible for higher number of classes.

We also have

$$\begin{aligned}\mathbb{P}\left(Y \mid t^f \circ f(X)\right) &= \mathbb{P}\left(Y \mid \mathbb{P}\left(Y \mid r \circ f(X) = r \circ f(X)\right)\right) \\ &= \mathbb{P}\left(Y \mid \mathbb{P}\left(Y \mid r \circ f(X)\right)\right) \\ &= \mathbb{P}\left(Y \mid r \circ f(X)\right) \\ &= t^f \circ f(X)\end{aligned} \tag{42}$$

Consequently, $t^f \circ f$ is calibrated.

If $\arg\max_k f_k(X) = \arg\max_k \mathbb{P}\left(Y=k \mid \arg\max_k f_k(X)\right)$, then $\arg\max_k f_k(X) = \arg\max_k \mathbb{P}\left(Y=k \mid r \circ f(X)\right) = \arg\max_k \mathbb{P}\left(Y=k \mid r \circ f(X) = r \circ f(X)\right) = \arg\max_k t^f_k \circ f(X)$, i.e. t^f is accuracy preserving. Recall that $\arg\max_k f_k(X)$ is the predicted top-label, making $\arg\max_k \mathbb{P}\left(Y=k \mid \arg\max_k f_k(X)\right)$ the most likely outcome given a predicted top-label. So, we can restate the above as: t^f is accuracy preserving if for every predicted top-label the most likely outcome is that label. This should hold in every meaningful practical setting, or else t^f might as well improve the accuracy.

$t^f \circ f$ has lower sharpness as f w.r.t. a proper score S. This is a special case of the following proposition, where we write $\text{SHARP}_S(f)$ as the sharpness of model f given by the sharpness term in Lemma 4.1 of a proper score S.

Proposition E.2. *Assume Lemma 4.1 holds given a proper score S. For a function $m\colon \mathscr{P}_n \to \mathscr{P}_n$ and model $f\colon \mathscr{X} \to \mathscr{P}_n$, we have*

$$\textit{SHARP}_S(f) \geq \textit{SHARP}_S(m \circ f).$$

Proof. Since we assumed Lemma 4.1 holds, the conditions for Fubini's theorem are met. We will use:

$$
\begin{aligned}
&\mathrm{SHARP}_S\left(f\right)\\
&\overset{\text{le 4.1}}{=} \mathbb{E}\left[d_S\left(\mathbb{P}_Y, \mathbb{P}_{Y|f(X)}\right)\right]\\
&\overset{\text{def } C.6}{=} \mathbb{E}\left[s_S\left(\mathbb{P}_Y, \mathbb{P}_{Y|f(X)}\right)\right] - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\
&\overset{\text{def } C.5}{=} \int\int S\left(\mathbb{P}_Y, y\right) \mathrm{d}\mathbb{P}_{Y|f(X)=z}\left(y\right) \mathrm{d}\mathbb{P}_{f(X)}\left(z\right) - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\
&= \int S\left(\mathbb{P}_Y, y\right) \mathrm{d}\mathbb{P}_{Y,f(X)}\left(y, z\right) - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\
&\overset{\text{Fubini}}{=} \int S\left(\mathbb{P}_Y, y\right) \int \mathrm{d}\mathbb{P}_{f(X)|Y=y}\left(z\right) \mathrm{d}\mathbb{P}_Y\left(y\right) - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\
&= \int S\left(\mathbb{P}_Y, y\right) \mathrm{d}\mathbb{P}_Y\left(y\right) - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\
&\overset{\text{def } C.6}{=} g_S\left(\mathbb{P}_Y\right) - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]
\end{aligned}
\tag{43}
$$

Now, we can show

$$
\begin{aligned}
&\mathrm{SHARP}_S\left(f\right)\\
&\overset{\text{eq (43)}}{=} g_S\left(\mathbb{P}_Y\right) - \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right)\right]\\
&= g_S\left(\mathbb{P}_Y\right) - \mathbb{E}_{m\circ f(X)}\left[\mathbb{E}_{f(X)}\left[g_S\left(\mathbb{P}_{Y|f(X)}\right) \mid m\circ f\left(X\right)\right]\right]\\
&\overset{\text{Jensen}}{\geq} g_S\left(\mathbb{P}_Y\right) - \mathbb{E}_{m\circ f(X)}\left[g_S\left(\mathbb{E}_{f(X)}\left[\mathbb{P}_{Y|f(X)} \mid m\circ f\left(X\right)\right]\right)\right]\\
&= g_S\left(\mathbb{P}_Y\right) - \mathbb{E}_{m\circ f(X)}\left[g_S\left(\mathbb{E}_{f(X)}\left[\mathbb{E}\left[e_Y \mid f\left(X\right)\right] \mid m\circ f\left(X\right)\right]\right)\right]\\
&= g_S\left(\mathbb{P}_Y\right) - \mathbb{E}_{m\circ f(X)}\left[g_S\left(\mathbb{E}\left[e_Y \mid m\circ f\left(X\right)\right]\right)\right]\\
&= g_S\left(\mathbb{P}_Y\right) - \mathbb{E}_{m\circ f(X)}\left[g_S\left(\mathbb{P}_{Y|m\circ f(X)}\right)\right]\\
&\overset{\text{eq (43)}}{=} \mathrm{SHARP}_S\left(m\circ f\right)
\end{aligned}
\tag{44}
$$

□

If the underlying score is the log score, then the sharpness is the mutual information between predictions and target random variable. Consequently, we can interpret the sharpness as generalized mutual information. This gives the proposition the following intuitive meaning: There exists no function, that can transform a random variable in a way such that the mutual information with another random variable is increased. Or, in other words, we cannot add 'information' to a random variable by transforming it in a deterministic way.

F Proper U-scores

In this section we introduce a generalization of proper scores. Based on U-statistics, we define proper U-scores. This allows us to naturally extend the definition of proper calibration errors to be based on proper U-scores instead of just proper scores. Consequently, we can cover more calibration errors with desired properties. For example, we can show that the squared KCE [55] is a proper calibration error based on a U-score (but not on a conventional score). The squared KCE has an unbiased estimator, thus, this extension of the definition of proper calibration errors has substantial practical value.

F.1 Background

Let $X_1, \ldots, X_n$ be n iid random variables and $\phi\left(x_1, \ldots, x_r\right)$ a function with $r \leq n$. Let $\mathbf{P} = \left\{a \in \{1, \ldots, n\}^r \mid a_1 < \cdots < a_r\right\}$ be the set of r sized ordered permutations out of n, i.e. $|\mathbf{P}| = \binom{n}{r}$.

Then $U = \frac{1}{|\mathbf{P}|} \sum_{a \in \mathbf{P}} \phi\left(X_{a_1}, \ldots, X_{a_r}\right)$ is a unbiased minimum-variance estimator (UMVE) of $\mathbb{E}\left[\phi\left(X_1, \ldots, X_r\right)\right]$ and called U-statistic [21].

F.2 Contributions

Assume we have two measure spaces $(\mathscr{X}, \mathscr{F}_X)$ and $(\mathscr{Y}, \mathscr{F}_Y)$, and corresponding $\mathscr{P}_X$ and $\mathscr{P}_Y$ sets of possible probability measures. We want to score a conditional distribution $P\colon \mathscr{X} \to \mathscr{P}_Y$ given another conditional distribution $Q\colon \mathscr{X} \to \mathscr{P}_Y$.

Definition F.1. A **U-scoring rule** S is a function of the form

$$S : \mathscr{P}_Y^r \times \mathscr{Y}^r \to \overline{\mathbb{R}}$$

with $r \in \mathbb{N}$ and $\overline{\mathbb{R}} :- \mathbb{R} \cup \{-\infty, \infty\}$.

It takes r predictions and events and returns a score. For $r = 1$, U-scoring rules are scoring rules in the common definition.

Definition F.2. A **U-scoring function** s_S based on a U-scoring rule S is defined as

$$\begin{aligned} s_S &: \mathscr{P}_Y^{2r} \to \overline{\mathbb{R}} \\ (P_1, \ldots, P_r, Q_1, \ldots, Q_r) &\mapsto \int_{\mathscr{Y}^r} S\left(P_1, \ldots, P_r, y_1, \ldots, y_r\right) \mathrm{d}\left(Q_1 \times \cdots \times Q_r\right)(y) \end{aligned} \tag{45}$$

For $r = 1$, U-scoring functions are scoring functions in the common definition. If $Q_1, \ldots, Q_r$ are the distributions of $Y_1, \ldots, Y_r$ we can also write $s\left(P_1, \ldots, P_r, Q_1, \ldots, Q_r\right) = \mathbb{E}\left[S\left(P_1, \ldots, P_r, Y_1, \ldots, Y_r\right)\right]$.

Definition F.3. A U-scoring function s_S (and its U-scoring rule S) is defined to be **proper** if and only if

$$\begin{aligned} &\forall \mathbb{P} \in \mathscr{P}_X, \quad X_1, \ldots, X_r \overset{iid}{\sim} \mathbb{P}, \quad \forall P, Q\colon \mathscr{X} \to \mathscr{P}_Y : \\ &\quad \mathbb{E} s_S\left(P\left(X_1\right), \ldots, P\left(X_r\right), Q\left(X_1\right), \ldots, Q\left(X_r\right)\right) \\ &\quad \geq \mathbb{E} s_S\left(Q\left(X_1\right), \ldots, Q\left(X_r\right), Q\left(X_1\right), \ldots, Q\left(X_r\right)\right) \end{aligned} \tag{46}$$

and **strictly proper** if and only if additionally

$$\begin{aligned} &\forall \mathbb{P} \in \mathscr{P}_X, \quad X_1, \ldots, X_r \overset{iid}{\sim} \mathbb{P}, \quad \forall P, Q\colon \mathscr{X} \to \mathscr{P}_Y : \\ &Q \neq P \\ &\quad \implies \mathbb{E} s_S\left(P\left(X_1\right), \ldots, P\left(X_r\right), Q\left(X_1\right), \ldots, Q\left(X_r\right)\right) \\ &\quad\quad > \mathbb{E} s_S\left(Q\left(P_1\right), \ldots, Q\left(P_r\right), Q\left(P_1\right), \ldots, Q\left(P_r\right)\right) \end{aligned} \tag{47}$$

In words, s_S (or S) is proper if comparing Q with itself gives the best expected value, and strictly proper if no other $P \neq Q$ can achieve this value. The U-statistic of a proper s_S is a UMVE [21]. For $r = 1$, proper U-scores are identical to proper scores if $\mathscr{P}_X$ is sufficiently large. This holds since for function $f\colon \mathscr{X} \to \mathbb{R}$ and appropriate $\mathscr{P}_X$ we have: $\left(\forall \mu \in \mathscr{P}_X : \int f \mathrm{d}\mu = 0\right) \iff f = 0$.

Definition F.4. $g\left(Q_1, \ldots, Q_r\right) = s\left(Q_1, \ldots, Q_r, Q_1, \ldots, Q_r\right)$ is called the (generalized or associated) entropy.

Definition F.5. Given a proper U-score S, the associated **U-divergence** d is defined as

$$\begin{aligned} d_S &: \mathscr{P}_Y^{2r} \to \overline{\mathbb{R}}_{\geq 0} \\ (P_1, \ldots, P_r, Q_1, \ldots, Q_r) &\mapsto s_S\left(P_1, \ldots, P_r, Q_1, \ldots, Q_r\right) - g_S\left(Q_1, \ldots, Q_r\right). \end{aligned} \tag{48}$$

If S is a strictly proper U-score, $Q_1, \ldots, Q_r$ iid and $P_1, \ldots, P_r$ iid, then $\mathbb{E} d_S$ is zero if and only if $\forall i \in \{1, \ldots, r\} : \; Q_i \overset{a.s.}{=} P_i$. This follows directly by setting $P_i = P\left(X_i\right)$ and $Q_i = Q\left(X_i\right)$ in equation (47).

Assuming $P_1, \ldots, P_r$ are random variables and $\mathbb{P}_{Y|P_1}, \ldots, \mathbb{P}_{Y|P_r} \in \mathscr{P}_Y$ are the conditional distribution of independent random variables $Y_1, \ldots, Y_r \sim$

$\mathbb{P}_Y$, where each Y_i only depends on P_i. Under the condition that $g_S(\mathbb{P}_Y, \ldots, \mathbb{P}_Y), \mathbb{E}\left[g_S\left(\mathbb{P}_{Y|P_1}, \ldots, \mathbb{P}_{Y|P_r}\right)\right], \mathbb{E}\left[|S(P_1, \ldots, P_r, Y_1, \ldots, Y_r)|\right], \mathbb{E}\left[|S(\mathbb{P}_Y, \ldots, \mathbb{P}_Y, Y_1, \ldots, Y_r)|\right] < \infty$, we have the decomposition

$$\begin{aligned}
&\mathbb{E}\left[S(P_1, \ldots, P_r, Y_1, \ldots, Y_r)\right] \\
&= \mathbb{E}\left[s_S\left(P_1, \ldots, P_r, \mathbb{P}_{Y|P_1}, \ldots, \mathbb{P}_{Y|P_r}\right)\right] \\
&= g_S(\mathbb{P}_Y, \ldots, \mathbb{P}_Y) \\
&\qquad + \mathbb{E}\left[d_S\left(P_1, \ldots, P_r, \mathbb{P}_{Y|P_1}, \ldots, \mathbb{P}_{Y|P_r}\right)\right] \\
&\qquad - \mathbb{E}\left[d_S\left(\mathbb{P}_Y, \ldots, \mathbb{P}_Y, \mathbb{P}_{Y|P_1}, \ldots, \mathbb{P}_{Y|P_r}\right)\right].
\end{aligned} \tag{49}$$

Proof is identical to proof of Lemma 4.1. The first term is the generalized entropy, the second the calibration, and the third the sharpness term.

Thus, every proper U-score S induces a proper calibration error defined as

$$\begin{aligned}
&\mathrm{CE}_S(f) \\
&= \mathbb{E}\left[d_S\left(f(X_1), \ldots, f(X_r), \mathbb{P}_{Y|f(X_1)}, \ldots, \mathbb{P}_{Y|f(X_r)}\right)\right] \\
&\text{with iid } X_1, \ldots, X_r.
\end{aligned} \tag{50}$$

Since proper U-scores are identical to proper scores for $r = 1$, this definition of proper calibration errors does not contradict definitions or findings in the main paper. For any strictly proper U-score S, CE_S of model f is zero if and only if f is calibrated. This directly follows from the property of the U-divergence. But, it should be noted that we cannot assume every property holding for $r = 1$ also holds for $r \in \mathbb{N}$. Investigating this can be seen as potential future work.

An example with $r = 2$:

For positive definite kernel matrix k, define

$$S(P_1, P_2, y_1, y_2) = (P_1 - e_{y_1})^\intercal k(P_1, P_2)(P_2 - e_{y_2}) \tag{51}$$

which gives

$$g_S(Q_1, Q_2) = 0 \tag{52}$$

and

$$\begin{aligned}
&d_S(P_1, P_2, Q_1, Q_2) \\
&= (P_1 - Q_1)^\intercal k(P_1, P_2)(P_2 - Q_2)
\end{aligned} \tag{53}$$

and the calibration term

$$\begin{aligned}
&\mathbb{E}\left[d_S\left(P_1, P_2, \mathbb{P}_{Y|P_1}, \mathbb{P}_{Y|P_2}\right)\right] \\
&= \mathbb{E}\left[\left(P_1 - \mathbb{P}_{Y|P_1}\right)^\intercal k(P_1, P_2)\left(P_2 - \mathbb{P}_{Y|P_2}\right)\right]
\end{aligned} \tag{54}$$

If $P_1, P_2 \sim \mathbb{P}_{f(X)}$, then this is the squared KCE (SKCE) of f [55]. Widmann et al. [55] proved that the SKCE is zero if and only if f is calibrated, and f is calibrated if $f(X) = \mathbb{P}_{Y|f(X)}$, which includes $f(X) = \mathbb{P}_{Y|X}$. Consequently, the associated divergence is not uniquely minimized by the target distribution. Thus, the score of the SKCE is proper but not strictly proper.

Interestingly, $\mathbb{E}\left[d_S\left(\mathbb{P}_Y, \mathbb{P}_Y, \mathbb{P}_{Y|f(X)}, \mathbb{P}_{Y|f(X')}\right)\right] = 0$ for X, X' iid, i.e. the score of the SKCE only measures calibration and ignores sharpness. This fact is consistent with all previous findings of the SKCE.

G Extended experiments

In this section, we provide further details of the experimental setup and report additional results. This includes results in the squared space, where the upper bound estimator is minimum-variance unbiased. Further, we present results on the Friedman 1 regression problem, which is also used by Widmann et al. [56].

G.1 Details on the ECE estimator simulation in Figure 2

We simulate model predictions of a 100 class classification problem with validation set size of 10'000 and test set size of 10'000. For this, we sample the model predictions from a multivariate logistic normal distribution [1], since it is a lot more flexible in its covariance matrix than a dirichlet distribution. This brings the samples closer to real-world model predictions. We sample the covariance matrix from an inverse-wishart distribution with a scale matrix of $I_{100}/0.01$. The scale matrix was tempered in such a way to receive model predictions with ≈ 87.6 % classification accuracy. We will explain the label sampling in the following. Again, we aimed for realistic values.

Now, we want a model-target relation of which we know that the model is calibrated. For this, we iterate over every model prediction and use each model prediction as a categorical distribution from which we sample the label. Consequently, each model prediction is the ground truth distribution of each label. This gives us calibrated prediction-target pairs, which we used to estimate the ECE of the perfectly calibrated 'model' in Figure 2 (blue line). Next, to gradually decrease the level of calibration, we scale the predictions via different temperatures in the logit space. Thus, we know that the 'model' of mediocre calibration (orange line) is worse than the 'model' of perfect calibration, and better than the 'model' of bad calibration (green line).

G.2 Details on experimental setup of Section 5

The experiments are conducted across several model-dataset combinations, for which logit sets are openly accessible [4] [29, 44]. This includes the models LeNet 5 [32], ResNet 50 (with and without pretraining), ResNet 50 NTS, ResNet 101 (with and without pretraining) ResNet 110, ResNet 110 SD, ResNet 152, ResNet 152 SD [20], Wide ResNet 32 [60], DenseNet 40, DenseNet 161 [22], and PNASNet5 Large [33] and the datasets CIFAR10, CIFAR100 [27], and ImageNet [9]. We did not conduct model training by ourselves, and refer to [29] and [44] for further details. Validation and test set splits are predefined in every logit set. We include TS, ETS, and DIAG as injective recalibration methods. For optimization of TS and ETS, we modified the available implementation of Zhang et al. [61] and used the validation set as calibration set. For DIAG, we used the exact implementation of Rahimi et al. [44].

For every dataset we investigate ten ticks of different (sampled) test set sizes. The ticks are determined to be equally apart in the $\log_2$ space. The minimum is always 100 and the maximum the full available test set size. We use repeated sampling with subsequent averaging to counteract the increased estimation variance for low test set sizes. The estimated standard errors are also shown in the plots, but they are often barely visible. The number of samples in each tick is along the following:

- Tick 1 ($n = 100$): 20000
- Tick 2: 15842
- Tick 3: 12168
- Tick 4: 8978
- Tick 5: 6272
- Tick 6: 4050
- Tick 7: 2312
- Tick 8: 1058
- Tick 9: 288
- Tick 10 (full test set): 2

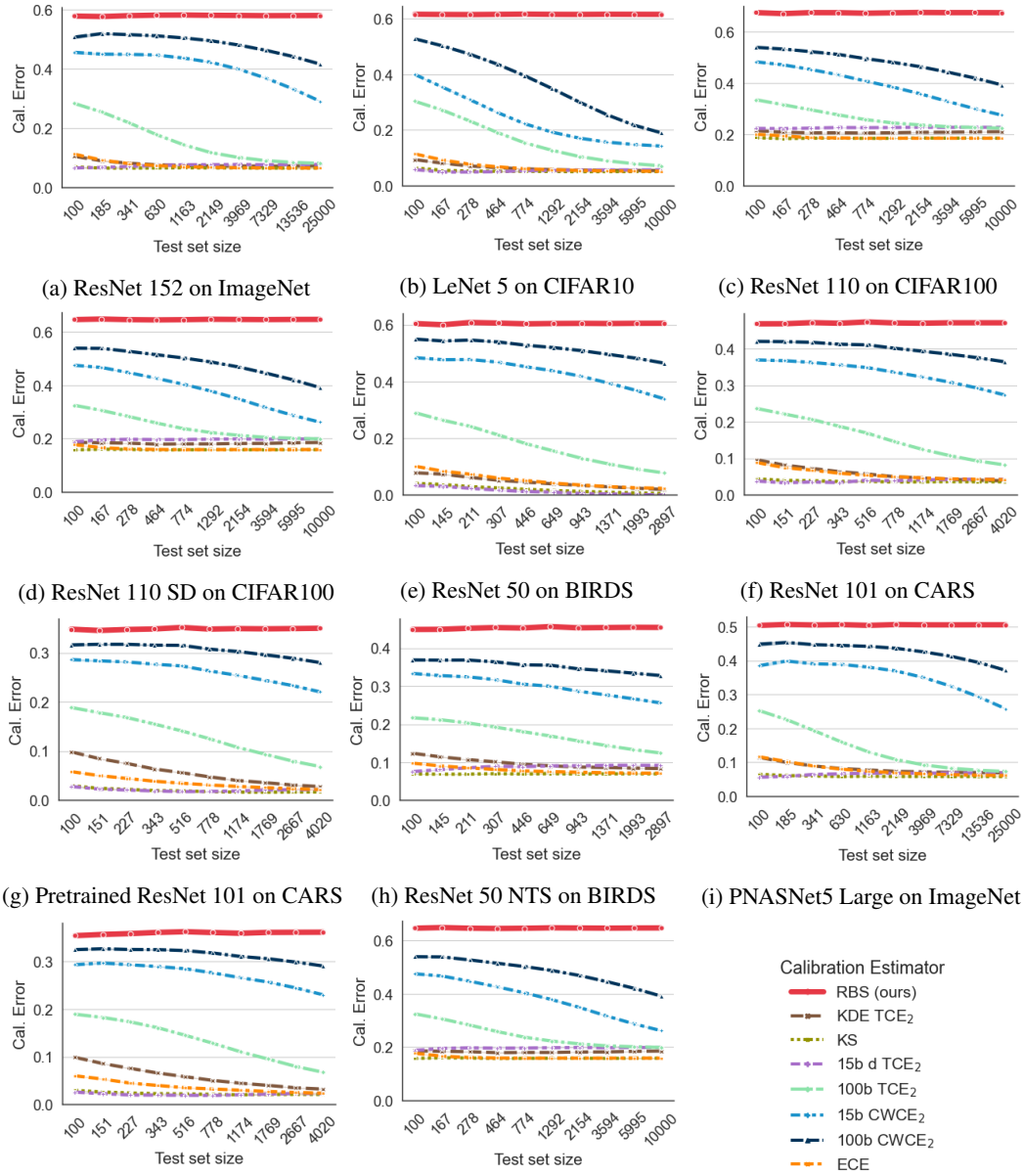
[4] https://github.com/markus93/NN_calibration/ and https://github.com/AmirooR/IntraOrderPreservingCalibration

The seeds for the sampling of the experiments have been saved. Since we choose the amount of samples such that the estimation standard error is low, we expect similar results no matter the chosen seed.

All experiments have been computed on a machine with 1007 GB RAM and two Intel(R) Xeon(R) Gold 6230R CPU @ 2.10GHz.

G.3 Estimated model calibration

Calibration errors according to different estimators and for different model-dataset combinations are shown in figure 5 and first row of figure 7 (in squared space). These experiments confirm that the proposed upper bound is stable across a multitude of settings.



(a) ResNet 152 on ImageNet (b) LeNet 5 on CIFAR10 (c) ResNet 110 on CIFAR100

(d) ResNet 110 SD on CIFAR100 (e) ResNet 50 on BIRDS (f) ResNet 101 on CARS

(g) Pretrained ResNet 101 on CARS (h) ResNet 50 NTS on BIRDS (i) PNASNet5 Large on ImageNet

(j) Pretrained ResNet 50 on CARS (k) ResNet 110 SD on CIFAR100

Figure 5: Different calibration error estimates versus the test set size. The red line corresponds to the square root of the Brier score which is an upper bound of the true $\sqrt{\text{CE}_2}$. The other estimators are lower bounds.

G.4 Recalibration improvement

In the main text we investigated recalibration improvement of common estimators for the calibration error and compared their reliability to RBS. According to Proposition 4.5 and since RBS is asymptotically unbiased and consistent, it can be regarded as a reliable approximation of the real improvement of the recalibration methods. However, if we move to the squared space, our proposed upper bound is even provably reliable since it has a minimum-variance unbiased estimator. This motivates further experiments comparing existing calibration errors in the squared space, which we describe in the following. Here, we first report additional results comparing common estimators to RBS; we then report results in the squared space. We start with a formal description of the problem and experimental setup.

Let D be a sampled subset of the full test set. Let f be the underlying model and h an optimized recalibration method. Let e be an calibration error estimator taking a dataset and a model as inputs. The recalibration improvement according to estimator e is estimated via $e(D, f) - e(D, h \circ f)$.

Recalibration improvement of common estimators We compute the recalibration improvement of common estimators on several test set samples of a given size and plot the average of these on the y-axis. We extend the results reported in the main text by covering additional datasets, models and architectures. These extended experiments confirm the findings reported in the main text, namely that only RBS reliably quantifies the improvement in calibration error after recalibration (Fig. 6; standard errors are shown).

Recalibration improvement in the squared space The recalibration improvement in the squared space according to estimator e is estimated via $(e(D, f))^2 - (e(D, h \circ f))^2$. Only our proposed upper bound yields provably unbiased estimates of the true recalibration improvement w.r.t. CE_2. In contrast to our approach, all other estimators are sensitive to test set size and/or substantially misestimate the true recalibration improvement in squared space (figure 7).

G.5 Variance regression

In the following, we give more details on the variance regression experiment in the main paper, but also add further results of the Friedman 1 regression problem.

In all following scenarios, we are interested in the effect of recalibration towards predictive uncertainty. For this, we use Platt scaling ($x \to wx + b$ with parameters $w, b \in \mathbb{R}$) of the variance output and optimize w and b with the L-BFGS optimizer on the validation set. Further, since Platt scaling is injective, we apply Theorem 4.3 and Proposition 4.5 to treat the DSS score as an calibration error for recalibration. Consequently, optimizing Platt scaling with the DSS score is equivalent to optimizing the associated calibration error.

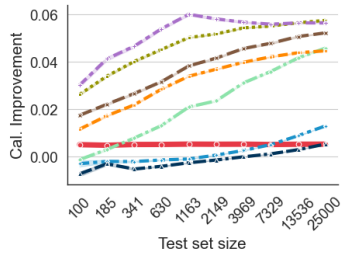
We will use this recalibration procedure in each iteration during model training, but without modifying the model for the next training step.

Widmann et al. [56] used the MSE as training objective, while we use the DSS, as it is a natural extension of the MSE to variance regression.

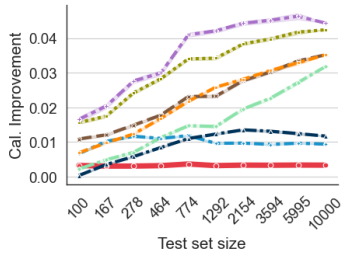
We repeat each experiment with five distinct seeds and aggregate the results, giving the characteristic error bands in each figure.

G.5.1 UCI dataset *Residential Building*

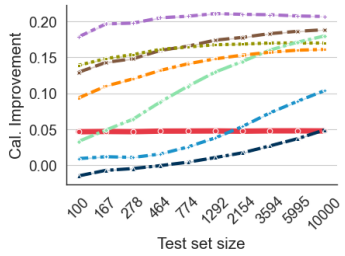
The Residential Building dataset consists of 107 features and 372 data instances. To have similar conditions as the Friedman 1 regression problem in the next section, we split the dataset into a training, validation, and test set with sizes 100, 100, and 173. We use a fully-connected mixture density network as Widmann et al. [56], except we are also using an output node for the variance prediction, and reduce its size for a more stable training. Specifically, it consists of 107 input nodes, 200 hidden nodes, and 2 output nodes. Similar to Widmann et al. [56], we use Adam as model optimizer with default parameters (0.001 learning rate, 0.9 first momentum decay, 0.999 second momentum decay).



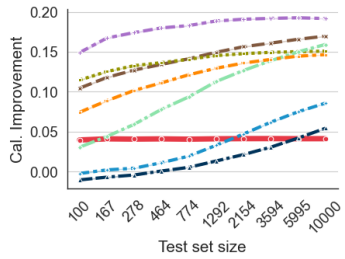
(a) TS applied to ResNet 152 on ImageNet



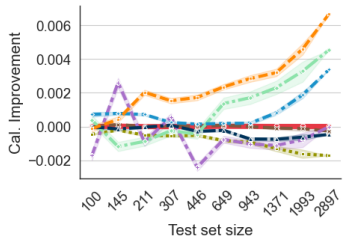
(b) TS applied to LeNet 5 on CIFAR10



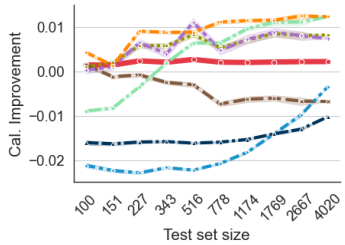
(c) TS applied to ResNet 110 on CIFAR100



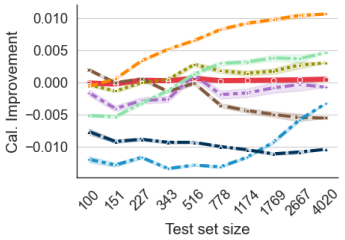
(d) TS applied to ResNet 110 SD on CIFAR100



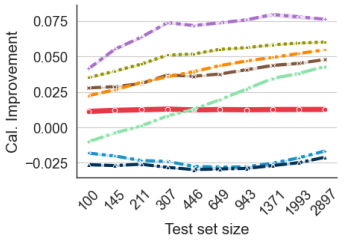
(e) TS applied to ResNet 50 on BIRDS



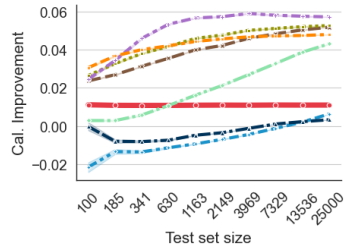
(f) DIAG applied to ResNet 101 on CARS



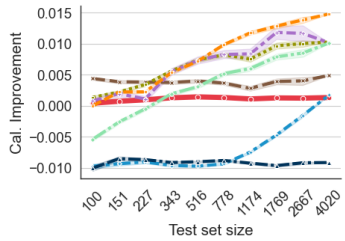
(g) DIAG applied to pretrained ResNet 101 on CARS



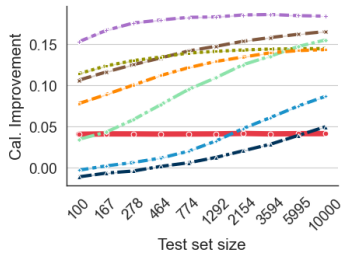
(h) DIAG applied to ResNet 50 NTS on BIRDS



(i) DIAG applied to PNASNet5 Large on ImageNet



(j) DIAG applied to pretrained ResNet 50 on CARS



(k) ETS applied to ResNet 110 SD on CIFAR100

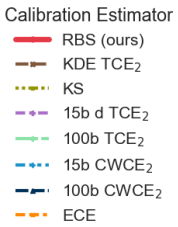


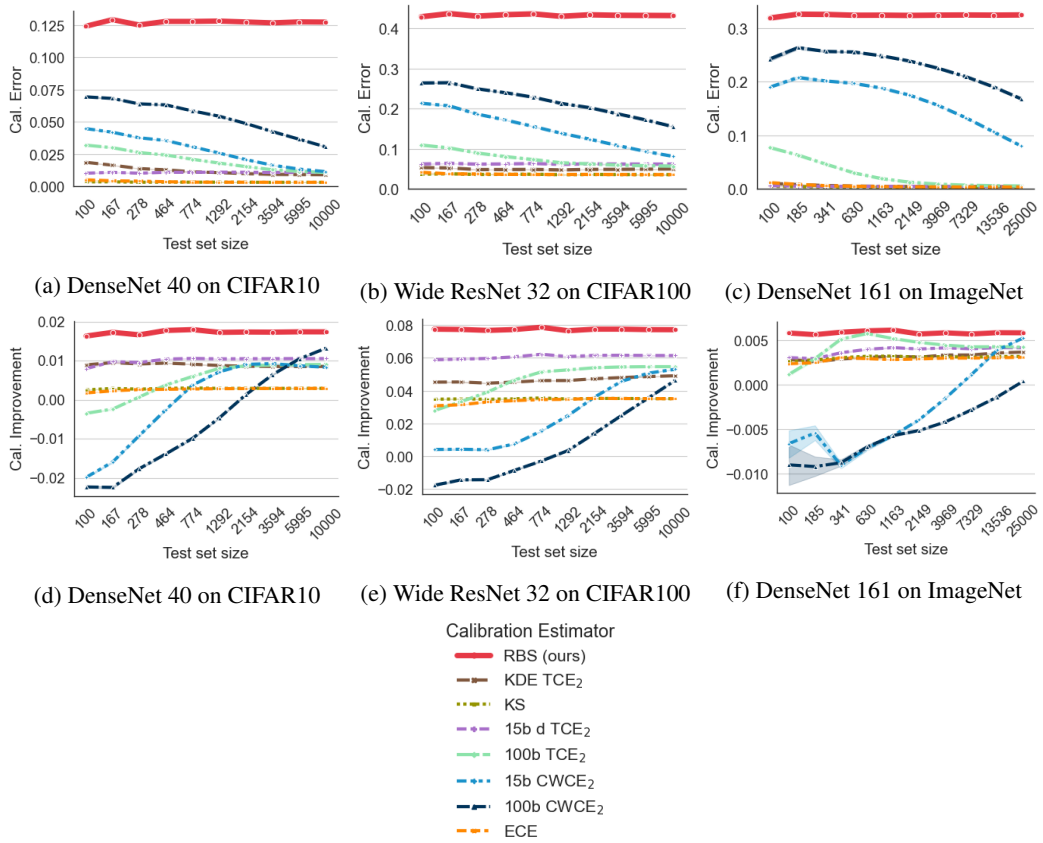
Figure 6: Different calibration improvement estimates versus the test set size. The red line corresponds to the square root of the Brier score.

We show similar results as in Figure 4 but with aggregations from different runs with distinct seeds. The evaluations are depicted in 8 and repeat the findings in the main paper.

G.5.2 Friedman 1

The Friedman 1 regression problem consists of ten feature variables but only five influence the target variable [11]. The target variable is given by

$$Y = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \epsilon \tag{55}$$



(a) DenseNet 40 on CIFAR10 (b) Wide ResNet 32 on CIFAR100 (c) DenseNet 161 on ImageNet

(d) DenseNet 40 on CIFAR10 (e) Wide ResNet 32 on CIFAR100 (f) DenseNet 161 on ImageNet

Figure 7: **First row:** Different squared calibration error estimates versus the test set size. The red line corresponds to the Brier score which is an upper bound of the true CE_2. The other errors are lower bounds. **Second row:** Estimated improvements in the squared space of injective recalibration methods in different settings. Our approach captures the true improvement w.r.t. CE_2 in an unbiased manner.

with input $X_i \sim U(0,1)$ independently uniformly distributed for $i = 1, \ldots, 10$, and noise $\epsilon \sim \mathcal{N}(0,1)$. It was used lately to investigate model calibration in the regression case [56]. We slightly modify the Friedman 1 regression problem to have varying variance depending on the sixth feature variable, i.e. $\epsilon \sim \mathcal{N}(0, 0.5 + X_6)$. This makes it non-trivial to give an estimate of the predictive uncertainty in the form of the predicted variance. We sample a training, validation, and test set, each of size 100 similar to Widmann et al. [56].

We use the same fully-connected mixture density network as Widmann et al. [56], except we are also using an output node for the variance prediction. Further, we use the same training details as Widmann et al. [56]. We repeat each run three times and aggregate the results.

We again compare DSS, SKCE, and average predicted variance throughout model training with and without recalibration. We depict the performance according to various errors during model training in Figure 9. As can be seen, recalibration adjusts overfitting of the predicted variance. Consequently, the uncertainty communication of the model is improved. Further, the SKCE seems to be less influenced by the variance calibration and more so by the mean calibration. This is a significant drawback when uncertainty communication is done via the predicted variance.

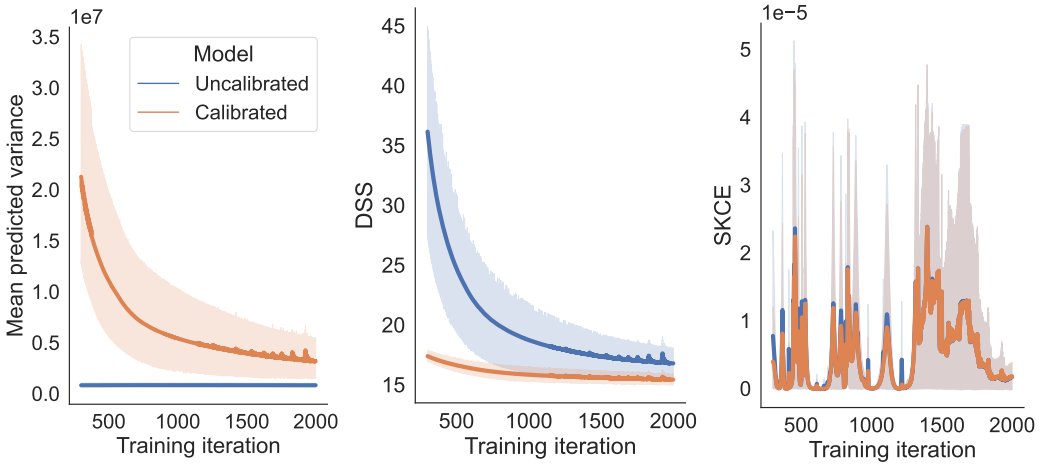


Figure 8: **Left:** Average predicted variance throughout model training before and after recalibration. Initially, due to a bad fit, recalibration adjusts the variance accordingly for better communicated uncertainty. Once the model fit improves, the predicted variance requires less adjustment due to less uncertainty in each prediction. **Middle:** DSS communicates reasonably changes in the variance due to recalibration. **Right:** SKCE fails to capture the variance trend and behaves erratically.

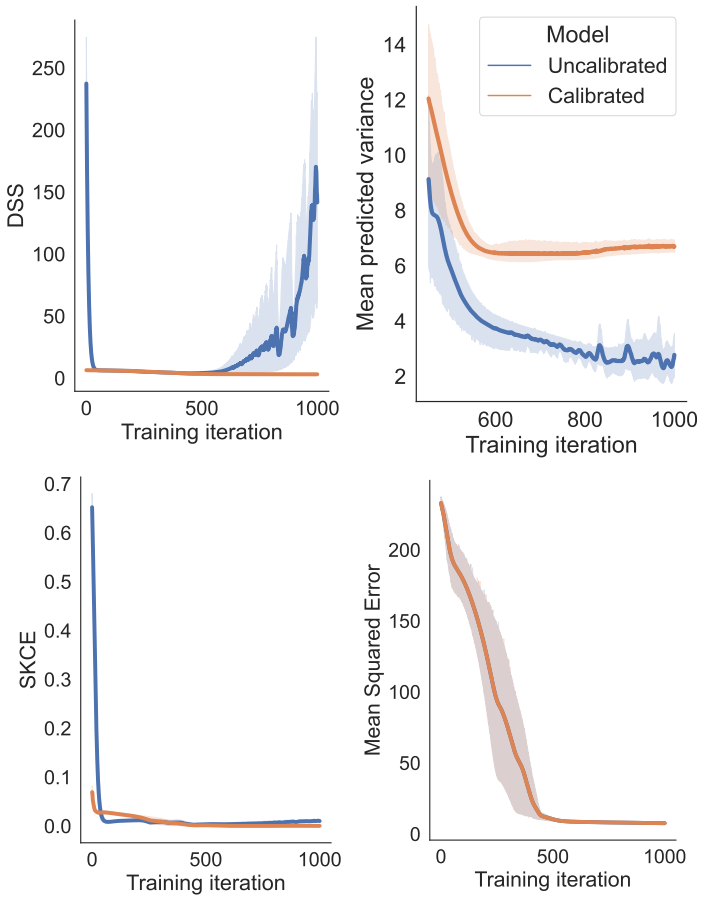


Figure 9: **Upper left:** DSS shows that overfitting occurs at some point during training. Recalibration successfully adjusts this overfit. This indicates that most of the overfit is regarding variance and not mean prediction. **Upper right:** Average predicted variance starting from the point of overfitting. Recalibration adjusts the steadily decreasing predicted variance to a constant level. **Lower left:** SKCE signals improved calibration at the start of training but remains relatively unchanged by the variance overfit. **Lower right:** The MSE curve confirms that the predicted mean is not overfitted and suggests the SKCE is more sensitive to the calibration of the mean than the calibration of the variance estimate. Our recalibration does not influence the predicted mean, thus we omit the recalibrated model from this subfigure.