Consistent and Asymptotically Unbiased Estimation of Proper Calibration Errors

Teodora Popordanoska[*]
ESAT-PSI
KU Leuven, Belgium

Sebastian G. Gruber[*]
German Cancer Research Center (DKFZ)
German Cancer Consortium (DKTK)
Goethe University Frankfurt, Germany

Aleksei Tiulpin
HST Unit
University of Oulu, Finland
Preon Health Oy, Finland

Florian Buettner
German Cancer Research Center (DKFZ)
German Cancer Consortium (DKTK)
Frankfurt Cancer Institute, Germany
Goethe University Frankfurt, Germany

Matthew B. Blaschko
ESAT-PSI
KU Leuven, Belgium

Abstract

Proper scoring rules evaluate the quality of probabilistic predictions, playing an essential role in the pursuit of accurate and well-calibrated models. Every proper score decomposes into two fundamental components – *proper calibration error* and refinement – utilizing a Bregman divergence. While uncertainty calibration has gained significant attention, current literature lacks a general estimator for these quantities with known statistical properties. To address this gap, we propose a method that allows consistent, and asymptotically unbiased estimation of *all* proper calibration errors and refinement terms. In particular, we introduce Kullback–Leibler calibration error, induced by the commonly used cross-entropy loss. As part of our results, we prove a relation between refinement and f-divergences, which implies information monotonicity in neural networks, regardless of which proper scoring rule is optimized. Our experiments validate empirically the claimed properties of the proposed estimator and suggest that the selection of a post-hoc calibration method should be determined by the particular calibration error of interest.

[*]Shared first authorship. The authors can change the order for their own purposes. Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

Risk minimization is the cornerstone of machine learning, where the goal is to develop models that are accurate and provide well-calibrated uncertainty estimates. Central to this pursuit are proper scoring rules (Gneiting & Raftery, 2007), which measure the quality of probabilistic predictions via dissimilarity measures of probability distributions, known as Bregman divergences (Bregman, 1967; Ovcharov, 2018). A significant breakthrough in this realm is the decomposition of the expected loss associated with a proper score into calibration and refinement (Murphy, 1973; DeGroot & Fienberg, 1981; Blattenberger & Lad, 1985; Bröcker, 2009). Facilitated by this result, understanding and mitigating calibration error (CE) has become one of the key concerns for applications like healthcare (Haggenmüller et al., 2021; Katsaouni et al., 2021), climate modelling (Gneiting & Raftery, 2005; Kashinath et al., 2021) and autonomous driving (Yurtsever et al., 2020).

On the most fundamental level, calibration errors compare a predictive distribution with a conditional target distribution (Gruber & Buettner, 2022). For this, the machine learning literature proposed a wide range of different calibration errors in the multi-class setting (Bröcker, 2009; Kull & Flach, 2015; Naeini et al., 2015; Vaicenavicius et al., 2019; Kumar et al., 2018, 2019; Widmann et al., 2019; Gupta et al., 2020; Zhang et al., 2020; Popordanoska et al., 2022; Gruber & Buettner, 2022). The most common ones are based on absolute or squared differences. However, current literature lacks a general estimator of calibration error and refinement induced by *any* Bregman divergence.

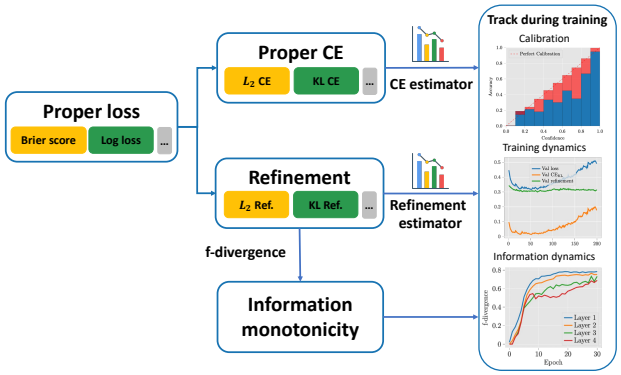


Figure 1: Proper losses decompose into calibration error (CE) and refinement. We propose consistent and asymtotically unbiased estimators for all proper CE and refinement terms. Moreover, we derive a novel connection between refinement and information monotonicity via f-divergences. The proposed estimators can be used to track training dynamics, information flow during training, and model calibration.

We approach the study of calibration errors through the notion of **proper calibration errors** (Gruber & Buettner, 2022), which is a general class of calibration errors derived from risk minimization via proper scores. For example, the Brier score induces the squared L_2 calibration error, for which there exists a consistent estimator (Popordanoska et al., 2022). Estimating the KL calibration error, which is induced by the most common proper score in classification – the categorical negative log likelihood[1], remains an open challenge.

In this work, we introduce a *consistent and asymptotically unbiased* estimator for all proper calibration errors and refinement terms. Additionally, our proposed estimator can also be used for estimating the so-called model sharpness. Similar to how every proper score generates a proper calibration error, a sharpness term is generated. For example, the sharpness induced by the log loss is the mutual information between prediction and target variable (Huszar, 2013). It is investigated in the context of general forecasts, but is not well understood for neural networks (DeGroot & Fienberg, 1983; Murphy & Winkler, 1977; Blattenberger & Lad, 1985; Bröcker, 2009; Murphy, 1973). We show that any model sharpness is identical to an f-divergence (Csiszár, 1972). Through the information monotonicity of the f-divergence, we then derive the concept of general information monotonicity in neural networks. This gives a novel perspective into the workings of neural networks and illustrates how the information bottleneck theory is a more general concept beyond mutual information.

[1]We use the terms categorical negative log-likelihood, log-loss and cross-entropy loss interchangeably.

Our **contributions** are depicted in Figure 1 and can be summarized as follows:

1. We provide a general estimator for all proper calibration errors and refinement terms in classification, which is consistent and its bias converges with rate $\mathcal{O}\left(n^{-1}\right)$.
2. We show that model sharpness can be formulated as a multi-distribution f-divergence. Based on this result, we derive the concept of information monotonicity in neural networks beyond mutual information.
3. We conduct experiments showcasing the empirical properties of our estimator, as well as its utility in selecting an appropriate post-hoc calibration method for the desired calibration error.

2 RELATED WORK

In this section, we give a brief overview of estimating calibration errors. Calibration errors are notoriously difficult to estimate since they compare a prediction $g(X)$ with the conditional expectation $\mathbb{E}[Y \mid g(X)]$, where Y is the one-hot encoded target variable, X the input variable, and g the predictive model. The difficulty arises since $g(X)$ is in general a multivariate continuous random variable. Originally, model calibration was only considered for a finite set of predictions (Murphy, 1973), simplifying the estimation since $g(X)$ is then a discrete random variable. Platt (1999) used histogram estimation, which transforms the continuous prediction space to a discrete one, to assess the calibration of a continuous binary model. Different ways to define the histogram bins are equal width or equal mass binning techniques (Nguyen & O'Connor, 2015). The number of bins and the binning scheme can significantly influence the estimated value (Kumar et al., 2019) and there is no optimal default since every setting has a different bias-variance tradeoff (Nixon et al., 2019). Further, using a fixed binning scheme represents a lower bound of the respective calibration error (Kumar et al., 2019; Vaicenavicius et al., 2019; Ma & Blaschko, 2021). In consequence, Vaicenavicius et al. (2019) propose to use adaptive binning similar to other approaches in histogram estimation (Nobel, 1996). Dimitriadis et al. (2021) and Roelofs et al. (2022) introduce approaches to optimize the number of bins. Zhang et al. (2020) circumvent binning schemes by using kernel density estimation via Bayes' theorem.
The first calibration error estimator for a multi-class model was given by Naeini et al. (2015) and is still the most commonly used measure to quantify calibration, known as expected calibration error (ECE) (Guo et al., 2017).
The ECE is a special case of top-label confidence

calibration since it only uses the predicted top-label confidence $\max_{i \in \mathscr{Y}} g_i(X)$ and compares it with the conditional accuracy $\mathbb{E}[Y_C \mid \max_{i \in \mathscr{Y}} g_i(X)]$ with $C = \arg\max_{i \in \mathscr{Y}} g_i(X)$. In contrast, class-wise calibration estimation uses all predicted classes, but only in isolation from each other, since it compares $g_i(X)$ with $\mathbb{E}[Y_i \mid g_i(X)]$ for each class i. Both notions depend on estimating a conditional distribution given a univariate continuous random variable. Consequently, estimation schemes of one notion can be applied to the other. For example, Kull et al. (2019) and Nixon et al. (2019) use histogram binning estimation to estimate class-wise calibration. To circumvent the need of binning, Kumar et al. (2018) propose the maximum mean calibration error, based on positive definite kernels, and Gupta et al. (2020) introduce the Kolmogorov-Smirnov calibration error. In contrast to top-label and class-wise calibration, canonical calibration refers to the case when the model prediction $g(X)$ matches the conditional target $\mathbb{E}[Y \mid g(X)]$ almost surely (Vaicenavicius et al., 2019; Popordanoska et al., 2022). It is substantially more difficult to estimate with increasing classes since $g(X)$ also increases in dimensionality. This makes histogram based approaches infeasible and kernel density estimation strongly dependent on the scalability of the kernel to higher dimensions. Popordanoska et al. (2022) propose a specific kernel choice to estimate canonical L_p calibration errors. Further, Widmann et al. (2019) introduce the kernel calibration error based on positive definite kernels. It can be seen as a canonical extension of the maximum mean calibration error.

3 PROPER CALIBRATION ERRORS

In classification, it is common to use a loss function of the form $L\colon \Delta^k \times \mathscr{Y}$, where Δ^k is the $(k-1)$ dimensional simplex and $\mathscr{Y}$ the sample space of the one-hot encoded target variable Y. Further, assume X is the feature variable with realizations in a space $\mathscr{X}$. To optimize a model $g\colon \mathscr{X} \to \Delta^k$ mapping from the feature space into the probability simplex, we use the expected loss

$$\mathscr{R}(g) \coloneqq \mathbb{E}[L(g(X), Y)], \tag{1}$$

which is referred to as **risk**. If L is minimized by the Bayes classifier $g^*(x) \coloneqq \mathbb{E}[Y \mid X = x]$, then $-L$ is a proper score (Gneiting & Raftery, 2007). In that case, we may also refer to L as a proper loss (Williamson, 2014). The best achievable (negative) risk for a given target distribution q is given by $F(q) \coloneqq -\inf_{p \in \Delta^k} \mathbb{E}_{Y \sim q}[L(p, Y)]$. The function F is convex if and only if $-L$ is a proper score (Gneiting & Raftery, 2007).

Every differentiable proper score is uniquely associated with a Bregman divergence. A Bregman divergence (Bregman, 1967) in a k-dimensional space $U \subset \mathbb{R}^k$ is characterized by a continuously differentiable, strictly convex function $F\colon U \to \mathbb{R}$, with $D_F(p, q) \coloneqq F(p) - F(q) - \langle \nabla F(q), p - q \rangle$. Special cases are the squared Euclidean distance with $F(p) = \|p\|_2^2$ as 2-norm, and the Kullback–Leibler divergence with $F(p) = \sum_{i=1}^k p_i \log(p_i)$ as negative Shannon entropy.

Following Ovcharov (2018), the risk related to a proper loss is connected to a Bregman divergence via

$$\mathscr{R}(g) + \mathbb{E}[F(\mathbb{E}[Y \mid X])] = \mathbb{E}[D_F(\mathbb{E}[Y \mid X], g(X))]. \tag{2}$$

Consequently, risk minimization is equivalent to minimizing an expected Bregman divergence since they only differ by a model independent constant. The most prominent example as risk is the expected log loss, which induces the Kullback–Leibler divergence as Bregman divergence (Ovcharov, 2018).

We can use Bregman divergences to assess the canonical calibration of a model. Bröcker (2009) showed that proper scores in classification can be decomposed into calibration and refinement terms. Similarly, we can do the same for the risk, namely

$$\begin{aligned}\mathscr{R}(g) = &\underbrace{\mathbb{E}[D_F(\mathbb{E}[Y \mid g(X)], g(X))]}_{\text{Calibration}} \\ &+ \underbrace{\mathbb{E}[-F(\mathbb{E}[Y \mid g(X)])]}_{=:\mathrm{REF}_F(g)\ \text{(Refinement)}}.\end{aligned} \tag{3}$$

The sharpness of a model is defined via (DeGroot & Fienberg, 1981)

$$\mathrm{SHARP}_F(g) = \mathbb{E}[D_F(\mathbb{E}[Y \mid g(X)], \mathbb{E}[Y])]. \tag{4}$$

It is zero for non-informative classifiers. If F is the negative Shannon entropy, then the sharpness is equivalent to the mutual information between output variable and target variable. Model sharpness is related to the refinement term (DeGroot & Fienberg, 1981, 1983; Kull & Flach, 2015; Kuleshov & Deshpande, 2022) by

$$-\mathrm{SHARP}_F(g) = F(\mathbb{E}[Y]) + \underbrace{\mathbb{E}[-F(\mathbb{E}[Y \mid g(X)])]}_{\text{(Refinement)}}. \tag{5}$$

Thus, sharpness is maximized via risk minimization.

Gruber & Buettner (2022) defined a calibration error of the form

$$\mathrm{CE}_F(g) \coloneqq \mathbb{E}[D_F(\mathbb{E}[Y \mid g(X)], g(X))] \tag{6}$$

as **proper calibration error**. Since any convex function defined on the simplex can be related to a proper score (Ovcharov, 2015), any Bregman divergence can be used to define a proper calibration error. Murphy

(1973) derived the squared calibration error from the Brier score. It is defined as

$$\mathrm{CE}_2^2(g) = \mathbb{E}\left[\left\|\mathbb{E}[Y \mid g(X)] - g(X)\right\|_2^2\right]. \tag{7}$$

With the square root applied, it is also member of the so-called L_p calibration errors, where the 2-norm is replaced by a general p-norm (Naeini et al., 2015; Kumar et al., 2019; Wenger et al., 2020; Popordanoska et al., 2022). Popordanoska et al. (2022) propose an estimator of L_p calibration errors via kernel density estimation and offer evaluations of the squared calibration error. Another example of a proper calibration error can be derived from the log-likelihood and results in a calibration error based on the Kullback–Leibler divergence D_{KL} given by

$$\mathrm{CE}_{\mathrm{KL}}(g) = \mathbb{E}\left[D_{\mathrm{KL}}\left(\mathbb{E}[Y \mid g(X)], g(X)\right)\right]. \tag{8}$$

As all differentiable proper calibration errors are induced by a Bregman divergence, non-negativity applies immediately, but the range is dependent on which divergence is employed: $\mathrm{CE}_2^2(g) \in [0, 2]$, while $\mathrm{CE}_{\mathrm{KL}}(g) \in [0, \infty)$. Since we are only using distributions as inputs for calibration errors, the Kullback–Leibler CE is more principled according to information theory than the squared calibration error (MacKay, 2003). The squared error implies a Euclidean geometry for its inputs since the input space is non-bounded. But, distributions are non-negative and normalized, and, consequently, exist in a bounded space. The Kullback–Leibler divergence is a better representation of these restrictions, since it is also not defined for negative inputs (MacKay, 2003). Currently, no estimator for general proper calibration errors exist, including the Kullback–Leibler case. As part of this work we propose a consistent, asymptotically unbiased and differentiable estimator for any proper calibration error.

4 ESTIMATING PROPER CALIBRATION ERRORS

Given an i.i.d. labeled data sample $\{(x_i, y_i)\}_{1 \le i \le n}$, a generic Bregman divergence calibration error estimator can be defined via

$$\mathrm{CE}_F(g) \approx \frac{1}{n}\sum_{h=1}^{n}\left(F\left(\widehat{\mathbb{E}[Y \mid g(x_h)]}\right) - F(g(x_h)) - \left\langle \nabla F\left(g(x_h)\right), \widehat{\mathbb{E}[Y \mid g(x_h)]} - g(x_h)\right\rangle\right). \tag{9}$$

It is straightforward to verify that the application of $F(x) = \|x\|_2^2$ recovers exactly the L_2 special case of the estimator given by (Popordanoska et al., 2022, Equation (9)), while setting $F(p) = \langle p, \log(p)\rangle$ yields

$$\mathrm{CE}_{\mathrm{KL}}(g) \approx \frac{1}{n}\sum_{h=1}^{n}\left\langle \widehat{\mathbb{E}[Y \mid g(x_h)]}, \log\left(\frac{\widehat{\mathbb{E}[Y \mid g(x_h)]}}{g(x_h)}\right)\right\rangle, \tag{10}$$

where log and division operations are taken element-wise, and we may interpret the inner product using $\lim_{y \searrow 0} y \log y = 0$ to ensure that it remains well defined on the vertices of the probability simplex. In Table 1 we show our derived estimators for calibration error and refinement induced by Brier score and log loss. Detailed derivations can be found in Appendix E. In the sequel, our notation will use capital letters for unbounded random variables, and lower case variables with subscripts for elements of our data sample. Note that we may still treat elements of the data sample as random variables.

We define the finite sample estimator for the conditional expectation as in Popordanoska et al. (2022, Equation (3))

$$\widehat{\mathbb{E}[Y \mid g(X)]} := \frac{\sum_{j=1}^{n} k(g(X), g(x_j)) y_j}{\sum_{j=1}^{n} k(g(X), g(x_j))}, \tag{11}$$

where a natural choice for k can be the Dirichlet kernel (Popordanoska et al., 2022), defined as

$$k_{Dir}(g(x_i), g(x_j)) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_{jk})}{\prod_{k=1}^{K} \Gamma(\alpha_{jk})} \prod_{k=1}^{K} g(x_i)_k^{\alpha_{jk}-1} \tag{12}$$

with $\alpha_j = \frac{g(x_j)}{h} + 1$ (Ouimet & Tolosana-Delgado, 2022). This results in a differentiable, asymptotically unbiased and consistent estimator of the conditional expectation.

Another common choice for estimating the conditional expectation is by using a binning kernel (which returns 1 if $g(x_i)$ and $g(x_j)$ fall in the same bin, and 0 otherwise), resulting in an estimator given by $\widehat{\mathbb{E}[Y \mid g(X)]} = \frac{1}{|B_{g(x_h)}|}\sum_{j \in B_{g(x_h)}} y_j$, where $B_{g(x_h)}$ denotes the bin into which $g(x_h)$ is assigned. Although this approach allows for faster computation, the estimator is not differentiable, thus preventing it to be directly used as part of calibration regularized training or differentiable recalibration methods. In addition, several papers (Vaicenavicius et al., 2019; Widmann et al., 2019; Ashukha et al., 2020) have raised concerns about asymptotic inconsistency of binned estimators, as well as sensitivity to the binning scheme, and limited scalability with the number of classes.

5 STATISTICAL PROPERTIES

We show here the existence of consistent and asymptotically unbiased estimators of arbitrary proper cali-

Table 1: Proposed estimators of calibration error and refinement, induced by Brier score (first row) and log loss (second row) for a classifier g, one-hot encoded label y, and dataset size n. The term $\mathbb{E}[\widehat{Y \mid g(x_h)}]$ is defined in Equation (11) via kernel density estimation.

Loss $L(g(x), y)$	Calibration error estimator $\widehat{\mathrm{CE}}_F(g)$	Refinement estimator $\widehat{\mathrm{REF}}_F(g)$
$\|g(x) - y\|_2^2$	$\frac{1}{n}\sum_{h=1}^{n} \left\| \mathbb{E}[\widehat{Y \mid g(x_h)}] - g(x_h) \right\|_2^2$	$-\frac{1}{n}\sum_{h=1}^{n} \left\| \mathbb{E}[\widehat{Y \mid g(x_h)}] \right\|_2^2$
$-\langle \log g(x), y \rangle$	$\frac{1}{n}\sum_{h=1}^{n} \left\langle \mathbb{E}[\widehat{Y \mid g(x_h)}], \log \frac{\mathbb{E}[\widehat{Y \mid g(x_h)}]}{g(x_h)} \right\rangle$	$-\frac{1}{n}\sum_{h=1}^{n} \left\langle \mathbb{E}[\widehat{Y \mid g(x_h)}], \log \mathbb{E}[\widehat{Y \mid g(x_h)}] \right\rangle$

bration errors. We first show in Section 5.1 that the optimal big-$\mathcal{O}$ convergence rates for estimators of CE_F and REF_F are the same and that a consistent estimator of one can be used to construct an estimator of the other. We then prove in Section 5.2 via a Taylor series approach that an estimator via REF_F gives asymptotic unbiasedness for *all* Bregman divergences, giving a constructive proof that a single software implementation can provide a good baseline estimator for any proper calibration error, using a function reference for F to parameterize the Bregman divergence. We first show this property for an estimate via REF_F, as it gives the core ideas and is a more compact proof. We subsequently extend our analysis to show the same rates for direct estimation via Equation (9) in Appendix B.

5.1 Estimation via refinement

Equation (9) provides a calibration error estimate directly from the definition of a Bregman divergence, but we show here that we can also use an estimate based on refinement using the decomposition (cf. Equation (53))

$$\mathbb{E}\left[D_F(Y, g(X))\right] - \mathrm{CE}_F(g) = \mathbb{E}[F(Y) - F(\mathbb{E}[Y \mid g(X)])]. \tag{13}$$

From this decomposition and that $\mathbb{E}\left[D_F(Y, g(X))\right]$ can be estimated with an empirical average achieving an unbiased estimator with $\mathcal{O}(n^{-1/2})$ rate of convergence, we see that the difficulty of estimating refinement or CE is essentially the same, as the rate of bias and convergence for an estimator of one can be transferred to the other by subtracting from the empirical estimate of the risk. Furthermore, we note that a simple empirical mean over $\{F(y_i)\}_{1 \le i \le n}$ is the Minimum-Variance Unbiased Estimator (MVUE) of $\mathbb{E}[F(Y)]$ for a finite sample, and it is the estimate of $-\mathbb{E}[F(\mathbb{E}[Y \mid g(X)])]$ that is the primary challenge.

To summarize, assuming an empirical estimator of the refinement (cf. Equation (15)) $\widehat{\mathrm{REF}}_F(g) \approx \mathbb{E}[-F(\mathbb{E}[Y \mid g(X)])]$, we may compute

$$\widehat{\mathrm{CE}}_F(g) \coloneqq -\widehat{\mathrm{REF}}_F(g) + \frac{1}{n}\sum_{i=1}^{n} \left(D_F(y_i, g(x_i)) - F(y_i) \right), \tag{14}$$

and the rates of convergence of $\widehat{\mathrm{CE}}_F(g)$ and its bias will be determined by the rates of $\widehat{\mathrm{REF}}_F(g)$.

5.2 Asymptotic unbiasedness and rate of bias

If we use the empirical estimator of $\mathbb{E}[Y \mid g(X)]$ in Equation (11), the bias converges as $\mathcal{O}(n^{-1})$ while the estimator itself has a rate of $\mathcal{O}(n^{-1/2})$. By the same argument as Gruber & Buettner (2022, Footnote 2),

$$\widehat{\mathrm{REF}}_F(g) \coloneqq -\frac{1}{n}\sum_{h=1}^{n} F\left(\frac{\sum_{j \neq h} k(g(x_h), g(x_j)) y_j}{\sum_{j \neq h} k(g(x_h), g(x_j))} \right) \tag{15}$$

is a consistent and asymptotically unbiased estimator of the refinement $\mathbb{E}[-F(\mathbb{E}[Y \mid g(X)])]$ for all F. We note that continuity of F, a condition required for the argument of Gruber & Buettner (2022, Footnote 2), is guaranteed for all Bregman divergences, as F is differentiable by assumption.

Proposition 5.1. *Estimation of proper calibration errors both by Equation* (14) *and by Equation* (9) *has a* $\mathcal{O}(n^{-1/2})$ *convergence rate.*

Proof. Both the empirical refinement of Equation (15) and the direct estimator of Equation (9) are differentiable non-linear functions of the same ratio estimator of conditional expectation. This ratio estimator has a known convergence of $\mathcal{O}(n^{-1/2})$ (Scott & Wu, 1981). Therefore, direct application of the multivariate delta method to each function yields the desired result. □

Proposition 5.2. *For all proper calibration errors parameterized by some F satisfying the requirements of a Bregman divergence, estimation of CE by Equations* (14) *&* (15) *has a bias that converges as* $\mathcal{O}(n^{-1})$.

Proof. We show the asymptotic rate of bias using a Taylor series expansion. First, define $\Delta = \widehat{\mathbb{E}[Y \mid g(x_h)]} - \mathbb{E}[Y \mid g(x_h)]$. The rate of bias of $\widehat{\mathrm{REF}}_F(g)$ is determined by the rate of bias of each of the summands in Equation (15)

$$\begin{aligned}
&\mathbb{E}\left[F\left(\mathbb{E}\left[\widehat{Y \mid g(x_h)}\right]\right)\right] = \mathbb{E}\left[F\left(\mathbb{E}[Y \mid g(x_h)] + \Delta\right)\right] \\
&\approx F\left(\mathbb{E}[Y \mid g(x_h)]\right) + \mathbb{E}\left[\mathrm{D}\, F\left(\mathbb{E}[Y \mid g(x_h)]\right) \Delta\right] \\
&\quad + \mathbb{E}\left[\frac{1}{2}\Delta^T\, \mathrm{D}^2\, F\left(\mathbb{E}[Y \mid g(x_h)]\right) \Delta\right] + \dots
\end{aligned} \tag{16}$$

where D is the differential operator (Magnus & Neudecker, 1999, Chapt. 5). It is well known that the bias of the 1st order term (a ratio estimator) is $\mathscr{O}(n^{-1})$ and the remaining bias will be dominated by the 2nd order term (Wolter, 2007, Theorem 6.2.5). When $\mathrm{D}^2\, F\left(\mathbb{E}[Y \mid g(x_h)]\right) \neq 0$, this yields

$$\begin{aligned}
&\mathbb{E}\left[\Delta^T\, \mathrm{D}^2\, F\left(\mathbb{E}[Y \mid g(x_h)]\right) \Delta\right] \\
&\asymp \mathrm{Trace}\left[\mathrm{Cov}\left(\mathbb{E}\left[\widehat{Y \mid g(x_h)}\right]\right)\right] \\
&\quad + \underbrace{\left\|\mathbb{E}\left[\mathbb{E}\left[\widehat{Y \mid g(x_h)}\right]\right] - \mathbb{E}[Y \mid g(x_h)]\right\|^2}_{= \left\|\mathrm{Bias}\left(\mathbb{E}\left[\widehat{Y \mid g(x_h)}\right]\right)\right\|^2 = \mathscr{O}(n^{-2})},
\end{aligned} \tag{17}$$

where the notation $\asymp$ is taken here to mean that the left and right side have the same asymptotic rate of convergence in n, and the r.h.s. is due to a bias-variance decomposition of the l.h.s. Finally, we have $\mathrm{Var}\left(\mathbb{E}\left[\widehat{Y \mid g(x_h)}\right]_i\right) = \mathscr{O}(n^{-1})$, which implies

$$\left|\mathrm{Bias}\left(\widehat{\mathrm{REF}}_F(g)\right)\right| = \mathscr{O}(n^{-1}) \tag{18}$$

irrespective of F. □

We therefore conclude that estimation of any proper calibration error via Equation (14) results in a consistent and asymptotically unbiased estimator with convergence $\mathscr{O}(n^{-1/2})$, and bias that converges as $\mathscr{O}(n^{-1})$. In Appendix B, we extend this result to show the same rates for estimation via Equation (9) as well.

6 RELATIONSHIP WITH INFORMATION MONOTONICITY IN NEURAL NETWORKS

A key part of this work is to relate uncertainty calibration to information theoretic principles. First, we summarize relevant concepts and provide the necessary foundation to derive our contributions.

6.1 Background on information monotonicity

In machine learning, information monotonicity is most commonly known through the information bottleneck theory (Slonim & Tishby, 1999; Bialek et al., 2001; Gilad-Bachrach et al., 2003; Chechik et al., 2003; Shamir et al., 2010; Shwartz-Ziv & Tishby, 2017; Saxe et al., 2018). Information monotonicity states that each layer in a neural network is indirectly optimized by the mutual information of the output, resulting in a so-called information flow, or information plane dynamics, throughout the network (Saxe et al., 2018; Goldfeld et al., 2019).

Further, Csiszár (1972) derived the class of f-divergences according to several principles, including information monotonicity. These divergences between distributions are widely applicable, for example throughout statistics (Liese & Vajda, 2006) and in generative modelling (Creswell et al., 2018). Similar to Garcia-Garcia & Williamson (2012) and Duchi et al. (2018), we use the following definition for multiple distributions. Given a convex function $f\colon [0,\infty)^k \to (-\infty,\infty]$ with $f(1,\dots,1) = 0$, the **f-divergence** between distributions $P_1,\dots,P_k$ and Q is defined by

$$I_f\left(P_1,\dots,P_k \parallel Q\right) = \int f\left(\frac{\mathrm{d}P_1}{\mathrm{d}Q},\dots,\frac{\mathrm{d}P_k}{\mathrm{d}Q}\right)\mathrm{d}Q. \tag{19}$$

Following the property of f, we have $I_f\left(P_1,\dots,P_k \parallel Q\right) \geq 0$ with equality if $P_1 = \dots = P_k = Q$. Let M be a Markov kernel transforming a distribution P into a distribution MP, then Garcia-Garcia & Williamson (2012) show that the information monotonicity is given by

$$I_f\left(MP_1,\dots,MP_k \parallel MQ\right) \leq I_f\left(P_1,\dots,P_k \parallel Q\right). \tag{20}$$

In the following, we make the novel connection between f-divergences and refinement via model sharpness.

6.2 A novel generalization of neural network information monotonicity

We now present our contribution regarding the existence of information monotonicity in neural networks. As a preliminary step, we first show that model sharpness has the form of an f-divergence. We defer proofs to Appendix C.

Proposition 6.1 (Sharpness as f-divergence)**.** *Let $F\colon \Delta^k \to \mathbb{R}$ be a convex function and $g\colon \mathscr{X} \to \Delta^k$ a classifier with prediction distributions $P_y \coloneqq \mathbb{P}\left(g(X) \mid Y = y\right)$, and $P \coloneqq \mathbb{P}\left(g(X)\right)$. Then, the model sharpness can be represented as an f-divergence via*

$$\mathrm{SHARP}_F\left(g\right) = I_{F^Y}\left(P_1,\dots,P_k \parallel P\right), \tag{21}$$

where $F^Y(x) \coloneqq F\left(\mathbb{E}\left[Y_1\right] x_1,\dots,\mathbb{E}\left[Y_k\right] x_k\right) - F\left(\mathbb{E}\left[Y_1\right],\dots,\mathbb{E}\left[Y_k\right]\right)$.

Thus, we can interpret the classifier sharpness as the f-divergence between the class-conditional prediction distributions and the marginal prediction distribution. The marginal class distribution determines the weight of each ratio. For a given classification task, it is constant across all models. We can now provide the key result of this section.

Theorem 6.2 (Information monotonicity in neural networks)**.** *For a neural network $g(X) = h_l(\dots(h_1(X)))$ with layers h_i, $i \in \{1,\dots,l\}$, conditional distributions $P_y^i \coloneqq \mathbb{P}(h_i(\dots(h_1(X))) \mid Y = y)$, and marginal distributions $P^i \coloneqq \mathbb{E}\left[P_Y^i\right]$, we have*

$$\begin{aligned} \mathrm{SHARP}_F(g) &= I_{F^Y}\left(P_1^l,\dots,P_k^l \parallel P^l\right) \\ &\leq I_{F^Y}\left(P_1^{l-1},\dots,P_k^{l-1} \parallel P^{l-1}\right) \\ &\leq \dots \leq I_{F^Y}\left(P_1^1,\dots,P_k^1 \parallel P^1\right). \end{aligned} \tag{22}$$

This theorem offers a generalization of the information flow in neural networks. Since sharpness is maximized via risk minimization, the information (as quantified by an f-divergence) is implicitly also maximized in each layer, no matter the proper loss. The signal, which is forward propagated in each layer, follows the known information flow of the information bottleneck theory. A rich literature exists on information bottleneck experiments (Shwartz-Ziv & Tishby, 2017; Saxe et al., 2018; Goldfeld et al., 2019; Wu & Fischer, 2020; Wu et al., 2020; Wang et al., 2022). In Section 7 we will extend on that by monitoring the model sharpness via the refinement term throughout training and by assessing information monotonicity beyond mutual information.

In Section 2, we have seen that calibration and sharpness are two different yet related concepts derived from risk minimization. Theorem 6.2 presents a novel link between information bottleneck theory and uncertainty calibration – two key research areas in deep learning, which consist of rich literature but little exchange. For example, according to our result, optimizing via the information bottleneck theory does not offer calibrated predictions and requires post-hoc uncertainty calibration for trustworthy probability forecasts. This underlines the general importance of calibration estimation.

7 EXPERIMENTS

The outline of the experiments is as follows. We (i) compare the choices of kernel for the conditional expectation estimator, discussed in Section 4; (ii) compare the direct estimator from Equation (9) with the estimator derived via risk in Equation (14); (iii) analyse the empirical properties of the proposed estimator, derived in Section 5.2; (iv) show new insights regarding the choice of post-hoc calibration method, depending on the chosen calibration error; (v) demonstrate the information monotonicity in neural networks, discussed in Section 6.2, for the L_2 case via our proposed estimator.

In all experiments we evaluate class-wise CE, which can be derived from One-vs-Rest risk minimization (cf. Appendix D). The CE estimator obtained by setting $F(x) = \|x\|_2^2$ in Equation 14 will be denoted as $\widehat{\mathrm{CE}}_2^2$, while the one derived from $F(p) = \langle p, \log(p)\rangle$ will be referred to as $\widehat{\mathrm{CE}}_{\mathrm{KL}}$. The bandwidth of the Dirichlet kernel is determined through a combination of a leave-one-out maximum likelihood estimation and visual inspection of the resulting density. Typical values range from 0.01 to 0.0001. The source code and trained models can be found at: `https://github.com/tpopordanoska/proper-calibration-error`.

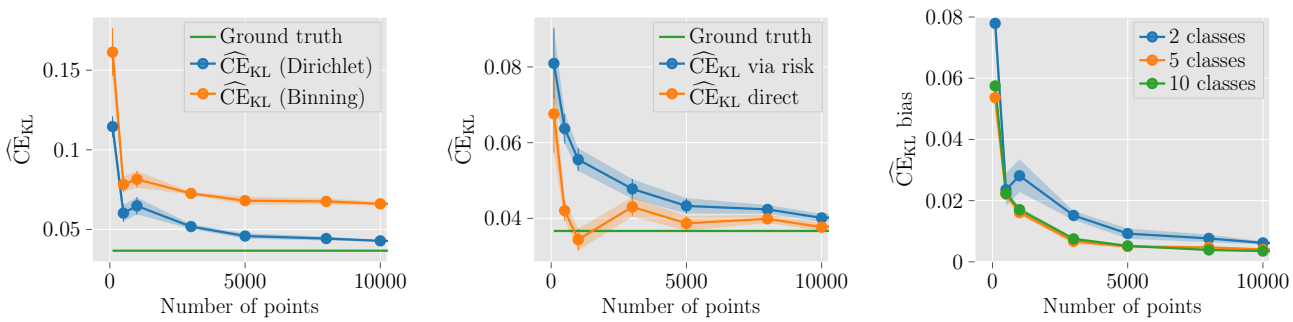
7.1 Empirical properties

To analyze the empirical properties of the proposed estimator, we create synthetic data with miscalibrated scores, for which the ground truth CE is known, following Popordanoska et al. (2022). First, we sample uniform points from the simplex and we apply temperature scaling with $t_1 = 0.9$ to ensure that the scores are closer to the boundaries of the simplex. Then, we generate ground truth labels based on the sampled probabilities, resulting in a perfectly calibrated classifier. Finally, to intentionally introduce miscalibration, we apply an additional temperature scaling with $t_2 = 0.6$.

In Figure 2a we compare the performance of two proposed choices of kernel, k_{Dir} and k_{bin}, for increasing number of samples used for the estimation. We observe that the Dirichlet-based estimator not only has better properties, like differentiability, consistency and asymptotic unbiasedness, but also has better empirical performance. Figure 2b compares the direct implementation of the estimator, as given in Equation (9), with the estimator derived via the risk, given in Equation (14). We derived theoretically the statistical properties of both estimators, while empirically the direct implementation (orange curve) performs better than the estimator derived via risk (blue curve). Finally, Figure 2c shows the convergence of the bias of $\widehat{\mathrm{CE}}_{\mathrm{KL}}$ as a function of the number of points used for the estimation. We observe that regardless of the number of classes, the estimator consistently provides reliable estimates of class-wise calibration error.

7.2 Post-hoc calibration

Here we demonstrate the application of our proposed estimator for evaluating CE on CIFAR10/100 (Krizhevsky & Hinton, 2009), after performing post-hoc calibration. Following standard practice, we trained various PreResNet (He et al., 2016),



(a) Dirichlet vs. Binning (b) Direct vs. estimation via risk (c) Bias convergence

Figure 2: (a) A comparison of the binning-based estimator and the KDE-based estimator with Dirichelt kernel. (b) A comparison of the CE estimator using Equation (14) and direct estimation via Equation (9). (c) Convergence of the bias as a function of the number of points used for the estimation.

Table 2: Performance evaluation ($\widehat{\mathrm{CE}}_{\mathrm{KL}} \times 100$ and $\widehat{\mathrm{CE}}_2^2 \times 100$, lower is better) of various network architectures on CIFAR-10/100 with no calibration, and after recalibrating with IR and TS. The number in the bracket represents the change of CE (in %) relative to the uncalibrated score. The results are averaged over 5 seeds.

		No calibration		**Isotonic regression**		**Temperature scaling**	
Dataset	Model	$\widehat{\mathrm{CE}}_{\mathrm{KL}}$	$\widehat{\mathrm{CE}}_2^2$	$\widehat{\mathrm{CE}}_{\mathrm{KL}}$	$\widehat{\mathrm{CE}}_2^2$	$\widehat{\mathrm{CE}}_{\mathrm{KL}}$	$\widehat{\mathrm{CE}}_2^2$
CIFAR-10	PreResNet20	$2.74_{\pm 0.06}$	$1.12_{\pm 0.01}$	$1.57_{\pm 0.05}(\downarrow 43\%)$	$0.58_{\pm 0.01}(\downarrow 48\%)$	$1.26_{\pm 0.02}(\downarrow 54\%)$	$0.66_{\pm 0.01}(\downarrow 41\%)$
	PreResNet56	$1.94_{\pm 0.04}$	$0.84_{\pm 0.03}$	$1.30_{\pm 0.03}(\downarrow 33\%)$	$0.50_{\pm 0.03}(\downarrow 41\%)$	$1.03_{\pm 0.02}(\downarrow 47\%)$	$0.54_{\pm 0.02}(\downarrow 36\%)$
	PreResNet110	$1.76_{\pm 0.03}$	$0.77_{\pm 0.01}$	$1.26_{\pm 0.03}(\downarrow 29\%)$	$0.49_{\pm 0.01}(\downarrow 37\%)$	$0.98_{\pm 0.01}(\downarrow 44\%)$	$0.51_{\pm 0.01}(\downarrow 33\%)$
	PreResNet164	$1.58_{\pm 0.02}$	$0.70_{\pm 0.01}$	$1.20_{\pm 0.02}(\downarrow 24\%)$	$0.43_{\pm 0.02}(\downarrow 38\%)$	$0.91_{\pm 0.02}(\downarrow 42\%)$	$0.47_{\pm 0.01}(\downarrow 32\%)$
	VGG16BN	$2.85_{\pm 0.05}$	$1.00_{\pm 0.02}$	$1.36_{\pm 0.03}(\downarrow 52\%)$	$0.57_{\pm 0.01}(\downarrow 42\%)$	$1.38_{\pm 0.03}(\downarrow 52\%)$	$0.80_{\pm 0.02}(\downarrow 20\%)$
	WideResNet28x10	$1.21_{\pm 0.02}$	$0.61_{\pm 0.01}$	$1.09_{\pm 0.03}(\downarrow 10\%)$	$0.41_{\pm 0.01}(\downarrow 34\%)$	$0.93_{\pm 0.01}(\downarrow 23\%)$	$0.49_{\pm 0.01}(\downarrow 19\%)$
CIFAR-100	PreResNet20	$0.76_{\pm 0.01}$	$0.38_{\pm 0.00}$	$0.96_{\pm 0.02}(\uparrow 26\%)$	$0.35_{\pm 0.00}(\downarrow 9\%)$	$0.70_{\pm 0.00}(\downarrow 8\%)$	$0.35_{\pm 0.00}(\downarrow 8\%)$
	PreResNet56	$0.78_{\pm 0.02}$	$0.35_{\pm 0.01}$	$0.88_{\pm 0.01}(\uparrow 13\%)$	$0.28_{\pm 0.00}(\downarrow 21\%)$	$0.61_{\pm 0.01}(\downarrow 22\%)$	$0.30_{\pm 0.00}(\downarrow 14\%)$
	PreResNet110	$0.76_{\pm 0.01}$	$0.34_{\pm 0.00}$	$0.85_{\pm 0.01}(\uparrow 12\%)$	$0.27_{\pm 0.00}(\downarrow 20\%)$	$0.60_{\pm 0.00}(\downarrow 21\%)$	$0.30_{\pm 0.00}(\downarrow 12\%)$
	PreResNet164	$0.74_{\pm 0.00}$	$0.33_{\pm 0.00}$	$0.86_{\pm 0.01}(\uparrow 16\%)$	$0.26_{\pm 0.00}(\downarrow 21\%)$	$0.59_{\pm 0.01}(\downarrow 20\%)$	$0.29_{\pm 0.00}(\downarrow 11\%)$
	VGG16BN	$1.23_{\pm 0.01}$	$0.43_{\pm 0.00}$	$0.95_{\pm 0.01}(\downarrow 23\%)$	$0.34_{\pm 0.00}(\downarrow 21\%)$	$0.75_{\pm 0.00}(\downarrow 39\%)$	$0.38_{\pm 0.00}(\downarrow 11\%)$
	WideResNet28x10	$0.62_{\pm 0.01}$	$0.30_{\pm 0.00}$	$0.72_{\pm 0.02}(\uparrow 15\%)$	$0.22_{\pm 0.00}(\downarrow 27\%)$	$0.60_{\pm 0.01}(\downarrow 3\%)$	$0.30_{\pm 0.00}(\downarrow 1\%)$

Table 3: Accuracy on CIFAR-10.

Model	No calibration	Isotonic regression
PreResNet20	$91.95_{\pm 0.05}$	$91.94_{\pm 0.07}$
PreResNet56	$94.38_{\pm 0.13}$	$94.34_{\pm 0.13}$
PreResNet110	$94.86_{\pm 0.04}$	$94.83_{\pm 0.05}$
PreResNet164	$95.24_{\pm 0.05}$	$95.14_{\pm 0.06}$
VGG16BN	$93.26_{\pm 0.04}$	$93.23_{\pm 0.04}$
WideResNet28x10	$95.54_{\pm 0.05}$	$95.53_{\pm 0.04}$

VGG16 (Simonyan & Zisserman, 2014) and WideResNet (Zagoruyko & Komodakis, 2016) architectures. Details about the training can be found in Appendix F.

In Table 2 we present a comparison of $\widehat{\mathrm{CE}}_{\mathrm{KL}}$ and $\widehat{\mathrm{CE}}_2^2$ for models before and after calibration with temperature scaling (TS) and isotonic regression (IR) (Guo et al., 2017). We focus on these methods because of their distinct optimization objectives during calibration: TS minimizes the NLL loss, while IR aims to optimize a weighted Brier score. It is notable that we obtain a much better $\widehat{\mathrm{CE}}_{\mathrm{KL}}$ score with TS across all architectures on both datasets. For instance, we report 42% improvement from the uncalibrated score using TS, compared with 24% decrease in CE using IR for PreResNet164 on CIFAR-10. The effect is opposite for $\widehat{\mathrm{CE}}_2^2$: IR is a better suited calibration technique if one aims to minimize this metric. For example, calibration with IR on VGG16BN results in 42% decrease in CE, compared to only 20% decrease with TS. On the other hand, if the goal is to optimize $\widehat{\mathrm{CE}}_{\mathrm{KL}}$, the results on CIFAR-100 indicate that IR may even harm this metric. Table 3 summarizes the accuracy on CIFAR-10 before and after calibration with IR. We notice that while TS is known to be accuracy-perserving, IR also retains accuracy in practice for this setting. In summary, these findings suggest that the choice of calibration method should be influenced by the specific calibration error of interest, i.e., IR is more suitable for minimizing $\widehat{\mathrm{CE}}_2^2$, whereas TS should be preferred for $\widehat{\mathrm{CE}}_{\mathrm{KL}}$.

The full table evaluating accuracy, NLL and Brier score is in Appendix F. Additional experiments involving convergence of bias, monitoring CE and sharpness during training, performing model selection, and measuring class-wise CE are also discussed in that Appendix.

7.3 Information monotonicity

We illustrate the information monotonicity in neural networks via our refinement estimator. We can apply our refinement estimator also to the sharpness of random vectors not located in the simplex. For this, note that $\mathbb{E}\left[f\left(\mathbb{E}\left[Y \mid X\right]\right)\right] = \mathbb{E}\left[f\left(\mathbb{E}\left[Y \mid g\left(X\right)\right]\right)\right]$ for any function f and injective function g (Gruber & Buettner, 2022, Appendix D.8). In our case, f is the convex function of a refinement term and g is chosen to be an invertible version of the softmax function and X are the activations of an intermediate layer.

We train a fully connected neural network on MNIST via stochastic gradient descent. The model has four hidden layers with nine nodes each. In Figure 3, we show that the model accuracy and L_2 sharpness of each intermediate layer throughout training. Note that sharpness can also be interpreted as mutual information. At the beginning of training, the information in each layer increases sharply similar to the accuracy. But, the initial increase of information holds no longer for layers one and two, while layers three and four experience a slow-down in information gain. This information gap is then reduced for longer periods of training. We conclude that the information in each layer is not optimized uniformly throughout training, which can be monitored via our refinement estimator.

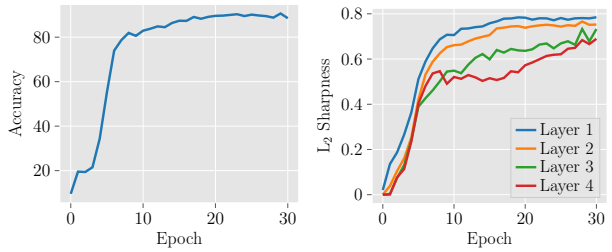


Figure 3: Accuracy and information monotonicity throughout training via our sharpness estimator (here: L_2). Each consecutive layer's sharpness is lower bounded by the next layer. The sharpness of the whole model is optimized implicitly via a proper score.

8 DISCUSSION AND CONCLUSION

In this paper, we proposed a consistent and differentiable estimator for all proper calibration errors. The estimator is asymptotically unbiased, has a convergence rate of $\mathcal{O}(n^{-1/2})$, and the bias decreases at a rate of $\mathcal{O}(n^{-1})$. Specifically, we introduce the Kullback–Leibler CE as a theoretically justified choice in standard neural network training procedures that utilize log loss. Furthermore, we showed that model sharpness, a generalization of mutual information, is equal to a multi-distribution f-divergence. Via this relation, we proved that information monotonicity in neural networks is a general concept beyond log minimization and can be monitored via sharpness during model training.

Our work has several limitations. It is an open research question to find bias corrections for proper CE estimators. Although Popordanoska et al. (2022) provided a debiasing strategy for $\widehat{\mathrm{CE}}_2^2$, it does not transfer directly to our more general case. Further, an intrinsic problem of density estimators for CE is the $\mathcal{O}(n^2)$ complexity. In our experiments, we demonstrate that evaluation can be effectively computed on subsets of reasonable size. However, assessing larger sets becomes computationally expensive. Improved debiasing could make block estimators feasible (Zaremba et al., 2013), leading also to improved computational properties.

In summary, we show that there exist asymptotically unbiased estimators for an entire landscape of proper CEs. The experimental results demonstrate the empirical behavior of the proposed approach and showcase its properties in assessing CE and sharpness. This makes it a valuable component for designing calibration methods which aim to minimize a specific CE.

Acknowledgements

This research received funding from the Flemish Government (AI Research Program) and the Research Foundation - Flanders (FWO) through project number G0G2921N. The publication was also supported by funding from the Academy of Finland (Profi6 336449 funding program), the University of Oulu strategic funds, the Wellbeing Services County of North Ostrobothnia (VTR project K62716), Terttu Foundation, and the Finnish Foundation for Cardiovascular Research. The authors wish to acknowledge CSC – IT Center for Science, Finland, for generous computational resources.
Co-funded by the European Union (ERC, TAIPO, 101088594 to FB). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020.

Bialek, W., Nemenman, I., and Tishby, N. Predictability, complexity, and learning. *Neural computation*, 13(11): 2409–2463, 2001.

Blattenberger, G. and Lad, F. Separating the brier score into calibration and refinement components: A graphical exposition. *The American Statistician*, 39(1):26–32, 1985.

Bregman, L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 – 217, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90040-7.

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.

Chechik, G., Globerson, A., Tishby, N., and Weiss, Y. Information bottleneck for Gaussian variables. *Advances in Neural Information Processing Systems*, 16, 2003.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

Csiszár, I. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4): 191–213, 1972.

DeGroot, M. H. and Fienberg, S. E. Assessing probability assessors: Calibration and refinement. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS, 1981.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22, 1983. ISSN 00390526, 14679884.

Dimitriadis, T., Gneiting, T., and Jordan, A. I. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8):e2016191118, 2021. doi: 10.1073/pnas.2016191118. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2016191118`.

Duchi, J., Khosravi, K., and Ruan, F. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018.

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.

Garcia-Garcia, D. and Williamson, R. C. Divergences and risks for multiclass experiments. In *Conference on Learning Theory*, pp. 28–1. JMLR Workshop and Conference Proceedings, 2012.

Gilad-Bachrach, R., Navot, A., and Tishby, N. An information theoretic tradeoff between complexity and accuracy. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pp. 595–609. Springer, 2003.

Gneiting, T. and Raftery, A. E. Weather forecasting with ensemble methods. *Science*, 310:248 – 249, 2005.

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.

Goldfeld, Z., Van Den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., and Polyanskiy, Y. Estimating information flow in deep neural networks. In *36th International Conference on Machine Learning, ICML 2019*, pp. 4153–4162. International Machine Learning Society (IMLS), 2019.

Gruber, S. G. and Buettner, F. Better uncertainty calibration via proper scores for classification and beyond. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

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.

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

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.

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.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 630–645. Springer, 2016.

Huszar, F. *Scoring rules, divergences and information in Bayesian machine learning.* PhD thesis, University of Cambridge, 2013.

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.

Janowczyk, A. and Madabhushi, A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7: 29, 07 2016. doi: 10.4103/2153-3539.186902.

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

Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., et al. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379 (2194):20200093, 2021.

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.

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

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

Kuleshov, V. and Deshpande, S. Calibrated and sharp uncertainties in deep learning via density estimation. In *International Conference on Machine Learning*, pp. 11683–11693. PMLR, 2022.

Kull, M. and Flach, P. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15*, pp. 68–85. Springer, 2015.

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, 2019.

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.

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.

Liese, F. and Vajda, I. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

Ma, X. and Blaschko, M. B. Meta-cal: Well-controlled post-hoc calibration by ranking. In *International Conference on Machine Learning*, 2021.

MacKay, D. J. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Magnus, J. R. and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, second edition, 1999. ISBN 0471986321 9780471986324 047198633X 9780471986331.

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

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.

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

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.

Murphy, A. H. and Winkler, R. L. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(1):41–47, 1977. ISSN 00359254, 14679876.

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*, pp. 2901–2907, 2015.

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.

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

Nobel, A. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3): 1084–1105, 1996.

Ouimet, F. and Tolosana-Delgado, R. Asymptotic properties of Dirichlet kernel density estimators. *Journal of Multivariate Analysis*, 187:104832, 2022.

Ovcharov, E. Y. Existence and uniqueness of proper scoring rules. *J. Mach. Learn. Res.*, 16:2207–2230, 2015.

Ovcharov, E. Y. Proper scoring rules and Bregman divergence. *Bernoulli*, 24(1):53 – 79, 2018. doi: 10.3150/16-BEJ857.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, 2019.

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.

Popordanoska, T., Sayer, R., and Blaschko, M. B. A consistent and differentiable lp canonical calibration error estimator. In Oh, A. H., Agarwal, A., Belgrave, D., and

Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

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.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4036–4054. PMLR, 2022.

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., and Cox, D. D. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*, 2018.

Scott, A. and Wu, C.-F. On the asymptotic distribution of ratio and regression estimators. *Journal of the American Statistical Association*, 76(373):98–102, 1981. ISSN 01621459. URL `http://www.jstor.org/stable/2287051`.

Shamir, O., Sabato, S., and Tishby, N. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Slonim, N. and Tishby, N. Agglomerative information bottleneck. *Advances in neural information processing systems*, 12, 1999.

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

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.

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.

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

Wang, Z., Huang, S.-L., Kuruoglu, E. E., Sun, J., Chen, X., and Zheng, Y. PAC-bayes information bottleneck. In *International Conference on Learning Representations*, 2022.

Wenger, J., Kjellström, H., and Triebel, R. Non-parametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 178–190, 2020.

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.

Williamson, R. C. The geometry of losses. In *Conference on Learning Theory*, pp. 1078–1108. PMLR, 2014.

Wolter, K. M. *Introduction to Variance Estimation*. Springer, 2007.

Wu, T. and Fischer, I. Phase transitions for the information bottleneck in representation learning. In *International Conference on Learning Representations*, 2020.

Wu, T., Fischer, I., Chuang, I. L., and Tegmark, M. Learnability for the information bottleneck. In *Uncertainty in Artificial Intelligence*, pp. 1050–1060. PMLR, 2020.

Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

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

Zaremba, W., Gretton, A., and Blaschko, M. B-tests: Low variance kernel two-sample tests. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26, 2013.

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 models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, we included a GitHub link.]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A UNCERTAINTY CALIBRATION IN CLASSIFICATION

In this section, we give a more detailed overview of calibration errors compared to the main paper. To introduce calibration formally, we consider a classifier $g\colon \mathscr{X} \to \Delta^k$, where $\Delta^k \coloneqq \left\{(p_1, \dots, p_k)^\intercal \in [0,1]^k \mid \sum_{i=1}^k p_i = 1\right\}$ is a probability simplex with k vertices, and $\mathscr{X}$ is a feature space with feature variable X. We denote class labels as one-hot encoded variables of $\mathscr{Y} = \{1, \dots, k\}$, i.e. the target variable Y has observations $y \in \{0,1\}^k \subset \Delta^k$ with $\|y\|_1 = 1$. In the literature, there exist multiple notions of calibration of different strength (Vaicenavicius et al., 2019; Kull et al., 2019). Calibration errors assess the degree of violation of a respective notion.

The strongest notion of calibration is the canonical calibration, which compares the probability vector prediction $g(X)$ with the target distribution $\mathbb{E}[Y \mid g(X)]$ given this prediction. A class of canonical calibration errors, which was studied recently in several works (Naeini et al., 2015; Kumar et al., 2019; Wenger et al., 2020; Popordanoska et al., 2022; Gruber & Buettner, 2022), is referred to as L_p calibration error and defined as

$$\mathrm{CE}_p(f) = \left(\mathbb{E}\left[\|\mathbb{E}[Y \mid g(X)] - g(X)\|_p^p\right]\right)^{\frac{1}{p}}. \tag{23}$$

The special case $\widehat{\mathrm{CE}}_2^2$, as given in Equation (7), can be derived from the Brier score (Murphy, 1973).

Canonical calibration errors are notoriously difficult to estimate and represent a calibration strictness which may not be necessary in practice (Vaicenavicius et al., 2019). One common and less constraining notion is class-wise calibration, which compares the individual prediction $g_i(X)$ of class $i \in \mathscr{Y}$ with the target class distribution $\mathbb{P}(Y = i \mid g_i(X))$ given the individual class prediction. The class-wise calibration error with respect to a given L_p space is defined as

$$\mathrm{CWCE}_p(g) = \left(\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\left(\mathbb{P}(Y = i \mid g_i(X)) - g_i(X)\right)^p\right]\right)^{\frac{1}{p}}. \tag{24}$$

The definition is formalized based on Kumar et al. (2019) and Gruber & Buettner (2022), while the class-wise concept was introduced independently by Kull et al. (2019) and Nixon et al. (2019).

While class-wise calibration is easier to evaluate than canonical calibration, it still scales linearly in complexity with the number of classes, which can be problematic for tasks with a very high number of classes. In contrast, the most common approach in the machine learning literature is top-label confidence calibration (Naeini et al., 2015; Guo et al., 2017; Joo et al., 2020; Kristiadi et al., 2020; Rahimi et al., 2020; Tomani et al., 2021; Minderer et al., 2021; Tian et al., 2021; Islam et al., 2021; Menon et al., 2021; Morales-Álvarez et al., 2021; Gupta et al., 2021; Wang et al., 2021; Fan et al., 2022), which is not affected by this issue. In this notion, we compare if the predicted top-label confidence $\max_{i\in\mathscr{Y}} g_i(X)$ matches the conditional accuracy $\mathbb{P}(Y = \arg\max_{i\in\mathscr{Y}} g_i(X) \mid \max_{i\in\mathscr{Y}} g_i(X))$ given the predicted top-label confidence. It is known in the literature that top-label confidence calibration represents the weakest notion of calibration (Vaicenavicius et al., 2019; Widmann et al., 2019; Gruber & Buettner, 2022). The top-label confidence calibration error based on an L_p space is defined as (Kumar et al., 2019; Gruber & Buettner, 2022)

$$\mathrm{TCE}_p(g) = \left(\mathbb{E}\left[\left(\mathbb{P}\left(Y = \arg\max_{j\in\mathscr{Y}} g_j(X) \mid \max_{j\in\mathscr{Y}} g_j(X)\right) - \max_{j\in\mathscr{Y}} g_j(X)\right)^p\right]\right)^{\frac{1}{p}}. \tag{25}$$

For $p = 1$, its binning based estimator is commonly referred to as expected calibration error (Naeini et al., 2015; Guo et al., 2017).

Besides the L_p based calibration errors presented above, there also exist other calibration errors, like maximum mean calibration error (Kumar et al., 2018), Kolmogorov-Smirnov calibration error (Gupta et al., 2020), and kernel calibration error (Widmann et al., 2019). We exclude these errors from our analysis and experiments as they are not related to risk minimization.

B RATE OF BIAS OF PROPER CALIBRATION ERROR ESTIMATION

We have already shown the existence of a consistent and asymptotically unbiased estimator via the refinement in Section 5.2. We now show that direct estimation via the Bregman divergence definition of a proper calibration error via Equation (9) also yields the same asymptotic rates.

$$\widehat{\mathrm{CE}}_F(g) := \frac{1}{n}\sum_{h=1}^{n}\left(F\left(\frac{\sum_{j\neq h}k(g(x_h),g(x_j))y_j}{\sum_{j\neq h}k(g(x_h),g(x_j))}\right) - F(g(x_h)) - \left\langle \nabla F\left(g(x_h)\right), \frac{\sum_{j\neq h}k(g(x_h),g(x_j))y_j}{\sum_{j\neq h}k(g(x_h),g(x_j))} - g(x_h)\right\rangle\right) \tag{26}$$

$$= -\widehat{\mathrm{REF}}_F(g) - \frac{1}{n}\sum_{h=1}^{n}\left(F(g(x_h)) + \left\langle \nabla F\left(g(x_h)\right), \frac{\sum_{j\neq h}k(g(x_h),g(x_j))y_j}{\sum_{j\neq h}k(g(x_h),g(x_j))}\right\rangle - \left\langle \nabla F\left(g(x_h)\right), g(x_h)\right\rangle\right). \tag{27}$$

We have already shown the empirical refinement estimator to have bias that decreases as $\mathscr{O}(n^{-1})$ in Section 5.2, and the sums over $F(g(x_h))$ and $\langle \nabla F\left(g(x_h)\right), g(x_h)\rangle$ are unbiased. We therefore focus on the term

$$\sum_{h=1}^{n}\left\langle \nabla F\left(g(x_h)\right), \frac{\sum_{j\neq h}k(g(x_h),g(x_j))y_j}{\sum_{j\neq h}k(g(x_h),g(x_j))}\right\rangle. \tag{28}$$

The second argument to the inner product is a ratio estimator with asymptotic Gaussian distribution (Scott & Wu, 1981), which is known to have bias that converges as $\mathscr{O}(n^{-1})$.

$$\mathbb{E}\left[\left\langle \nabla F\left(g(x_h)\right), \frac{\sum_{j\neq h}k(g(x_h),g(x_j))y_j}{\sum_{j\neq h}k(g(x_h),g(x_j))}\right\rangle\right] = \left\langle \mathbb{E}\left[\nabla F\left(g(x_h)\right)\right], \mathbb{E}\left[\frac{\sum_{j\neq h}k(g(x_h),g(x_j))y_j}{\sum_{j\neq h}k(g(x_h),g(x_j))}\right]\right\rangle \tag{29}$$

for each h, as the arguments to the inner product are independent of each other. We consequently conclude that the rate of bias of the inner product is the same as the rate of bias of the ratio estimator, as the first argument is unbiased. Therefore, the overall rate of bias for direct estimation of arbitrary proper calibration errors using Equation (9) is also $\mathscr{O}(n^{-1})$.

C PROOFS FOR INFORMATION MONOTONICITY

In this section, we provide detailed proofs about the f-divergence representation of the model sharpness and the information monotonicity in neural networks. Since we will require the target variable Y in categorical encoding and one-hot encoding, we will write Y for the former and $\vec{Y}$ for the latter. This notation is exclusive to this section.

C.1 Proof of Proposition 6.1

Let $F\colon \Delta^k \to \mathbb{R}$ be a convex function and $g\colon \mathscr{X} \to \Delta^k$ a classifier with prediction distributions $P_y \coloneqq \mathbb{P}\left(g\left(X\right) \mid Y = y\right)$, and $P \coloneqq \mathbb{P}\left(g\left(X\right)\right)$. Then, the model sharpness can be represented as an f-divergence

via

$$
\begin{aligned}
&\mathrm{SHARP}_F\left(g\right)\\
&= \mathbb{E}\left[D_F\left(\mathbb{E}\left[\vec{Y} \mid g\left(X\right)\right], \mathbb{E}\left[\vec{Y}\right]\right)\right]\\
&\overset{\text{def}}{=} \mathbb{E}\left[F\left(\mathbb{E}\left[\vec{Y} \mid g\left(X\right)\right]\right) - F\left(\mathbb{E}\left[\vec{Y}\right]\right) - \left\langle \nabla F\left(\mathbb{E}\left[\vec{Y}\right]\right), \mathbb{E}\left[\vec{Y}\right] - \mathbb{E}\left[\vec{Y} \mid g\left(X\right)\right]\right\rangle\right]\\
&= \mathbb{E}\left[F\left(\mathbb{E}\left[\vec{Y} \mid g\left(X\right)\right]\right)\right] - F\left(\mathbb{E}\left[\vec{Y}\right]\right)\\
&= \int_{\mathscr{X}} F\left(\mathbb{E}\left[\vec{Y} \mid g\left(X\right)\right]\right) - F\left(\mathbb{E}\left[\vec{Y}\right]\right) \mathrm{d}\mathbb{P}\left(g\left(X\right)\right)\\
&= \int_{\mathscr{X}} F\left(\mathbb{E}\left[\vec{Y}_1\right] \frac{\mathrm{d}\mathbb{P}\left(g\left(X\right) \mid Y = 1\right)}{\mathrm{d}\mathbb{P}\left(g\left(X\right)\right)}, \ldots, \mathbb{E}\left[\vec{Y}_k\right] \frac{\mathrm{d}\mathbb{P}\left(g\left(X\right) \mid Y = k\right)}{\mathrm{d}\mathbb{P}\left(g\left(X\right)\right)}\right)\\
&\quad - F\left(\mathbb{E}\left[\vec{Y}\right]\right) \mathrm{d}\mathbb{P}\left(g\left(X\right)\right)\\
&= \int_{\mathscr{X}} F\left(\mathbb{E}\left[\vec{Y}_1\right] \frac{\mathrm{d}P_1}{\mathrm{d}P}, \ldots, \mathbb{E}\left[\vec{Y}_k\right] \frac{\mathrm{d}P_k}{\mathrm{d}P}\right) - F\left(\mathbb{E}\left[\vec{Y}\right]\right) \mathrm{d}P\\
&= I_{F^Y}\left(P_1, \ldots, P_k \parallel P\right)
\end{aligned}
\tag{30}
$$

where $F^Y\left(x\right) \coloneqq F\left(\mathbb{E}\left[\vec{Y}_1\right] x_1, \ldots, \mathbb{E}\left[\vec{Y}_k\right] x_k\right) - F\left(\mathbb{E}\left[\vec{Y}_1\right], \ldots, \mathbb{E}\left[\vec{Y}_k\right]\right)$ and by using Bayes' theorem. The function I_{F^Y} is a multi-distribution f-divergence since F^Y is convex (follows from F being convex) and $F^Y\left(1, \ldots, 1\right) = F\left(\mathbb{E}\left[\vec{Y}_1\right], \ldots, \mathbb{E}\left[\vec{Y}_k\right]\right) - F\left(\mathbb{E}\left[\vec{Y}_1\right], \ldots, \mathbb{E}\left[\vec{Y}_k\right]\right) = 0$.

C.2 Proof of Theorem 6.2

For a neural network $g\left(X\right) = h_l\left(\ldots\left(h_1\left(X\right)\right)\right)$ with layers h_i, $i \in \{1, \ldots, l\}$, conditional distributions $P_y^i \coloneqq \mathbb{P}\left(h_i\left(\ldots\left(h_1\left(X\right)\right)\right) \mid Y = y\right)$, and marginal distributions $P^i \coloneqq \mathbb{E}\left[P_Y^i\right]$, we have

$$
\begin{aligned}
\mathrm{SHARP}_F\left(g\right) = I_{F^Y}\left(P_1^l, \ldots, P_k^l \parallel P^l\right) &\leq I_{F^Y}\left(P_1^{l-1}, \ldots, P_k^{l-1} \parallel P^{l-1}\right)\\
&\leq \cdots \leq I_{F^Y}\left(P_1^1, \ldots, P_k^1 \parallel P^1\right) \leq \mathrm{SI}_F\left(X; Y\right).
\end{aligned}
\tag{31}
$$

Proof. Since a neural network in our context is simply a chain of function, it is sufficient to prove that the inequality holds for a single arbitrary function f transforming a sample space Ω with σ-field $\mathscr{F}$. For this, assume we are in the context of a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and measurable space $(\Omega_f, \mathscr{F}_f)$ such that $f\colon \Omega \to \Omega_f$ is a measurable function. Define the function $M_f\colon \Omega \times \mathscr{F}_f \to \mathbb{R}$ via $M_f\left(\omega, A\right) = \mathbf{1}_{\{\omega \in \Omega \mid f(\omega) \in A\}}\left(\omega\right)$, where $\mathbf{1}$ is the indicator function. Similar as Garcia-Garcia & Williamson (2012), we write $M_f\mathbb{P}\left(A\right) = \int_\Omega M_f\left(\omega, A\right) \mathrm{d}\mathbb{P}\left(\omega\right)$. Now, we simply have to show that M_f is a Markov kernel, which then proves the statement via the information monotonicity of f-divergences.
For $A \in \mathscr{F}_f$ we have

$$
\begin{aligned}
M_f\mathbb{P}\left(A\right) &= \int_\Omega M_f\left(\omega, A\right) \mathrm{d}\mathbb{P}\left(\omega\right)\\
&= \int_\Omega \mathbf{1}_{\{\omega \in \Omega \mid f(\omega) \in A\}}\left(\omega\right) \mathrm{d}\mathbb{P}\left(\omega\right)\\
&= \int_{\{\omega \in \Omega \mid f(\omega) \in A\}} \mathrm{d}\mathbb{P}\\
&= \mathbb{P}\left(\{\omega \in \Omega \mid f\left(\omega\right) \in A\}\right).
\end{aligned}
\tag{32}
$$

The last line shows that $M_f\mathbb{P}$ is a distribution, which fulfills the definition of a Markov kernel as given in (Garcia-Garcia & Williamson, 2012).
Now, using the information monotonicity of f-divergences, we get for $i \in \{2, \ldots, l\}$

$$
\begin{aligned}
I_{F^Y}\left(P_1^i, \ldots, P_k^i \parallel P^i\right) &= I_{F^Y}\left(M_{g_i}P_1^{i-1}, \ldots, M_{g_i}P_k^{i-1} \parallel M_{g_i}P^{i-1}\right)\\
&\leq I_{F^Y}\left(P_1^{i-1}, \ldots, P_k^{i-1} \parallel P^{i-1}\right),
\end{aligned}
\tag{33}
$$

which proves the inequality chain. It is upper bounded by the statistical information $\mathrm{SI}_F(X;Y) \coloneqq \mathbb{E}\left[F\left(\mathbb{E}\left[\vec{Y} \mid X\right]\right)\right] - F\left(\mathbb{E}\left[\vec{Y}\right]\right)$ since

$$\begin{aligned}
&I_{F^Y}\left(P_1^1, \ldots, P_k^1 \parallel P^1\right) \\
&= I_{F^Y}\left(M_{g_1}\mathbb{P}\left(X \mid Y=1\right), \ldots, M_{g_1}\mathbb{P}\left(X \mid Y=k\right) \parallel M_{g_1}\mathbb{P}\left(X\right)\right) \\
&\leq I_{F^Y}\left(\mathbb{P}\left(X \mid Y=1\right), \ldots, \mathbb{P}\left(X \mid Y=k\right) \parallel \mathbb{P}\left(X\right)\right) \\
&= \int_{\mathscr{X}} F\left(\mathbb{E}\left[\vec{Y}_1\right]\frac{\mathrm{d}\mathbb{P}\left(X \mid Y=1\right)}{\mathrm{d}\mathbb{P}\left(X\right)}, \ldots, \mathbb{E}\left[\vec{Y}_k\right]\frac{\mathrm{d}\mathbb{P}\left(X \mid Y=k\right)}{\mathrm{d}\mathbb{P}\left(X\right)}\right) - F\left(\mathbb{E}\left[\vec{Y}\right]\right)\mathrm{d}\mathbb{P}\left(X\right) \\
&= \mathbb{E}\left[F\left(\mathbb{E}\left[\vec{Y} \mid X\right]\right)\right] - F\left(\mathbb{E}\left[\vec{Y}\right]\right).
\end{aligned} \tag{34}$$

□

D CLASS-WISE CALIBRATION INDUCED BY ONE-VS-REST RISK

In this section, we derive class-wise calibration errors from One-vs-Rest risk minimization. We do so by decomposing the One-vs-Rest risk into calibration and sharpness terms analogous to the standard risk minimization case. This is a novel contribution towards better understanding of class-wise calibration errors. Specifically, this suggests to use class-wise calibration in predictive scenarios when the multi-class prediction consists of probabilities, which do not sum up to one.

For a binary loss $L\colon [0,1] \times \{0,1\} \to \mathbb{R}$, we define the risk of class i vs the rest as

$$\mathscr{R}_i\left(g\right) \coloneqq \mathbb{E}\left[L\left(g\left(X\right), \mathbf{1}\left\{Y=i\right\}\right)\right], \tag{35}$$

where $g\colon \mathscr{X} \to [0,1]$ is a binary classifier and Y takes values in $\{1, \ldots, k\}$.

In the following, we assume $-L$ is a proper score, i.e. it is minimized by the Bayes classifier.

Analogous, the negative Bayes risk for a $q \in [0,1]$ is given by

$$F\left(q\right) \coloneqq -\inf_{p \in [0,1]} \mathbb{E}_{Y \sim q}\left[L\left(p, Y\right)\right]. \tag{36}$$

Like in the canonical case, F is convex as long as $-L$ is a proper score. Consequently, D_F is a Bregman divergence. The Brier score and the squared Euclidean distance are recovered by $F\left(q\right) = q^2 + \left(1-q\right)^2$. The negative log likelihood and the Kullback–Leibler divergence are given by the case $F\left(q\right) = q\log q + \left(1-q\right)\log\left(1-q\right)$. As a novel contribution, we derive class-wise calibration errors from the mean of the class-wise One-vs-Rest risk of binary classifiers $g_1, \ldots, g_k$ as

$$\frac{1}{k}\sum_{i=1}^{k}\mathscr{R}_i\left(g_i\right) = \frac{1}{k}\sum_{i=1}^{k} -F\left(\mathbb{P}\left(Y=i\right)\right) + \mathrm{CWCE}_F\left(g_1, \ldots, g_k\right) - \mathrm{SHARP}_F\left(g_1, \ldots, g_k\right) \tag{37}$$

where we have a class-wise calibration error $\mathrm{CWCE}_F\left(g_1, \ldots, g_k\right) \coloneqq \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[D_F\left(\mathbb{P}\left(Y=i \mid g_i\left(X\right)\right), g_i\left(X\right)\right)\right]$ and a class-wise sharpness $\mathrm{SHARP}_F\left(g_1, \ldots, g_k\right) \coloneqq \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[D_F\left(\mathbb{P}\left(Y=i \mid g_i\left(X\right), \mathbb{P}\left(Y=i\right)\right)\right)\right]$. Analogous to the canonical case, we have $\mathrm{CWCE}_F\left(g_1, \ldots, g_k\right) = \mathrm{CWCE}_2^2\left(g\right)$ for $F\left(q\right) = q^2$ and $g\left(x\right) \coloneqq \left(g_1\left(x\right), \ldots, g_k\left(x\right)\right)^\intercal$. Surprisingly, the associated negative Bayes risk is not symmetric unlike other common cases (Gneiting & Raftery, 2007). Note that the factor $\frac{1}{k}$ systematically decreases the class-wise calibration error with growing k if the class-wise predictions are normalized to a probability vector (Gruber & Buettner, 2022).

Proof. We first show that the decomposition holds. Note that we implied F is differentiable, but the decomposition still holds under general conditions by replacing Bregman divergences with score divergences (Gneiting & Raftery, 2007). Since by assumption L is a negative proper score, we have $L\left(p, y\right) = -F\left(p\right) - F'\left(p\right)\left(y-p\right)$ for $p \in [0,1]$

and $y \in \{0, 1\}$ (c.f. Schervish representation (Gneiting & Raftery, 2007)).
Further, note that for $i \in \{1, \ldots, k\}$ we have

$$\begin{aligned}
&\mathscr{R}_i\left(g_i\right) \\
&\overset{\text{def}}{=} \mathbb{E}\left[L\left(g_i\left(X\right), \mathbf{1}\left\{Y = i\right\}\right)\right] \\
&= \mathbb{E}\left[-F\left(g_i\left(X\right)\right) - F'\left(g_i\left(X\right)\right)\left(\mathbf{1}\left\{Y = i\right\} - g_i\left(X\right)\right)\right] \\
&\overset{\text{LOTUS}}{=} \mathbb{E}\left[-F\left(g_i\left(X\right)\right) - F'\left(g_i\left(X\right)\right)\left(\mathbb{P}\left(Y = i \mid g_i\left(X\right)\right) - g_i\left(X\right)\right)\right] \\
&= \mathbb{E}\left[D_F\left(\mathbb{P}\left(Y = i \mid g_i\left(X\right)\right), g_i\left(X\right)\right)\right] - \mathbb{E}\left[F\left(\mathbb{E}\left(Y = i \mid g_i\left(X\right)\right)\right)\right] \\
&= \mathbb{E}\left[D_F\left(\mathbb{P}\left(Y = i \mid g_i\left(X\right)\right), g_i\left(X\right)\right)\right] - \mathbb{E}\left[D_F\left(\mathbb{P}\left(Y = i \mid g_i\left(X\right)\right), \mathbb{P}\left(Y = i\right)\right)\right] \\
&\quad - \mathbb{E}\left[F\left(\mathbb{P}\left(Y = i \mid g_i\left(X\right)\right)\right)\right],
\end{aligned} \tag{38}$$

where we used the law of the unconscious statistician (LOTUS) and the definition of Bregman divergences.
Now, we use this result to get

$$\begin{aligned}
\frac{1}{k}\sum_{i=1}^{k}\mathscr{R}_i\left(g_i\right) = \frac{1}{k}\sum_{i=1}^{k} &-F\left(\mathbb{P}\left(Y = i\right)\right) \\
&+ \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[D_F\left(\mathbb{P}\left(Y = i \mid g_i\left(X\right)\right), g_i\left(X\right)\right)\right] \\
&- \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[D_F\left(\mathbb{P}\left(Y = i \mid g_i\left(X\right)\right), \mathbb{P}\left(Y = i\right)\right)\right] \\
\overset{\text{def}}{=} \frac{1}{k}\sum_{i=1}^{k} &-F\left(\mathbb{P}\left(Y = i\right)\right) + \mathrm{CWCE}_F\left(g_1, \ldots, g_k\right) - \mathrm{SHARP}_F\left(g_1, \ldots, g_k\right).
\end{aligned} \tag{39}$$

Last, we show $\mathrm{CWCE}_F\left(g_1, \ldots, g_k\right) = \mathrm{CWCE}_2^2\left(g\right)$ for $F\left(q\right) = q^2$ and $g\left(x\right) \coloneqq \left(g_1\left(x\right), \ldots, g_k\left(x\right)\right)^\intercal$. Since $D_F\left(x, y\right) = \left(x - y\right)^2$, we have

$$\begin{aligned}
\mathrm{CWCE}_F\left(g_1, \ldots, g_k\right) &\overset{\text{def}}{=} \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[D_F\left(\mathbb{P}\left(Y = i \mid g_i\left(X\right)\right), g_i\left(X\right)\right)\right] \\
&= \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\left[\left(\mathbb{P}\left(Y = i \mid g_i\left(X\right)\right) - g_i\left(X\right)\right)^2\right] \\
&\overset{\text{def}}{=} \mathrm{CWCE}_2^2\left(g\right).
\end{aligned} \tag{40}$$

□

E PROPER CALIBRATION ERROR ESTIMATORS VIA KDE

E.1 The Dirichlet kernel in estimating $\mathbb{E}[Y \mid g(x)]$

The Dirichlet kernel is defined as:

$$k_{Dir}(g(x_h), g(x_j)) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_{jk})}{\prod_{k=1}^{K}\Gamma(\alpha_{jk})}\prod_{k=1}^{K}g(x_h)_k^{\alpha_{jk}-1} \tag{41}$$

with $\alpha_j = \frac{g(x_j)}{\gamma} + 1$, $\gamma > 0$ being a bandwidth parameter (Ouimet & Tolosana-Delgado, 2022; Popordanoska et al., 2022). We note that popular libraries such as PyTorch provide only $\log\Gamma$[2] and not Γ directly (Paszke et al.,

[2] `https://pytorch.org/docs/stable/generated/torch.lgamma.html`

2019). It will therefore be useful to consider

$$\log(k(g(x_h), g(x_j))) = \log \frac{\Gamma(\sum_{k=1}^{K} \alpha_{jk})}{\prod_{k=1}^{K} \Gamma(\alpha_{jk})} \prod_{k=1}^{K} g(x_h)_k^{\alpha_{jk}-1} \tag{42}$$

$$= \log \Gamma \left(\sum_{k=1}^{K} \alpha_{jk} \right) - \sum_{k=1}^{K} \log \Gamma (\alpha_{jk}) + \sum_{k=1}^{K} (\alpha_{jk} - 1) \log g(x_h)_k \tag{43}$$

$$= \log \Gamma \left(K + \sum_{k=1}^{K} \frac{g(x_j)_k}{\gamma} \right) - \sum_{k=1}^{K} \log \Gamma \left(\frac{g(x_j)_k}{\gamma} + 1 \right) + \frac{1}{\gamma} \langle g(x_j), \log(g(x_h)) \rangle \tag{44}$$

and we can further apply the log to the softmax function in the last $\log(g(x_h))$ inside the inner product.

We subsequently focus on terms of the form

$$\log \left(\sum_{j \neq h} k(g(x_h), g(x_j)) \right) = \text{LogSumExp}_{j \neq h} \left(\log(k(g(x_h), g(x_j))) \right). \tag{45}$$

Computation of terms of the form $\log \left(\sum_{j \neq h} k(g(x_h), g(x_j)) y_j \right)$ are essentially the same, but the LogSumExp operation should only be performed over indices where $y_{jk} \neq 0$.

E.2 Bregman derivation of the L_2 calibration error estimator

For the Bregman formulation of squared L_2 error, we have $F(x) = \|x\|_2^2$, and the r.h.s. of Equation (9) becomes

$$\frac{1}{n} \sum_{h=1}^{n} \left(\left\| \frac{\sum_{j \neq h} k(g(x_h), g(x_j)) y_j}{\sum_{j \neq h} k(g(x_h), g(x_j))} \right\|_2^2 + \|g(x_h)\|_2^2 - 2 \left\langle g(x_h), \frac{\sum_{j \neq h} k(g(x_h), g(x_j)) y_j}{\sum_{j \neq h} k(g(x_h), g(x_j))} \right\rangle \right)$$

$$= \frac{1}{n} \sum_{h=1}^{n} \left\| \frac{\sum_{j \neq h} k(g(x_h), g(x_j)) y_j}{\sum_{j \neq h} k(g(x_h), g(x_j))} - g(x_h) \right\|_2^2. \tag{46}$$

We see that this recovers exactly the L_2 special case of the estimator given by (Popordanoska et al., 2022, Equation(9)). That paper shows that the resulting estimator is consistent, and has a convergence of $\mathcal{O}(n^{-1/2})$ with a bias that converges as $\mathcal{O}(n^{-1})$.

E.3 Bregman derivation of the KL calibration error estimator

Recall from above that the Bregman KL divergence is generated by $F(p) = \langle p, \log(p) \rangle$, where the log operation is applied element-wise.

The Bregman formulation of KL calibration error is

$$\frac{1}{n} \sum_{h=1}^{n} \left(\left\langle \frac{\sum_{j \neq h} k(g(x_h), g(x_j)) y_j}{\sum_{j \neq h} k(g(x_h), g(x_j))}, \log \left(\frac{\sum_{j \neq h} k(g(x_h), g(x_j)) y_j}{\sum_{j \neq h} k(g(x_h), g(x_j))} \right) \right\rangle - \langle g(x_h), \log(g(x_h)) \rangle - \right.$$

$$\left. \left\langle \log(g(x_h)) + e, \frac{\sum_{j \neq h} k(g(x_h), g(x_j)) y_j}{\sum_{j \neq h} k(g(x_h), g(x_j))} - g(x_h) \right\rangle \right)$$

$$= \frac{1}{n} \sum_{h=1}^{n} \left\langle \frac{\sum_{j \neq h} k(g(x_h), g(x_j)) y_j}{\sum_{j \neq h} k(g(x_h), g(x_j))}, \log \left(\frac{\sum_{j \neq h} k(g(x_h), g(x_j)) y_j}{\sum_{j \neq h} k(g(x_h), g(x_j)) g(x_h)} \right) \right\rangle \tag{47}$$

where e is a vector of all ones and division of two vectors is assumed to be element-wise. The result is an estimator identical to Equation (10).

E.4 General Sharpness-Calibration error decompositions of expectations of Bregman divergences

In general, we can define a statistical risk measure based on a Bregman divergence as

$$\mathbb{E}_{(X,Y)\sim p}[D_F(Y, g(X))] = \mathbb{E}_{(X,Y)}[F(Y) - F(g(X)) - \langle \nabla F(g(X)), Y - g(X)\rangle]. \tag{48}$$

If we subtract out the associated Bregman calibration error (cf. Equation (9)), we have

$$\begin{aligned}
\mathbb{E}[D_F(Y, g(X))] - \mathrm{CE}_F(g) =& \mathbb{E}_{(X,Y)}[D_F(Y, g(X))] - \mathbb{E}_X\left[D_F(\mathbb{E}[Y \mid g(X)], g(X))\right] && (49)\\
=& \mathbb{E}[F(Y) - F(g(X)) - \langle \nabla F(g(X)), Y - g(X)\rangle] && (50)\\
& - \mathbb{E}[F(\mathbb{E}[Y \mid g(X)]) - F(g(X)) - \langle \nabla F(g(X)), \mathbb{E}[Y \mid g(X)] - g(X)\rangle]\\
=& \mathbb{E}[F(Y) - F(\mathbb{E}[Y \mid g(X)])] - \mathbb{E}[\langle \nabla F(g(X)), Y - \mathbb{E}[Y \mid g(X)]\rangle]. && (51)
\end{aligned}$$

It is a well known property of conditional expectation that $\mathbb{E}[f(Z) \cdot Y \mid Z] = f(Z)\mathbb{E}[Y \mid Z]$, which yields

$$\mathbb{E}[\langle \nabla F(g(X)), Y - \mathbb{E}[Y \mid g(X)]\rangle] = \mathbb{E}[\langle \nabla F(g(X)), Y\rangle] - \underbrace{\mathbb{E}[\mathbb{E}[\langle \nabla F(g(X)), Y\rangle \mid g(X)]]}_{=\mathbb{E}[\langle \nabla F(g(X)), Y\rangle] \text{ by law of total expectation}} = 0, \tag{52}$$

and our equation simplifies to

$$\mathbb{E}[D_F(Y, g(X))] - \mathrm{CE}_F(g) = \mathbb{E}[F(Y) - F(\mathbb{E}[Y \mid g(X)])]. \tag{53}$$

E.4.1 Recovery of L_2 refinement

Setting $F(p) = \|p\|^2$, the risk becomes the Brier score and we obtain

$$\mathbb{E}[\|Y\|^2 - \|\mathbb{E}[Y \mid g(X)]\|^2] = 1 - \mathbb{E}_X[\|\mathbb{E}[Y \mid g(X)]\|^2] \tag{54}$$

Popordanoska et al. (2022) show the L_2 refinement to be $\mathbb{E}[(1 - \mathbb{E}[Y \mid g(X))]\mathbb{E}[Y \mid g(X)]] = \mathbb{E}[\mathbb{E}[Y \mid g(X)] - \mathbb{E}[Y \mid g(X)]^2] = \mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y \mid g(X)]^2]$. This is in fact the first term of the above if we expand the norm as a sum over elements and split into expectations of terms from each dimension of Y.

E.4.2 Recovery of KL refinement

Setting $F(p) = \langle p, \log(p)\rangle$, the risk corresponds with cross-entropy loss, and we obtain

$$\begin{aligned}
\mathbb{E}[\underbrace{\langle Y, \log(Y)\rangle}_{=0} - \langle \mathbb{E}[Y \mid g(X)], \log(\mathbb{E}[Y \mid g(X)]\rangle] =& - \mathbb{E}[\langle \mathbb{E}[Y \mid g(X)], \log(\mathbb{E}[Y \mid g(X)]\rangle] && (55)\\
=& \mathbb{E}[H(\mathbb{E}[Y \mid g(X)])], && (56)
\end{aligned}$$

where H denotes entropy, we interpret the first inner product involving Y using $\lim_{Y \searrow 0} Y \log Y = 0$ as it is otherwise undefined, and we additionally assume categorical labels Y without uncertainty.

F EXPERIMENTS

In this section, we first provide a detailed description of the empirical setup. Subsequently, we further investigate the bias convergence of our estimator, both on simulated and real-world datasets. Then, we show a table evaluating accuracy, NLL and Brier score on CIFAR 10/100 before and after calibration with isotonic regression and temperature scaling. Finally, we present additional results for several use-cases of the estimator: monitoring model training, model selection, and assessing calibration error.

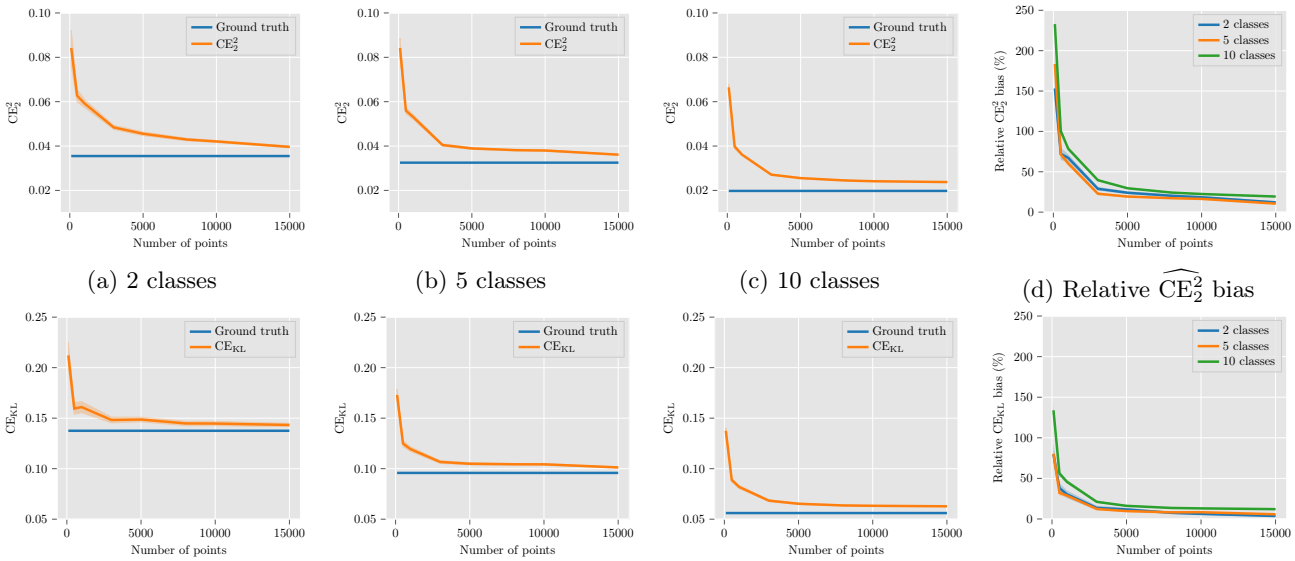
F.1 Empirical setup

Datasets The experiments in the main text rely on two widely used benchmark datasets, CIFAR-10/100 (Krizhevsky & Hinton, 2009), which consist of 32×32 natural images divided into 10 and 100 classes, respectively. We split the data into train/validation/test sets of 45000/5000/10000.

Models We trained PreResNet20, PreResNet56, PreResNet110, PreResNet164 (He et al., 2016), VGG16 (with BatchNorm) (Simonyan & Zisserman, 2014) and WideResNet28x10 (Zagoruyko & Komodakis, 2016) for 250 epochs with Stochastic Gradient Descent optimizer using PyTorch (Paszke et al., 2017). The learning rate was reduced by a factor of 10 at 150^{th} and 225^{th} epochs. The WideResNet and the PreResNet models were trained with the learning rate of 1e−1, batch size of 128, weighted decay (WD) of 1e−10 and Nesterov's momentum of 0.9. During the first epoch, we warmed up the training with the learning rate of 0.01. The VGG model was trained with the learning rate of 5e−2, WD of 5e−5, batch size of 128, and momentum of 0.05. The training was carried out with NVIDIA V100 GPU.

F.2 Convergence of bias

In addition to our empirical analysis for the convergence of bias on simulated data in the main part, here we present the calibration error and the relative bias, computed as $\frac{\widehat{\text{CE}}-\text{CE}}{\text{CE}} * 100$, with $\widehat{\text{CE}}$ the estimated and CE the ground truth calibration error. The averaged results across 20 iterations of sampling new points for the estimation, together with the standard errors, are shown in Figure 4.



(a) 2 classes (b) 5 classes (c) 10 classes (d) Relative $\widehat{\text{CE}}_2^2$ bias

(e) 2 classes (f) 5 classes (g) 10 classes (h) Relative $\widehat{\text{CE}}_{\text{KL}}$ bias

Figure 4: Calibration error and relative bias (%) on simulated data for different number of classes. Each plot shows the estimate as a function of the sample size. The top row evaluates $\widehat{\text{CE}}_2^2$, while the bottom row $\widehat{\text{CE}}_{\text{KL}}$.

Similarly, in Figure 5 we evaluate the calibration error, bias and relative bias on the test set of CIFAR-10, as a function of the number of points used for the estimation. The ground truth is calculated from the whole test set (10000 images). The bandwidth of the Dirichlet kernel is set to 0.02. The results reveal that the estimator achieves values that closely align with the ground truth, even with as few as a (couple of) hundred points.

F.3 Post-hoc calibration

In addition to the experiment in the main text, where we evaluate $\widehat{\text{CE}}_2^2$ and $\widehat{\text{CE}}_{\text{KL}}$ before and after calibration, in Table 4 we show accuracy, NLL and Brier score of various network architectures on CIFAR-10/100.

F.4 Additional experiments

Monitoring calibration error during training Monitoring accuracy and loss during the training of a deep neural network is essential for various reasons, including performance evaluation and model selection. We argue that monitoring calibration error and sharpness/refinement, in addition to the standard metrics, provides additional insights into the reliability and performance of the model.

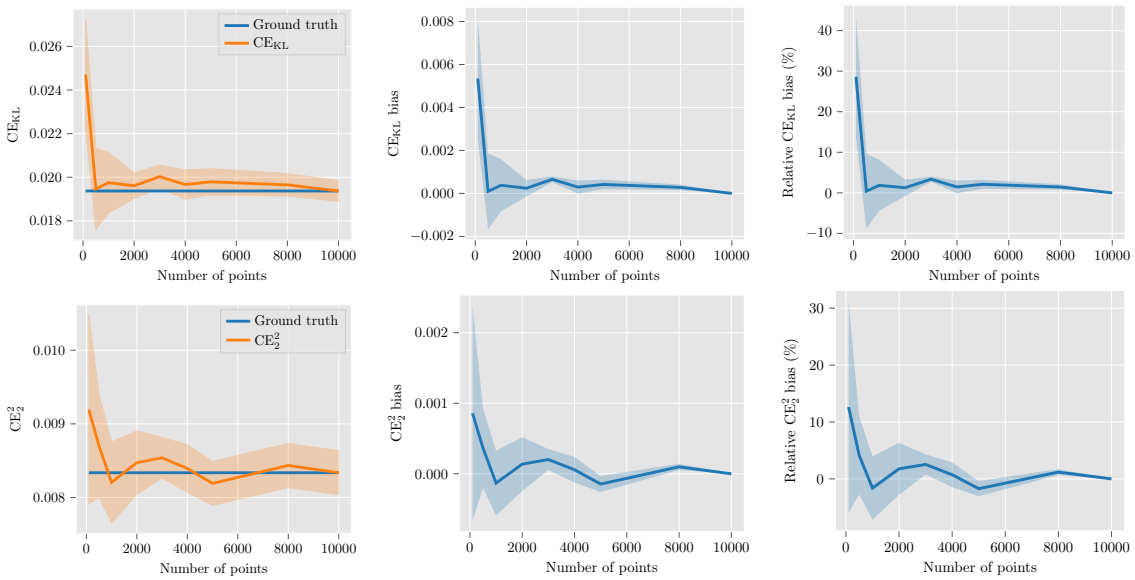


Figure 5: Calibration error, bias, and relative bias (%) evaluated on the test set of CIFAR 10, using predictions of trained PreResNet56 architectures with four different seeds. The ground truth is obtained from the whole test set.

Table 4: Performance evaluation (Acc ×100, NLL ×100 and Brier score ×100) of various network architectures on CIFAR-10/100 with no calibration, and recalibration with Isotonic Regression and Temperature Scaling.

		No calibration			**Isotonic Regression**			**Temperature Scaling**		
Dataset	Model	Acc	NLL	Brier	Acc	NLL	Brier	Acc	NLL	Brier
CIFAR-10	PreResNet20	$91.95_{\pm 0.05}$	$32.65_{\pm 0.45}$	$12.75_{\pm 0.08}$	$91.94_{\pm 0.07}$	$26.92_{\pm 0.09}$	$11.99_{\pm 0.07}$	$91.95_{\pm 0.05}$	$24.11_{\pm 0.14}$	$11.74_{\pm 0.07}$
	PreResNet56	$94.38_{\pm 0.13}$	$22.46_{\pm 0.25}$	$9.03_{\pm 0.18}$	$94.34_{\pm 0.13}$	$19.58_{\pm 0.19}$	$8.61_{\pm 0.14}$	$94.38_{\pm 0.13}$	$17.48_{\pm 0.17}$	$8.41_{\pm 0.13}$
	PreResNet110	$94.86_{\pm 0.04}$	$20.42_{\pm 0.18}$	$8.24_{\pm 0.06}$	$94.83_{\pm 0.05}$	$18.06_{\pm 0.34}$	$7.91_{\pm 0.04}$	$94.86_{\pm 0.04}$	$16.11_{\pm 0.07}$	$7.72_{\pm 0.06}$
	PreResNet164	$95.24_{\pm 0.05}$	$18.61_{\pm 0.29}$	$7.58_{\pm 0.06}$	$95.14_{\pm 0.06}$	$17.39_{\pm 0.33}$	$7.33_{\pm 0.07}$	$95.24_{\pm 0.05}$	$14.96_{\pm 0.17}$	$7.13_{\pm 0.06}$
	VGG16BN	$93.26_{\pm 0.04}$	$33.76_{\pm 0.29}$	$11.61_{\pm 0.10}$	$93.23_{\pm 0.04}$	$25.15_{\pm 0.33}$	$10.42_{\pm 0.08}$	$93.26_{\pm 0.04}$	$24.55_{\pm 0.16}$	$10.48_{\pm 0.09}$
	WideResNet28x10	$95.54_{\pm 0.05}$	$15.69_{\pm 0.16}$	$7.00_{\pm 0.07}$	$95.53_{\pm 0.04}$	$16.49_{\pm 0.44}$	$6.86_{\pm 0.08}$	$95.54_{\pm 0.05}$	$14.50_{\pm 0.17}$	$6.80_{\pm 0.09}$
CIFAR-100	PreResNet20	$68.01_{\pm 0.15}$	$121.48_{\pm 1.12}$	$44.38_{\pm 0.15}$	$67.58_{\pm 0.13}$	$134.53_{\pm 1.92}$	$44.60_{\pm 0.07}$	$68.01_{\pm 0.15}$	$113.17_{\pm 0.30}$	$42.95_{\pm 0.09}$
	PreResNet56	$74.38_{\pm 0.10}$	$112.80_{\pm 2.88}$	$38.24_{\pm 0.43}$	$74.00_{\pm 0.09}$	$115.75_{\pm 1.44}$	$36.88_{\pm 0.17}$	$74.38_{\pm 0.10}$	$93.10_{\pm 1.02}$	$35.41_{\pm 0.21}$
	PreResNet110	$75.62_{\pm 0.14}$	$106.74_{\pm 0.53}$	$36.44_{\pm 0.17}$	$75.26_{\pm 0.09}$	$108.38_{\pm 1.00}$	$35.25_{\pm 0.15}$	$75.62_{\pm 0.14}$	$88.82_{\pm 0.45}$	$33.83_{\pm 0.15}$
	PreResNet164	$76.54_{\pm 0.05}$	$103.19_{\pm 0.55}$	$35.12_{\pm 0.13}$	$76.11_{\pm 0.10}$	$108.24_{\pm 0.85}$	$34.18_{\pm 0.12}$	$76.54_{\pm 0.05}$	$86.30_{\pm 0.54}$	$32.68_{\pm 0.13}$
	VGG16BN	$71.36_{\pm 0.08}$	$167.78_{\pm 0.98}$	$46.45_{\pm 0.16}$	$71.08_{\pm 0.18}$	$137.77_{\pm 0.91}$	$40.96_{\pm 0.14}$	$71.36_{\pm 0.08}$	$120.41_{\pm 0.51}$	$39.77_{\pm 0.13}$
	WideResNet28x10	$79.56_{\pm 0.22}$	$84.17_{\pm 0.80}$	$29.93_{\pm 0.31}$	$79.23_{\pm 0.21}$	$99.28_{\pm 1.49}$	$29.92_{\pm 0.28}$	$79.56_{\pm 0.22}$	$81.83_{\pm 0.68}$	$29.44_{\pm 0.29}$

We trained VGG16 (Simonyan & Zisserman, 2014) on a binary classification task, using a publicly available dataset of breast histopathology images (Janowczyk & Madabhushi, 2016). We used 10% of the image patches from the dataset, ensuring that the original 70:30 ratio of negative to positive points is maintained. We apply a smoothing technique (exponential moving average) to improve clarity. In Figure 6 we show the training and validation metrics per epoch and we observe several trends. For instance, we notice that as the model overfits and validation loss increases, the refinement remains fairly flat, and the increase in validation loss is only due to the increasing calibration error. $\widehat{\mathrm{CE}}_{\mathrm{KL}}$ not only correctly uncovers an early stopping point (same as the loss), but it also offers a more refined view of the nature of overfitting: while the validation accuracy remains constant, the CE considerably increases. For these reasons, we advocate for incorporating the calibration metric induced by the given loss as part of the standard practice for monitoring the training process.

Model selection In practical settings, achieving high accuracy alone is often not sufficient for deploying machine learning models in decision-making pipelines. Obtaining a low calibration error becomes an equally important aspect in the evaluation of the model's performance. In such multi-objective optimization problems, the Pareto front is a useful tool to determine the set of optimal solutions. Figure 7 shows the Pareto front of snapshots evaluated at every 10th epoch of training a VGG16 architecture on CIFAR-10 for a total of 250 epochs. Each point represents the mean and standard error across four seeds. The orange points are Pareto optimal (or Pareto efficient), meaning that no other snapshots can improve one objective without degrading the other. In

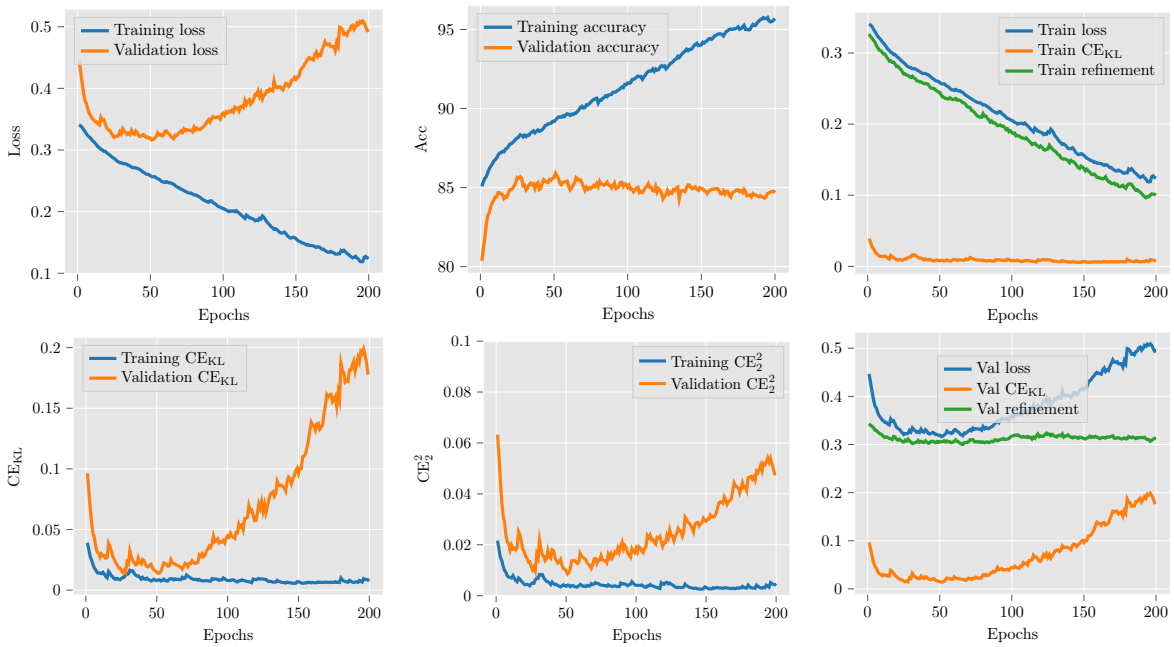


Figure 6: Training and validation trends monitoring loss, accuracy, calibration error and refinement. The calibration error is an effective tool for detecting overfitting. Observing the loss together with the induced calibration error and refinement offers unique insights for the performance of the model.

other words, the orange points represent the best possible trade-off between accuracy and calibration error.

Assessing calibration error In this part, we investigate the ranking performance of the class-wise versions of $\widehat{\text{CE}}_{\text{KL}}$ and $\widehat{\text{CE}}_2^2$ via kernel density estimation, and CE_1 with a binned estimator (Kull et al., 2019; Nixon et al., 2019). The latter can be seen as the class-wise version of the commonly used ECE (Naeini et al., 2015; Guo et al., 2017). Specifically, we evaluate calibration error using the three estimators for each of the ten classes within the CIFAR-10 dataset. The similarity of their performance in terms of ranking the classes can be observed in Table 5, where the numbers in the brackets represent the ranking order. We used 15 bins with equal-width binning scheme for CE_1, and the bandwidth for the KDE estimators was set to 0.02. Additionally, Figure 8 show the corresponding class-wise reliability diagrams for the models evaluated in the table. The blue bars represent the accuracy per bin. The red bars represent the gap of each bin to perfect calibration, i.e., the difference between accuracy and confidence for a given bin (darker shades signify under-confidence, while brighter red colors denote over-confidence).

Table 5: Calibration error evaluated using three estimators for each of the classes within CIFAR-10. The values are averaged over four seeds. The numbers in the brackets represent the ranking order.

Model	Metric	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8	Class 9
PreResNet56	$\widehat{\text{CE}}_{\text{KL}} \times 100$	1.58 (6)	0.91 (1)	2.13 (8)	4.28 (10)	1.78 (7)	3.59 (9)	1.30 (4)	1.09 (2)	1.12 (3)	1.34 (5)
	$\widehat{\text{CE}}_2^2 \times 1000$	7.25 (6)	4.04 (1)	9.27 (8)	16.96 (10)	8.26 (7)	14.58 (9)	5.34 (3)	5.77 (4)	5.31 (2)	5.99 (5)
	$\text{CE}_1 \times 1000$	6.25 (6)	3.36 (1)	7.65 (8)	15.62 (10)	6.51 (7)	12.76 (9)	4.34 (3)	3.96 (2)	4.52 (4)	4.97 (5)
WideResNet28x10	$\widehat{\text{CE}}_{\text{KL}} \times 100$	0.86 (5)	0.72 (3)	1.43 (8)	2.76 (10)	0.96 (7)	2.43 (9)	0.71 (2)	0.53 (1)	0.75 (4)	0.93 (6)
	$\widehat{\text{CE}}_2^2 \times 1000$	4.80 (6)	3.65 (2)	7.52 (8)	13.39 (10)	5.22 (7)	11.76 (9)	3.73 (3)	2.91 (1)	4.01 (4)	4.73 (5)
	$\text{CE}_1 \times 1000$	3.47 (5)	2.80 (3)	5.42 (8)	11.61 (10)	3.57 (7)	10.12 (9)	2.50 (2)	1.83 (1)	2.91 (4)	3.49 (6)

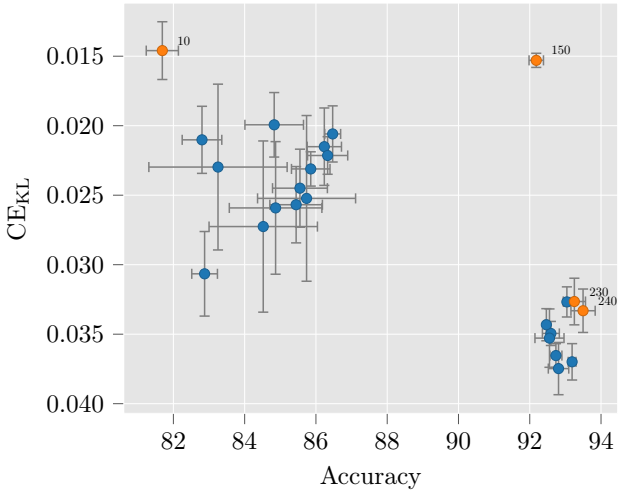


Figure 7: Pareto front of VGG16 snapshots evaluated at every 10th epoch. The model is trained for a total of 250 epochs on CIFAR-10. The orange points Pareto-dominate the rest. The numbers represent the corresponding epoch. Note that the $y-$axis is inverted.

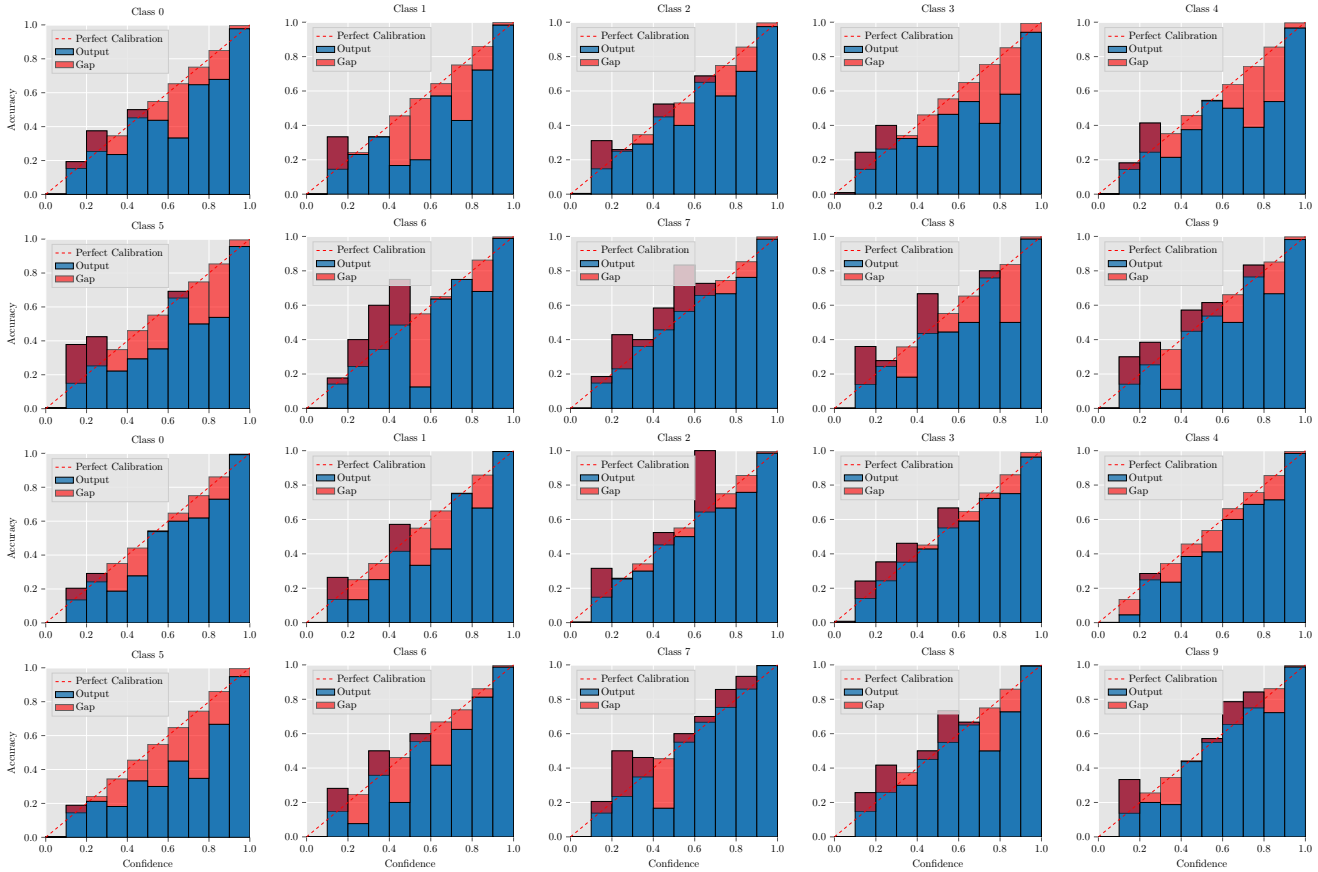


Figure 8: Reliability diagrams for each class of CIFAR-10 using PreResNet56 (top two rows) and WideResNet28x10 (bottom two rows).