A Note on "Assessing Generalization of SGD via Disagreement"

Anonymous authors
Paper under double-blind review

Abstract

Jiang et al. (2022) find empirically that the average test error of deep neural networks can be estimated via the prediction disagreement of two separately trained networks, which does not require labels. They show that this 'Generalization Disagreement Equality' follows from the well-calibrated nature of deep ensembles under the notion of a proposed 'class-aggregated calibration'. In this reproduction, we show on two datasets that the suggested theory might be impractical because a deep ensemble's calibration can deteriorate as prediction disagreement increases, which is precisely when the coupling of test error and disagreement is of interest, and labels are needed to estimate the calibration on new datasets. Further, we simplify the theoretical statements and proofs, showing them to be straightforward within a probabilistic context unlike the original hypothesis space view employed by Jiang et al. (2022).

1 Introduction

Machine learning models can cause great harm when their predictions become unreliable, yet we trust them blindly. Thus, finding ways to bound the test error of a trained deep neural network without access to the labels is of great importance: one could estimate the performance of models in the wild where unlabeled data is ubiquitous, labeling is expensive, and the data often does not match the training distribution. Crucially, it would provide a signal on when to trust the output of a model and when to defer to human experts instead.

Several recent works (Chen et al., 2021; Garg et al., 2022; Jiang et al., 2022) look at the question of how model predictions in a non-Bayesian setting can be used to estimate model accuracy.

In a Bayesian setting, *epistemic uncertainty* (Der Kiureghian & Ditlevsen, 2009) captures the uncertainty of a model about the reliability of its predictions, that is epistemic uncertainty quantifies the uncertainty of a model about its predictive distribution, while *aleatoric uncertainty* quantifies the ambiguity within the predictive distribution, c.f. Kendall & Gal (2017). Epistemic uncertainty thus tells us whether we can trust a model's predictions or not. Assuming a *well-specified* and *well-calibrated* Bayesian model, when its predictive distribution has low epistemic uncertainty for an input, it can be trusted. But likewise, a Bayesian model's calibration ought to deteriorate as epistemic uncertainty increases: in that case, the predictions become less reliable and so does the model's calibration.

In this context, calibration is an *aleatoric* metric for a model's reliability (Gopal, 2021). Calibration captures how well a model's confidence for a given prediction matches the actual frequency of that prediction in the limit of observations: when a model is 70% confident about assigning label A, does A indeed occur with 70% probability in these instances?

Jiang et al. (2022), while not Bayesian, make the very interesting empirical and theoretical discovery that deep ensembles satisfy a '*Generalization Disagreement Equality*' when they are well-calibrated according to a proposed '*class-aggregated calibration*' (or a '*class-wise calibration*') and empirically find that the respective calibration error generally bounds the absolute difference between the test error and '*disagreement rate*.' Jiang et al. (2022)'s theory builds upon Nakkiran & Bansal (2020)'s '*Agreement Property*' and provides backing for an empirical connection between the *test error* and *disagreement rate* of two separately trained networks on the same training data. Yet, while Nakkiran & Bansal (2020) limit the applicability of their Agreement

Property to in-distribution data, Jiang et al. (2022) carefully extend it: '*our theory is general and makes no restrictions on the hypothesis class, the algorithm, the source of stochasticity, or the test distributions (which may be different from the training distribution)*' with qualified evidence: '*we present preliminary observations showing that GDE is approximately satisfied even for certain distribution shifts within the PACS (Li et al., 2017) dataset.*'

In this paper, we present a new perspective on the theoretical results using a standard probabilistic approach for discriminative (Bayesian) models, whereas Jiang et al. (2022) use a hypothesis space of models which output one-hot predictions. Indeed, their theory does not require one-hot predictions, separately trained models (deep ensembles) or Bayesian models, and as remarked by the authors, their theoretical results also apply to a single model that outputs softmax probabilities. We will see that our perspective greatly simplifies the results and proofs.

This also means that the employed notion of disagreement rate *does not capture epistemic uncertainty but overall uncertainty*, similar to the predictive entropy, which is a major difference to Bayesian approaches which can evaluate epistemic uncertainty separately (Smith & Gal, 2018).

Importantly, we find that the connection between the proposed calibration metrics and the gap between test error and disagreement rate exists because the introduced notion of class-aggregated calibration is so strong that this connection follows almost at once.

Moreover, the suggested approach is circular[1]: calibration must be measured on the data distribution that we want to evaluate. Otherwise, we cannot bound the difference between the test error and the disagreement rate and obtain a signal on how trustworthy our model is. This reintroduces the need for labels on the unlabeled dataset, limiting practicality. Alternatively, one would have to assume that these calibration metrics do not change for different datasets or under distribution shifts, which we show to not hold: deep ensembles are less calibrated, the more the ensemble members disagree (even on in-distribution data).

Lastly, we draw connections and show that the 'class-aggregated calibration error' and the 'class-wise calibration error'[2] are equivalent to the 'adaptive calibration error' and 'static calibration error' introduced in Nixon et al. (2019).

Outline. We introduce the necessary background and notation in §2. In §3 we rephrase the theoretical statements from Jiang et al. (2022) using a parameter distribution (instead of a version space) and auxiliary random variables. This allows us to simplify the theoretical statements and proofs greatly in §4 and to examine the connection to Nixon et al. (2019). Finally, in §5, we provide empirical evidence that deep ensembles are less calibrated exactly when their ensemble members disagree.

2 Background & Setting

We introduce relevant notation, the initial Bayesian formalism, the connection to deep ensembles, and the probabilistic model. We restate the statements from Jiang et al. (2022) using this formalism in §3.

Notation. We use an implicit notation for expectations $\mathbb{E}[f(X)]$ when possible. For additional clarity, we also use $\mathbb{E}_X[f(X)]$ and $\mathbb{E}_{\mathrm{p}(x)} f(x)$, which fix the random variables and distribution, respectively, when needed.

We will use nested probabilistic expressions of the form $\mathbb{E}[\mathrm{p}(\hat{Y} = Y \mid X)]$. Prima facie, this seems unambiguous, but is $\mathrm{p}(\hat{Y} = Y \mid X)$ a transformed random variable of only X or also of Y (and $\hat{Y}$): what are we taking the expectation over? This is not always unambiguous, so we disambiguate between the probability for an event defined by an expression $\mathrm{p}[\dots] = \mathbb{E}[\mathbb{1}\{\dots\}]$, where $\mathbb{1}\{\dots\}$ is the indicator function[3], and a probability given specific outcomes for various random variables $\mathrm{p}(\hat{y} \mid x)$, c.f.:

$$\mathrm{p}[Y = \hat{Y} \mid X] = \mathbb{E}_{Y,\hat{Y}}[\mathbb{1}\{Y = \hat{Y}\} \mid X] = \mathbb{E}_{\mathrm{p}(y,\hat{y} \mid X)} \mathbb{1}\{y = \hat{y}\}, \tag{1}$$

---

[1] This was also added as a caveat to the camera-ready version of Jiang et al. (2022) after reviewing a preprint of this paper.

[2] Which is not explicitly introduced in Jiang et al. (2022) but can be analogously constructed.

[3] The indicator function is 1 when the predicate '...' is true and 0 otherwise.

which is a transformed random variable of X, while $\mathrm{p}(\hat{Y} = Y \mid X)$ is a (transformed) random variable that depends both on Y and X. That is, the difference is that Y is bound within the former but not the latter. In other words, $\mathrm{p}[\ldots \mid X]$ is a random variable in X, and any random variable that appears within the $\ldots$ is bound within that expression.

Probabilistic Model. We assume classification with K classes. For inputs X with ground-truth labels Y, we have a Bayesian model with parameters Ω that makes predictions $\hat{Y}$. What makes the model Bayesian is that the parameters follow a distribution $\mathrm{p}(\omega)$:

$$\mathrm{p}(y, \hat{y}, \omega \mid x) = \mathrm{p}(y \mid x)\ \mathrm{p}(\hat{y} \mid x, \omega)\ \mathrm{p}(\omega). \tag{2}$$

We focus on model evaluation. (Input) samples x can come either from 'in-distribution data' which follows the training set or from samples under covariate shift (distribution shift). The expected prediction over the model parameters is the *marginal predictive distribution*:

$$\mathrm{p}(\hat{y} \mid x) = \mathbb{E}_{\Omega}[\mathrm{p}(\hat{y} \mid x, \Omega)]. \tag{3}$$

On $\mathrm{p}(\omega)$. The main emphasis in Bayesian modelling can be Bayesian inference or Bayesian model averaging (Wilson & Izmailov, 2020). Here we concentrate on the model averaging perspective, and for simplicity take the model averaging to be with respect to *some* distribution $\mathrm{p}(\omega)$. Hence, we will use $\mathrm{p}(\omega)$ as the push-forward of models initialized with different initial seeds through SGD to minimize the negative log likelihood with weight decay and a specific learning rate schedule (MLE or MAP) (Mukhoti et al., 2021):

Assumption 1. We assume that $\mathrm{p}(\omega)$ is a distribution of possible models we obtain by training with a specific training regime on the training data with different seeds. A single ω identifies a single trained model.

Deep Ensembles. We cast deep ensembles (Hansen & Salamon, 1990; Lakshminarayanan et al., 2016), which refer to training multiple models and averaging predictions, into the introduced Bayesian perspective above by viewing them as an empirical finite sample estimate of the parameter distribution $\mathrm{p}(\omega)$. Then, $\omega_1, \ldots, \omega_N \sim \mathrm{p}(\omega)$ drawn i.i.d. are the *ensemble members*.

Again, the implicit model parameter distribution $\mathrm{p}(w)$ is given by the models that are obtained through training. Hence, we can view the predictions of a deep ensemble or the ensemble's prediction disagreement for specific x (or over the data) as empirical estimates of the predictions or the model disagreement using the implicit model distribution, respectively.

Calibration. The overall model's calibration for a given x measures how well the model's *(top-1) confidence*

$$\mathrm{Conf}_{\mathrm{Top1}} := \mathrm{p}(\hat{Y} = \arg\max_k \mathrm{p}(\hat{Y} = k \mid X) \mid X) \tag{4}$$

matches its *top-1 accuracy*

$$\mathrm{Acc}_{\mathrm{Top1}} := \mathrm{p}(Y = \arg\max_k \mathrm{p}(\hat{Y} = k \mid X) \mid X), \tag{5}$$

where we define both as transformed random variables of X. The calibration error is usually defined as the absolute difference between the two:

$$\mathrm{CE} := |\mathrm{Acc}_{\mathrm{Top1}} - \mathrm{Conf}_{\mathrm{Top1}}|. \tag{6}$$

In general, we are interested in the *expected calibration error (ECE)* over the data distribution (Guo et al., 2017) where we bin samples by their top-1 confidence. Intuitively, the ECE will be low when we can trust the model's top-1 confidence on the given data distribution.

Note that the accuracy if we were to draw $\hat{Y}$ according to $\mathrm{p}(\hat{y} \mid x)$ instead of taking the top-1 is

$$\begin{aligned} \mathrm{Acc} &:= \mathrm{p}[Y = \hat{Y} \mid X] && (7) \\ &= \sum_k \mathrm{p}(Y = k \mid X)\ \mathrm{p}(\hat{Y} = k \mid X) && (8) \end{aligned}$$

$$= \mathbb{E}_Y[\mathrm{p}(\hat{Y} = Y \mid X) \mid X], \tag{9}$$

and usually, we are interested in the accuracy over the whole dataset:

$$\mathrm{p}[\hat{Y} = Y] = \mathbb{E}[\mathrm{Acc}] = \mathbb{E}_X[\mathrm{p}[\hat{Y} = Y \mid X]] = \mathbb{E}_{X,Y}[\mathrm{p}(\hat{Y} = Y \mid X)]. \tag{10}$$

3 Rephrasing Jiang et al. (2022) in a Probabilistic Context

We present the same theoretical results as Jiang et al. (2022) but initially use a Bayesian formulation instead of a hypothesis space and define the relevant quantities as (transformed) random variables. As such, our definitions and theorems are equivalent and follow the paper but look different. We show these equivalences in §A in the appendix and prove the theorems and statements themselves in the next section.

First, however, we note a distinctive property of Jiang et al. (2022). It is assumed that each $\mathrm{p}(\hat{y} \mid x, \omega)$ is always one-hot for any ω. In practice, this could be achieved by turning a neural network's softmax probabilities into a one-hot prediction for the $\arg\max$ class. We call this the *Top1-Output-Property* (TOP).

Assumption 2. The Bayesian model $\mathrm{p}(\hat{y}, \omega \mid x)$ satisfies TOP when $\mathrm{p}(\hat{y}, \omega \mid x)$ is one-hot for all x and ω.

Definition 3.1. The *test error* and *disagreement rate*, as transformed random variables of Ω (and Ω'), are:

$$\begin{aligned}
\mathrm{TestError} &:= \mathrm{p}[\hat{Y} \neq Y \mid \Omega] && (11)\\
&= 1 - \mathrm{p}[\hat{Y} = Y \mid \Omega] && (12)\\
&= 1 - \mathbb{E}_{X,Y}[\mathrm{p}(\hat{Y} = Y \mid X, \Omega)], && (13)\\
&= 1 - \mathbb{E}_{\mathrm{p}(x,y)}\, \mathrm{p}(\hat{Y} = y \mid x, \Omega), && (14)\\
\mathrm{Dis} &:= \mathrm{p}[\hat{Y} \neq \hat{Y}' \mid \Omega, \Omega'] && (15)\\
&= 1 - \mathrm{p}[\hat{Y} = \hat{Y}' \mid \Omega, \Omega'] && (16)\\
&= 1 - \mathbb{E}_{X,\hat{Y}}[\mathrm{p}(\hat{Y}' = \hat{Y} \mid X, \Omega') \mid \Omega, \Omega'] && (17)\\
&= 1 - \mathbb{E}_{\mathrm{p}(x,\hat{y} \mid \Omega)}\, \mathrm{p}(\hat{Y}' = \hat{y} \mid x, \Omega'), && (18)
\end{aligned}$$

where for the disagreement rate, we expand our probabilistic model to take a second model Ω' with prediction $\hat{Y}'$ into account (and which uses the same parameter distribution), so:

$$\mathrm{p}(y, \hat{y}, \omega, \hat{y}', \omega' \mid x) := \mathrm{p}(y \mid x)\ \mathrm{p}(\hat{y} \mid x, \omega)\ \mathrm{p}(\omega)\, \mathrm{p}(\hat{Y} = \hat{y}' \mid x, \Omega = \omega')\ \mathrm{p}(\Omega = \omega').$$

Jiang et al. (2022) then introduce the property of interest:

Definition 3.2. The Bayesian model $\mathrm{p}(\hat{y}, \omega \mid x)$ satisfies the *Generalization Disagreement Equality (GDE)* when:

$$\mathbb{E}_\Omega[\mathrm{TestError}(\Omega)] = \mathbb{E}_{\Omega,\Omega'}[\mathrm{Dis}(\Omega, \Omega')] \quad (\Leftrightarrow \mathbb{E}[\mathrm{TestError}] = \mathbb{E}[\mathrm{Dis}]). \tag{19}$$

When this property holds, we seemingly do not require knowledge of the labels to estimate the test error: computing the disagreement rate is sufficient.

Two different types of calibration are then introduced, *class-wise* and *class-aggregated* calibration, and it is shown that they imply the GDE:

Definition 3.3. The Bayesian model $\mathrm{p}(\hat{y}, \omega \mid x)$ satisfies *class-wise calibration* when for any $q \in [0, 1]$ and any class $k \in [K]$:

$$\mathrm{p}(Y = k \mid \mathrm{p}(\hat{Y} = k \mid X) = q) = q. \tag{20}$$

Similarly, the Bayesian model $\mathrm{p}(\hat{y}, \omega \mid x)$ satisfies *class-aggregated calibration* when for any $q \in [0, 1]$:

$$\sum_k \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) = q \sum_k \mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q). \tag{21}$$

Theorem 3.4. *When the Bayesian model* $\mathrm{p}(\hat{y}, \omega \mid x)$ *satisfies class-wise or class-aggregated calibration, it also satisfies GDE.*

Finally, Jiang et al. (2022) introduce the *class-aggregated calibration error* similar to the ECE and then use it to bound the magnitude of any GDE gap:

Definition 3.5. The *class-aggregated calibration error (CACE)* is the integral of the absolute difference of the two sides in eq. (21) over possible $q \in [0,1]$:

$$\mathrm{CACE} := \int_{q \in [0,1]} \Big| \sum_k \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) - q \sum_k \mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q) \Big| dq. \tag{22}$$

Theorem 3.6. *For any Bayesian model* $\mathrm{p}(\hat{y}, \omega \mid x)$*, we have:*

$$|\mathbb{E}[\mathrm{TestError}] - \mathbb{E}[\mathrm{Dis}]| \leq \mathrm{CACE}.$$

4 GDE is Class-Aggregated Calibration in Expectation

Here, we show that proof for Theorem 3.6 is trivial if we use different but equivalent definitions of the class-wise and class-aggregate calibration. First though, we establish a better understanding for these definitions by examining the GDE property $\mathbb{E}[\mathrm{TestError}] = \mathbb{E}[\mathrm{Dis}]$. For this, we expand the definitions of $\mathbb{E}[\mathrm{TestError}]$ and $\mathbb{E}[\mathrm{Dis}]$, and use random variables to our advantage.

We define a quantity which will be of intuitive use later on: the *predicted accuracy*

$$\mathrm{PredAcc} := \mathbb{E}_{\hat{Y}}[\mathrm{p}(\hat{Y} \mid X) \mid X] = \sum_k \mathrm{p}(\hat{Y} = k \mid X)\, \mathrm{p}(\hat{Y} = k \mid X), \tag{23}$$

as a random variable of X. It measures the expected accuracy assuming the model's predictions are correct, that is the true labels follow $\mathrm{p}(\hat{y} \mid x)$.

Revisiting GDE. On the one hand, we have:

$$\begin{aligned}
\mathbb{E}[\mathrm{TestError}] &= \mathbb{E}_\Omega[\mathrm{p}[\hat{Y} \neq Y \mid \Omega]] && (24)\\
&= 1 - \mathrm{p}[\hat{Y} = Y] && (25)\\
&= 1 - \mathbb{E}_{X,\hat{Y}}[\mathrm{p}(Y = \hat{Y} \mid X)] && (26)\\
&= 1 - \mathbb{E}[\mathrm{Acc}] && (27)
\end{aligned}$$

and on the other hand, we have:

$$\begin{aligned}
\mathbb{E}[\mathrm{Dis}] &= \mathbb{E}_{\Omega,\Omega'}[\mathrm{p}[\hat{Y} \neq \hat{Y}' \mid \Omega, \Omega']] && (28)\\
&= 1 - \mathbb{E}_{\Omega,\Omega'}[\mathrm{p}[\hat{Y} = \hat{Y}' \mid \Omega, \Omega']] && (29)\\
&= 1 - \mathrm{p}[\hat{Y} = \hat{Y}'] && (30)\\
&= 1 - \mathbb{E}_{X,\hat{Y}}[\mathrm{p}(\hat{Y}' = \hat{Y} \mid X)] && (31)\\
&= 1 - \mathbb{E}_{X,\hat{Y}}[\mathrm{p}(\hat{Y} \mid X)] && (32)\\
&= 1 - \mathbb{E}[\mathrm{PredAcc}]. && (33)
\end{aligned}$$

The step from (30) to (31) is valid because $\hat{Y} \perp\!\!\!\perp \hat{Y}' \mid X$, and the step from (31) to (32) is valid because $\mathrm{p}(\hat{y}' \mid x) = \mathrm{p}(\hat{y} \mid x)$. Thus, we can rewrite Theorem 3.4 as:

Lemma 4.1. *The model* $\mathrm{p}(\hat{y} \mid x)$ *satisfies GDE, when*

$$\mathbb{E}[\mathrm{Acc}] = \mathbb{E}[\mathrm{p}(Y = \hat{Y} \mid X)] = \mathbb{E}[\mathrm{p}(\hat{Y} \mid X)] = \mathbb{E}[\mathrm{PredAcc}], \tag{34}$$

i.e. the accuracy of the model equals the predicted accuracy of the model, or equivalently, the error of the models equals the predicted error.

Crucially, while Jiang et al. (2022) calls $1 - \mathbb{E}_{X,\hat{Y}}[\mathrm{p}(\hat{Y} \mid X)]$ the expected disagreement rate $\mathbb{E}[\mathrm{Dis}]$, it actually is the predicted error of the (Bayesian) model as a whole.

Equally important, all dependencies on Ω have vanished. Indeed, we will not use Ω anymore for the remainder of this section. This reproduces the corresponding remark from Jiang et al. (2022)[4]:

Conclusion 1. The theoretical statements in Jiang et al. (2022) can be made about any discriminative model with predictions $\mathrm{p}(y \mid x)$.

When is $\mathbb{E}_{X,\hat{Y}}[\mathrm{p}(Y = \hat{Y} \mid X)] = \mathbb{E}_{X,\hat{Y}}[\mathrm{p}(\hat{Y} \mid X)]$? Or in other words: when does $\mathrm{p}(y = \hat{y} \mid x)$ equal $\mathrm{p}(\hat{y} \mid x)$ in expectation over $\mathrm{p}(x, y, \hat{y})$?

As a trivial sufficient condition, when the predictive distribution matches our data distribution—*i.e. when the model* $\mathrm{p}(\hat{y} \mid x)$ *is perfectly calibrated on average for all classes—and not only for the top-1 predicted class.* $ECE = 0$ is not sufficient because the standard calibration error only ensures that the data distribution and predictive distribution match for the top-1 predicted class (Nixon et al., 2019). But class-wise calibration entails this equality.

Class-Wise and Class-Aggregated Calibration. To see this, we rewrite class-wise and class-aggregated calibration slightly by employing the following tautology:

$$\mathrm{p}(\hat{Y} = k \mid \mathrm{p}(\hat{Y} = k \mid X) = q) = q, \tag{35}$$

which is obviously true due its self-referential nature. We provide a formal proof in §C in the appendix. Then we have the following equivalent definition:

Lemma 4.2. *The model* $\mathrm{p}(\hat{y} \mid x)$ *satisfies* class-wise calibration *when for any* $q \in [0, 1]$ *and any class* $k \in [K]$*:*

$$\mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) = \mathrm{p}(\hat{Y} = k, \mathrm{p}(\hat{Y} = k \mid X) = q). \tag{36}$$

Similarly, the model $\mathrm{p}(\hat{y} \mid x)$ *satisfies* class-aggregated calibration *when for any* $q \in [0, 1]$*:*

$$\mathrm{p}(\mathrm{p}(\hat{Y} = Y \mid X) = q) = \mathrm{p}(\mathrm{p}(\hat{Y} \mid X) = q), \tag{37}$$

and class-wise *calibration implies* class-aggregate *calibration.*

The straightforward proof is found in §C in the appendix.

Jiang et al. (2022) mention ‘level sets’ as intuition in their proof sketch. Here, we have been able to make this even clearer: class-aggregated calibration means that level-sets for accuracy $\mathrm{p}(\hat{Y} = Y \mid X)$ and predicted accuracy $\mathrm{p}(\hat{Y} \mid X)$—as random variables of Y and X, and $\hat{Y}$ and X, respectively—have equal measure, that is probability.

---

[4] The remark did not exist in the first preprint version.

GDE. Now, class-aggregated calibration immediately and trivially implies GDE. To see this, we use the following property of expectations:

Lemma 4.3. *For a random variable X, a function $t(x)$, and the random variable $T = t(X)$, it holds that*

$$\mathbb{E}_T[T] = \mathbb{E}[T] = \mathbb{E}_X[t(X)]. \tag{38}$$

This basic property states that we can either compute an expectation over T by integrating over $\mathrm{p}(T = t)$ or by integrating $t(x)$ over $\mathrm{p}(X = x)$. This is just a change of variable (push-forward of a measure).

We can use this property together with the class-aggregated calibration to see:

$$\begin{array}{ccc}
\mathbb{E}[\mathrm{Acc}] & & \mathbb{E}[\mathrm{PredAcc}] \\
\| & & \| \\
\mathbb{E}_{X,Y}[\mathrm{p}(\hat{Y} = Y \mid X)] & & \mathbb{E}_{X,\hat{Y}}[\mathrm{p}(\hat{Y} \mid X)] \\
\| & & \| \\
\mathbb{E}[\mathrm{p}(\hat{Y} = Y \mid X)] & & \mathbb{E}[\mathrm{p}(\hat{Y} \mid X)] \\
\| & & \| \\
\mathbb{E}_{q \sim \mathrm{p}(\hat{Y} = Y \mid X)}[q] & = & \mathbb{E}_{q \sim \mathrm{p}(\mathrm{p}(\hat{Y} \mid X))}[q],
\end{array} \tag{39}$$

which is exactly Lemma 4.1, where we start with the equality following from class-aggregated calibration and then apply Lemma 4.3 along each side. Thus, GDE is but an expectation over class-aggregated calibration; we have:

Theorem 4.4. *When a model $\mathrm{p}(\hat{y} \mid x)$ satisfies class-wise or class-aggregated calibration, it satisfies GDE.*

Proof. We can formalize the proof above slightly. We introduce two auxiliary random variables:

$$S := \mathrm{p}(\hat{Y} = Y \mid X), \tag{40}$$

as a transformed random variable of Y and X, and

$$T := \mathrm{p}(\hat{Y} \mid X), \tag{41}$$

as a transformed random variable of $\hat{Y}$ and X. Class-wise calibration implies class-aggregated calibration. Class-aggregated calibration then is $\mathrm{p}(S = q) = \mathrm{p}(T = q)$ (*). Writing out eq. (39), we have

$$\begin{aligned}
\mathbb{E}[\mathrm{p}(\hat{Y} = Y \mid X)] &= \mathbb{E}_{X,Y}[S] = \mathbb{E}[S] = \mathbb{E}_S[S] && (42) \\
&= \int \mathrm{p}(S = q)\, q \, dq && (43) \\
&\overset{(*)}{=} \int \mathrm{p}(T = q)\, q \, dq && (44) \\
&= \mathbb{E}_T[T] = \mathbb{E}[T] = \mathbb{E}_{X,\hat{Y}}[T] = \mathbb{E}[\mathrm{p}(\hat{Y} \mid X)], && (45)
\end{aligned}$$

which concludes the proof. ☐

The reader is invited to compare this derivation to the corresponding longer proof in the appendix of Jiang et al. (2022). The fully probabilistic perspective greatly simplifies the results, and the proofs are straightforward.

CACE. Showing that CACE bounds the gap between test error and disagreement is also straightforward:

Theorem 4.5. *For any model $\mathrm{p}(\hat{y} \mid x)$, we have:*

$$|\mathbb{E}[\mathrm{TestError}] - \mathbb{E}[\mathrm{Dis}]| \leq \mathrm{CACE}.$$

Proof. First, we note that

$$\mathrm{CACE} = \int_{q \in [0,1]} \big| \mathrm{p}(\mathrm{p}(\hat{Y} = Y \mid X) = q) - \mathrm{p}(\mathrm{p}(\hat{Y} \mid X) = q) \big| dq. \tag{46}$$

following the equivalences in the proof of Lemma 4.2. Then using the triangle inequality for integrals and $0 \leq q \leq 1$, we obtain:

$$\begin{aligned}
&\mathrm{CACE} && (47)\\
&= \int_{q\in[0,1]} \left|\, \mathrm{p}(\mathrm{p}(\hat{Y} = Y \mid X) = q) - \mathrm{p}(\mathrm{p}(\hat{Y} = \hat{Y} \mid X) = q) \right| dq && (48)\\
&\geq \int_{q\in[0,1]} q \left|\, \mathrm{p}(\mathrm{p}(\hat{Y} = Y \mid X) = q) - \mathrm{p}(\mathrm{p}(\hat{Y} = \hat{Y} \mid X) = q) \right| dq && (49)\\
&\geq \left| \int_{q\in[0,1]} q \; \mathrm{p}(\mathrm{p}(\hat{Y} = Y \mid X) = q) \, dq - \int_{q\in[0,1]} q \; \mathrm{p}(\mathrm{p}(\hat{Y} \mid X) = q) \, dq \right|. && (50)\\
&= \left| \mathbb{E}[S] - \mathbb{E}[T] \right| && (51)\\
&= \left| \mathbb{E}[\mathrm{TestError}] - \mathbb{E}[\mathrm{Dis}] \right|, && (52)
\end{aligned}$$

where we have used the monotonicity of integration in (49) and the triangle inequality in (50). □

The bound also serves as another—even simpler—proof for Theorem 4.4:

Conclusion 2. When the Bayesian model satifies class-wise or class-aggregated calibration, we have $\mathrm{CACE} = 0$ and thus $\mathbb{E}[\mathrm{TestError}] = \mathbb{E}[\mathrm{Dis}]$, i.e. the model satisfies GDE.

Furthermore, note again that a Bayesian model was not necessary for the last two theorems. The model parameters Ω were not mentioned—except for the specific definitions of TestError and Dis which depend on Ω following Jiang et al. (2022) but which we only use in expectation.

Moreover, we see that we can easily upper-bound CACE using the triangle inequality by 2, narrowing the statement in Jiang et al. (2022) that CACE can lie anywhere in $[0, K]$:

Conclusion 3. $\mathrm{CACE} \leq 2$.

Additionally, for completeness' sake, we can also define the class-wise calibration error formally as:

Definition 4.6. The *class-wise calibration error (CWCE)* is defined as:

$$\mathrm{CWCE} := \sum_k \int_{q\in[0,1]} \left|\, \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) - \mathrm{p}(\hat{Y} = k, \mathrm{p}(\hat{Y} = k \mid X) = q) \right|. \tag{53}$$

Using the triangle inequality, we have:

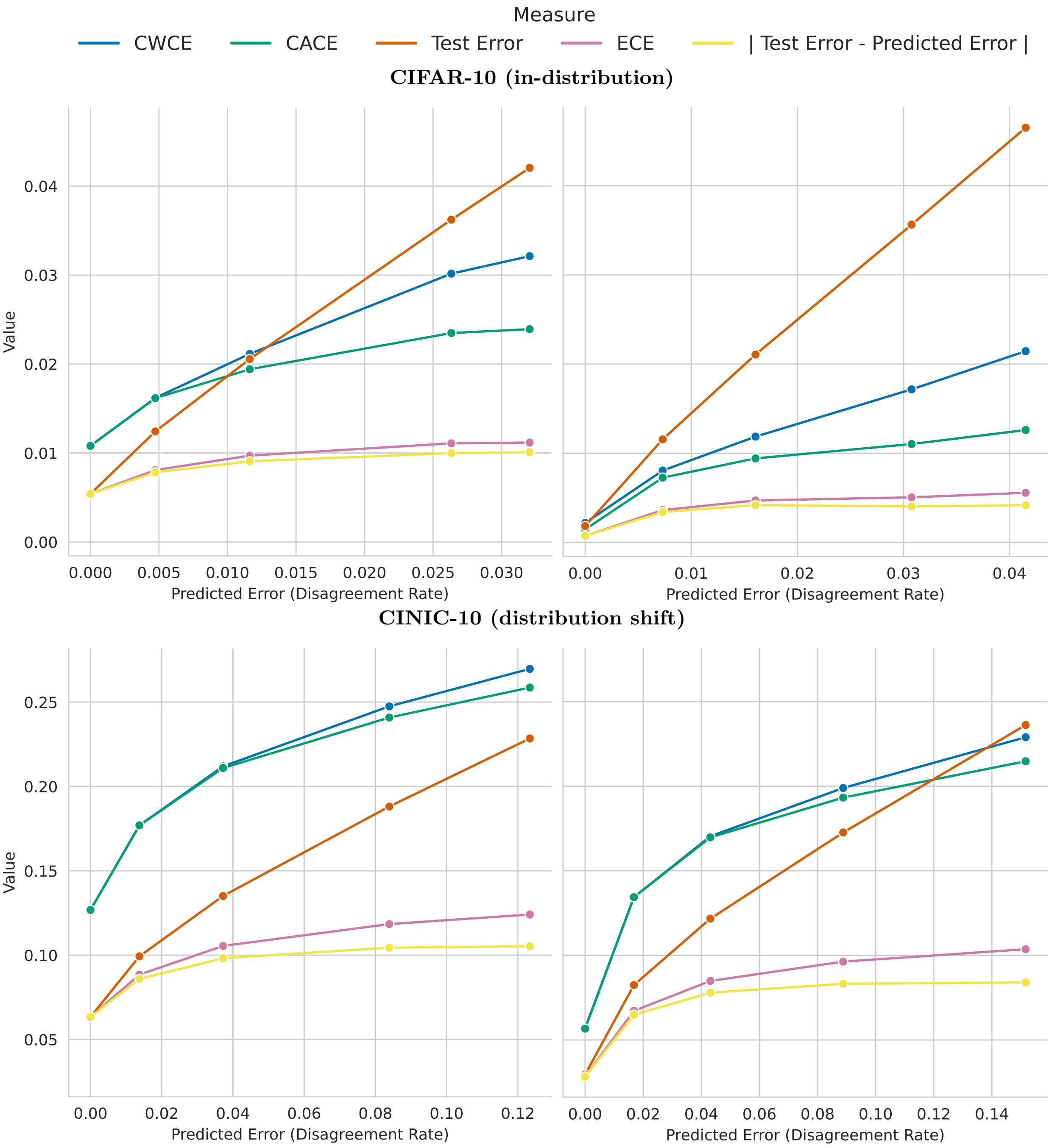
Lemma 4.7. $\mathrm{CWCE} \geq \mathrm{CACE} \geq |\, \mathbb{E}[\mathrm{Acc}] - \mathbb{E}[\mathrm{PredAcc}]|$.

Note that when we compute CACE empirically, we divide the dataset into several bins for different intervals of $\mathrm{p}(\hat{Y} = k \mid X)$. Jiang et al. (2022) use 15 bins. If we were to use a single bin, we would compute $|\, \mathbb{E}[\mathrm{Acc}] - \mathbb{E}[\mathrm{PredAcc}]|$ directly.

In §B we show that CWCE has previously been introduced as 'adaptive calibration error' in Nixon et al. (2019) and CACE as 'static calibration error' (with noteworthy differences between Nixon et al. (2019) and its implementation).

5 Deterioration of Calibration under Increasing Disagreement

Generally, we can only hope to trust model calibration for in-distribution data, while under distribution shift, the calibration ought to deteriorate. In our empirical falsification using models trained on CIFAR-10 and evaluated on the test sets of CIFAR-10 and CINIC-10 (as a dataset with a distribution shift), we find in both cases that calibration deteriorates under increasing disagreement. More importantly though, calibration markedly worsens under distribution shift.



(a) Ensemble with TOP

(b) Ensemble without TOP (same underlying models)

Figure 1: *Rejection Plot of Calibration Metrics for Increasing Disagreement In-Distribution (CIFAR-10) and Under Distribution Shift (CINIC-10).* Different calibration metrics (ECE, CWCE, CACE) vary across CIFAR-10 and CINIC-10 on an ensemble of 25 Wide-ResNet-28-10 model trained on CIFAR-10, depending on the rejection threshold of the predicted error (disagreement rate). Thus, calibration cannot be assumed constant for in-distribution data or under distribution shift. The test error increases almost linearly with the predicted error (disagreement rate), leading to 'GDE gap' |Test Error − Predicted Error| becoming almost flat, providing evidence for the empirical observations in Nakkiran & Bansal (2020); Jiang et al. (2022). The mean predicted error (disagreement rate) is shown on the x-axis. **(a)** shows results for an ensemble using TOP (following Jiang et al. (2022)), and **(b)** for a regular deep ensemble without TOP. The regular deep ensemble is better calibrated but has higher test error overall and lower test error for samples with small predicted error.

Specifically, we examine an ensemble of 25 WideResNet models (Zagoruyko & Komodakis, 2016) trained on CIFAR-10 (Krizhevsky et al., 2009) and evaluated on CIFAR-10 and CINIC-10 test data. CINIC-10 (Darlow et al., 2018) consists of CIFAR-10 and downscaled ImageNet samples for the same classes, and thus includes a distribution. The models are trained using the setup described in Mukhoti et al. (2021).

Figure 1 shows rejection plots under increasing disagreement for in-distribution data (CIFAR-10) and under distribution shift (CINIC-10). The rejection plots threshold the dataset on increasing levels of the predicted error (disagreement rate)—which is a measure of epistemic uncertainty when there is no expected aleatoric uncertainty in the dataset. We examine ECE, class-aggregated calibration error (CACE), class-wise calibration error (CWCE), error $\mathbb{E}[\text{TestError}]$, and 'GDE gap', $|\mathbb{E}[\text{Acc}] - \mathbb{E}[\text{PredAcc}]|$, as the predicted error (disagreement rate), $\mathbb{E}[\text{Dis}] = 1 - \mathbb{E}[\text{PredAcc}]$, increases.

We observe that all calibration metrics, ECE, CACE and CWCE, deteriorate under increasing disagreement, both in distribution and under distribution shift, and also worsen under distribution shift overall. This is consistent with the experimental results of Ovadia et al. (2019) which examines dataset shifts. However, given that the calibration metrics change with the quantity of interest, we conclude that:

Conclusion 4. The bound from Theorem 3.6 might not have as much expressive power as hoped since the calibration metrics themselves deteriorate as the model becomes more 'uncertain' about the data.

At the same time, the 'GDE gap', which is the actual gap between test error and predicted error, flattens, and the test error develops an almost linear relationship with the predicted error (up to a bias). This shows that there seem to be intriguing empirical properties of deep ensemble as observed previously (Nakkiran & Bansal, 2020; Jiang et al., 2022). However, they are not explained by the proposed calibration metrics[5].

As described previously, the results are not limited to Bayesian or version-space models but also apply to any model $\mathrm{p}(\hat{y} \mid x)$, including a regular deep ensembles without TOP. In our experiment, we find that a regular deep ensemble is better calibrated than the same ensemble made to satisfy TOP. We hypothesize that each ensemble member's own predictive distribution is better calibrated than its one-hot outputs, yielding a better calibrated ensemble overall.

Given that all these calibration metrics require access to the labels, and we cannot assume the model to be calibrated under distribution shift, we might just as well use the labels directly to asses the test error.

6 Conclusion

We have found that the theoretical statements in Jiang et al. (2022) can be expressed and proven more concisely when using probabilistic notation for (Bayesian) models that output softmax probabilities.

Moreover, we empirically found the proposed calibration metrics to deteriorate under increasing disagreement for in-distribution data, and as expected, we have found the same behavior under distribution shifts.

While Jiang et al. (2022) are careful to qualify their results for distribution shifts, above results should give us pause: strong assumptions are still needed to conjecture about model generalization, and we need to beware of circular arguments.

References

Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles, 2021.

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2): 105–112, 2009.

---

[5] The simplest explanation is that very few samples have high predicted error and thus the rejection plots flatten. This is not true. For CINIC-10, the first bucket contains 50k samples, and each latter buckets adds additional ∼10k samples.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance, 2022.

Achintya Gopal. Why calibration error is wrong given model uncertainty: Using posterior predictive checks with deep learning, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WvOGCEAQhxl.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision?, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*, 2021.

Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

A Equivalent Definitions

Jiang et al. (2022) defines a *hypothesis space* $\mathcal{H}$. In the literature, this is also sometimes called a version space. The hypothesis space induced by a stochastic training algorithm $\mathcal{A}$ is named $\mathcal{H}_{\mathcal{A}}$.

We can identify each hypothesis $h : \mathcal{X} \to [K]$ with itself as parameter $\omega_h = h$ and define $\mathrm{p}(\omega_h)$ as a uniform distribution over all parameters/hypotheses in $\mathcal{H}_{\mathcal{A}}$. This has the advantage of formalizing the distribution from which hypothesis are drawn ($h \sim \mathcal{H}_{\mathcal{A}}$), which is not made explicit in Jiang et al. (2022). $h(x) = k$ then becomes $\arg\max_{\hat{y}} \mathrm{p}(\hat{y} \mid x, \omega_h) = k$. Moreover, as $\mathrm{p}(\hat{y}, \omega \mid x)$ satisfies TOP, we have[6]

$$“\mathbb{1}\{h(x) = k\}'' = \mathrm{p}(\hat{Y} = k \mid x, \omega_h). \tag{54}$$

Thus, “$\mathrm{TestErr}_{\mathscr{D}}(h) \triangleq \mathbb{E}_{\mathscr{D}}[\mathbb{1}[h(X) \neq Y]]$” is equivalent to:

$$\begin{aligned}
“\,\mathrm{TestErr}_{\mathscr{D}}(h) &= \mathbb{E}_{\mathscr{D}}[\mathbb{1}[h(X) \neq Y]]'' && (55)\\
&= \mathbb{E}_{X,Y}[\mathrm{p}(\hat{Y} \neq Y \mid X, \omega_h)] && (56)\\
&= \mathbb{E}_{X}[\mathrm{p}[\hat{Y} \neq Y \mid X, \omega_h]] && (57)\\
&= \mathrm{p}[\hat{Y} \neq Y \mid \omega_h] && (58)\\
&= \mathrm{TestError}(\omega_h). && (59)
\end{aligned}$$

Similarly, “$\mathrm{Dis}_{\mathscr{D}}(h, h') \triangleq \mathbb{E}_{\mathscr{D}}[\mathbb{1}[h(X) \neq h'(X)]]$” is equivalent to:

$$\begin{aligned}
“\,\mathrm{Dis}_{\mathscr{D}}(h, h') &\triangleq \mathbb{E}_{\mathscr{D}}[\mathbb{1}[h(X) \neq h'(X)]]'' && (60)\\
&= \mathbb{E}_{\hat{Y},\hat{Y}'}[\mathrm{p}[\hat{Y} \neq \hat{Y}' \mid \omega_h, \omega_{h'}]] && (61)\\
&= \mathrm{Dis}(\omega_h, \omega_{h'}). && (62)
\end{aligned}$$

Further, “$\tilde{h}_k(x) \triangleq \mathbb{E}_{\mathscr{H}_{\mathcal{A}}}[\mathbb{1}[h(x) = k]]$” is equivalent to:

$$\begin{aligned}
“\tilde{h}_k(x) &\triangleq \mathbb{E}_{\mathscr{H}_{\mathcal{A}}}[\mathbb{1}[h(x) = k]]'' && (63)\\
&= \mathbb{E}_{\Omega}[\mathrm{p}(\hat{Y} = k \mid x, \Omega)] && (64)\\
&= \mathrm{p}(\hat{Y} = k \mid x). && (65)
\end{aligned}$$

For the GDE, “$\mathbb{E}_{h,h' \sim \mathscr{H}_{\mathcal{A}}}[\mathrm{Dis}_{\mathscr{D}}(h, h')] = \mathbb{E}_{h \sim \mathscr{H}_{\mathcal{A}}}[\mathrm{TestErr}(h)]$” is equivalent to:

$$\begin{aligned}
&“\mathbb{E}_{h,h' \sim \mathscr{H}_{\mathcal{A}}}[\mathrm{Dis}_{\mathscr{D}}(h, h')] = \mathbb{E}_{h \sim \mathscr{H}_{\mathcal{A}}}[\mathrm{TestErr}(h)]''\\
\Leftrightarrow\ &\mathbb{E}_{\Omega,\Omega'}[\mathrm{Dis}(\Omega, \Omega')] = \mathbb{E}_{\Omega}[\mathrm{TestError}(\Omega)]. && (66)
\end{aligned}$$

For the class-wise calibration, “$p(Y = k \mid \tilde{h}_k(X) = q) = q$” is equivalent to:

$$\begin{aligned}
&“p(Y = k \mid \tilde{h}_k(X) = q) = q'' && (67)\\
\Leftrightarrow\ &\mathrm{p}(Y = k \mid \mathrm{p}(\hat{Y} = k \mid X) = q) = q. && (68)
\end{aligned}$$

For the class-aggregated calibration, “$\frac{\sum_{k=0}^{K-1} p(Y=k, \tilde{h}_k(X)=q)}{\sum_{k=0}^{K-1} p(\tilde{h}_k(X)=q)} = q$” (and note in Jiang et al. (2022), class indices run from $0..K-1$) is equivalent to:

$$“\frac{\sum_{k=0}^{K-1} p(Y = k, \tilde{h}_k(X) = q)}{\sum_{k=0}^{K-1} p(\tilde{h}_k(X) = q)} = q'' \tag{69}$$

---

[6] We put definitions and expressions written using the notation and variables from Jiang et al. (2022) inside quotation marks “” to avoid ambiguities.

$$\Leftrightarrow \frac{\sum_{k=1}^{K} \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q)}{\sum_{k=1}^{K} \mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q)} = q \tag{70}$$

$$\begin{aligned} &\Leftrightarrow \sum_{k=1}^{K} \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) \\ &\qquad = q \sum_{k=1}^{K} \mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q). \end{aligned} \tag{71}$$

Finally, for the class-aggregated calibration error, the definition is equivalent to:

$$\begin{aligned} &\text{“ } \mathrm{CACE}_{\mathscr{D}}(\tilde{h}) \\ &\triangleq \int_{q \in [0,1]} \left| \frac{\sum_k p\left(Y = k, \tilde{h}_k(X) = q\right)}{\sum_k p\left(\tilde{h}_k(X) = q\right)} - q \right| \cdot \sum_k p\left(\tilde{h}_k(X) = q\right) dq && (72) \\ &= \int_{q \in [0,1]} \Big| \sum_k p\left(Y = k, \tilde{h}_k(X) = q\right) - q \sum_k p\left(\tilde{h}_k(X) = q\right) \Big| dq'' && (73) \\ &= \int_{q \in [0,1]} \Big| \sum_k \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) - q \sum_k \mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q) \Big| dq && (74) \end{aligned}$$

B Comparison of CACE and CWCE with calibration metrics with ‘adaptive calibration error’ and ‘static calibration error’

Nixon et al. (2019) examine shortcomings of the ECE metric and identify a lack of class conditionality, adaptivity and the focus on the maximum probability (argmax class) as issues. They suggest an adaptive calibration error which uses adaptive binning and averages of the calibration error separately for each class, thus equivalent to the class-wise calibration error and class-wise calibration (up to adaptive vs. even binning). In the paper, the static calibration error is defined as ACE with even instead of adaptive binning. However, in the widely used implementation[7], SCE is defined as equivalent to the class-aggregated calibration error.

C Additional Proofs

Lemma C.1. *For a model* $\mathrm{p}(\hat{y} \mid x)$*, we have for all* $k \in [K]$ *and* $q \in [0, 1]$*:*

$$\mathrm{p}(\hat{Y} = k \mid \mathrm{p}(\hat{Y} = k \mid X) = q) = q, \tag{75}$$

when the left-hand side is well-defined.

Proof. This is equivalent to

$$\mathrm{p}(\hat{Y} = k, \mathrm{p}(\hat{Y} = k \mid X) = q) = q\ \mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q), \tag{76}$$

as the conditional probability is either defined or $\mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q) = 0$. Assume the former. Let $\mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q) > 0$. Introducing the auxiliary random variable $T_k := \mathrm{p}(\hat{Y} = k \mid X)$ as a transformed random variable of X, we have

$$\mathrm{p}(\hat{Y} = k, T_k = q) = q\ \mathrm{p}(T_k = q). \tag{77}$$

We can write the probability as an expectation over an indicator function

$$\mathrm{p}(\hat{Y} = k, T_k = q) \tag{78}$$

---

[7] https://github.com/google-research/robustness_metrics/blob/baa47fbe38f80913590545fe7c199898f9aff349/robustness_metrics/metrics/uncertainty.py#L1585, added in April 2021

$$= \mathbb{E}_{X,\hat{Y}}[\mathbb{1}\{\hat{Y} = k, T_k(X) = q\}] \tag{79}$$

$$= \mathbb{E}_{X,\hat{Y}}[\mathbb{1}\{\hat{Y} = k\}\,\mathbb{1}\{T_k(X) = q\}] \tag{80}$$

$$= \mathbb{E}_X[\mathbb{1}\{T_k(X) = q\}\ \mathbb{E}_{\hat{Y}}[\mathbb{1}\{\hat{Y} = k\} \mid X]] \tag{81}$$

$$= \mathbb{E}_X[\mathbb{1}\{T_k(X) = q\}\ \mathrm{p}(\hat{Y} = k \mid X)]. \tag{82}$$

Importantly, if $\mathbb{1}\{T_k(x) = q\} = 1$ for an x, we have $T_k(x) = \mathrm{p}(\hat{Y} = k \mid x) = q$, and otherwise, we multiply with 0. Thus, this is equivalent to

$$= \mathbb{E}_X[\mathbb{1}\{T_k(X) = q\}\,q] \tag{83}$$

$$= q\,\mathbb{E}_X[\mathbb{1}\{T_k(X) = q\}] \tag{84}$$

$$= q\,\mathrm{p}(T_k(X) = q). \tag{85}$$

□

Lemma 4.2. *The model* $\mathrm{p}(\hat{y} \mid x)$ *satisfies* class-wise calibration *when for any* $q \in [0, 1]$ *and any class* $k \in [K]$*:*

$$\mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) = \mathrm{p}(\hat{Y} = k, \mathrm{p}(\hat{Y} = k \mid X) = q). \tag{36}$$

Similarly, the model $\mathrm{p}(\hat{y} \mid x)$ *satisfies* class-aggregated calibration *when for any* $q \in [0, 1]$*:*

$$\mathrm{p}(\mathrm{p}(\hat{Y} = Y \mid X) = q) = \mathrm{p}(\mathrm{p}(\hat{Y} \mid X) = q), \tag{37}$$

and class-wise *calibration implies* class-aggregate *calibration.*

Proof. Beginning from

$$\mathrm{p}(Y = k \mid \mathrm{p}(\hat{Y} = k \mid X) = q) = q, \tag{86}$$

we expand the conditional probability to

$$\Leftrightarrow \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) = q\ \mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q), \tag{87}$$

and substitute eq. (35) into the outer q, obtaining the first equivalence

$$\Leftrightarrow \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) = \mathrm{p}(\hat{Y} = k, \mathrm{p}(\hat{Y} = k \mid X) = q). \tag{88}$$

For the second equivalence, we follow the same approach. Beginning from

$$\sum_k \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) = q \sum_k \mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q), \tag{89}$$

we pull the outer q into the sum and expand using (35)

$$\Leftrightarrow \sum_k \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = k \mid X) = q) = \sum_k q\,\mathrm{p}(\mathrm{p}(\hat{Y} = k \mid X) = q) = \sum_k \mathrm{p}(\hat{Y} = k, \mathrm{p}(\hat{Y} = k \mid X) = q). \tag{90}$$

In the inner expression, k is tied to Y on the left-hand side and $\hat{Y}$ on the right-hand side, so we have

$$\Leftrightarrow \sum_k \mathrm{p}(Y = k, \mathrm{p}(\hat{Y} = Y \mid X) = q) = \sum_k \mathrm{p}(\hat{Y} = k, \mathrm{p}(\hat{Y} \mid X) = q). \tag{91}$$

Summing over k, marginalizes out $Y = k$ and $\hat{Y} = k$ respectively, yielding the second equivalence

$$\Leftrightarrow \mathrm{p}(\mathrm{p}(\hat{Y} = Y \mid X) = q) = \mathrm{p}(\mathrm{p}(\hat{Y} \mid X) = q). \tag{92}$$

Finally, class-wise calibration implies class-aggregated calibration as summing over different k in (88), which is equivalent to class-wise calibration, yields (90), which is equivalent to class-aggregated calibration. □