Collision Probability Matching Loss for Disentangling Epistemic Uncertainty from Aleatoric Uncertainty

Hiromi Narimatsu
NTT Communication Science Laboratories

Mayuko Ozawa
Ritsumeikan University

Shiro Kumano
NTT Communication Science Laboratories

Abstract

Two important aspects of machine learning, uncertainty and calibration, have previously been studied separately. The first aspect involves knowing whether inaccuracy is due to the epistemic uncertainty of the model, which is theoretically reducible, or to the aleatoric uncertainty in the data per se, which thus becomes the upper bound of model performance. As for the second aspect, numerous calibration methods have been proposed to correct predictive probabilities to better reflect the true probabilities of being correct. In this paper, we aim to obtain the squared error of predictive distribution from the true distribution as epistemic uncertainty. Our formulation, based on second-order Rényi entropy, integrates the two problems into a unified framework and obtains the epistemic (un)certainty as the difference between the aleatoric and predictive (un)certainties. As an auxiliary loss to ordinary losses, such as cross-entropy loss, the proposed collision probability matching (CPM) loss matches the cross collision probability between the true and predictive distributions to the collision probability of the predictive distribution, where these probabilities correspond to accuracy and confidence in confidence calibration, respectively. Unlike previous Shannon-entropy-based uncertainty methods, the proposed method makes the aleatoric uncertainty directly measurable as test-retest reliability, which is a summary statistic of the true distribution frequently used in scientific research on humans. We provide mathematical proof and strong experimental evidence for our formulation using both a real dataset consisting of real human ratings toward emotional faces and simulation.

1 Introduction

With the advances in machine learning techniques, computational power, and data availability, the performance of machine classifiers/regressors has dramatically improved in various tasks such as object recognition and speech recognition [Zhang et al., 2017, Zhao et al., 2019b]. However, accuracy depends not only on modeling imperfections but also on the uncertainty in the data [Mukhoti et al., 2021]. For tasks that inherently involve uncertainty and difficulty in determining a single ground truth [Kramer et al., 2018], even advanced neural models have difficulty achieving high performance. Consequently, without further examination it is unclear whether low performance stems from model imperfections or uncertainty of the ground truth.

One example is the prediction of human subjective judgments, such as emotion recognition and medical diagnosis. Human judgments are known to be ambiguous [Truong et al., 2009, Flexer and Lallai, 2019, Marmpena et al., 2018], meaning that the same person does not necessarily give the same label to the same item [Kramer et al., 2018]. Therefore, in scientific studies on humans, it is crucial to rigorously measure such intrapersonal (un)certainty, which is called intrarater reliability or test-retest reliability. Test-retest reliability is typically calculated using labels of the same item from the same person at two time points. Such measures include, for example, the percentage of their observed agreement, i.e., a kappa statistic [Sim and Wright, 2005] or Pearson's and intraclass correlations [Koo and Li, 2016]. Imperfect test-retest reliability implies that it is difficult to develop a model with an accuracy of 100%. In fact, in the task of personalized facial emotion perception recognition, i.e., the prediction of how a target respondent judges the emotional state of a target face, accuracy remains around 0.5 in five-class classification [Zhou et al., 2021].

In the artificial intelligence community, recent research has focused on two types of uncertainty to better understand the causes of model inaccuracies [Hüllermeier and Waegeman, 2021, Abdar et al., 2021]. The first is aleatoric uncertainty,

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

which is essentially the variability of data due to inherent randomness. The other is epistemic uncertainty due to the lack of knowledge about the true model, i.e., model imperfection. When modeling human impressions and emotions, the ambiguities contained in the target of modeling are regarded as aleatoric uncertainties. By definition, aleatoric uncertainty is model-agnostic and forms the upper bound of model performance[1]. Accordingly, no matter how much the modeling performance is improved, it is impossible to reduce the aleatoric uncertainty.

The predictive uncertainty of model is usually treated as the summation of two uncertainties [Ghahramani, 2015]. For instance, in a pioneering work, Smith and Gal [Smith and Gal,] defined epistemic uncertainty as the information gain or mutual information (MI) between model and data, and they derived the following: predictive uncertainty (entropy) = epistemic uncertainty (MI) + aleatoric uncertainty (entropy). Their aleatoric entropy is defined as the expected predictive entropy over Monte-Carlo dropout samples. However, their definition of aleatoric uncertainty unfavorably depends on the model. Actually, their aleatoric uncertainty becomes zero when all Monte-Carlo dropout samples give the same predictive distribution. This contradicts the model-agnostic nature of aleatoric uncertainty.

Another separately tackled yet related topic is confidence calibration. Confidence is commonly defined as the probability that the decision maker's answer is correct. By calibrating the probability of the predicted label, the final probability output from the model can be measured as a predictive uncertainty [Ovadia et al., 2019]. Although some training methods themselves produce well, though not perfectly, calibrated results [Thulasidasan et al., 2019], post-hoc methods are often used, including temperature scaling [Guo et al., 2017] and intra order-preserving, which are applicable to neural networks with softmax layers as output layers.

Confidence calibration is a straightforward solution if the ground truth of a respondent's confidence, i.e., the true confidence level, is available in judgment tasks. In such cases, the epistemic uncertainty can be quantified by simply subtracting the true confidence from the calibrated predictive uncertainty. Unfortunately, this is not always the case in practice. As metacognitive researchers have demonstrated, people cannot accurately report their own confidence in two-alternative forced-choice (2AFC) tasks [Jogan and Stocker, 2014]. Consequently, self-reported confidence does not necessarily reflect the probability of their judgment being correct. This discrepancy has been observed even more clearly in three-class settings [Li and Ma, 2020].

Therefore, we propose *collision probability matching (CPM), or κ-matching (kappa matching),* to measure how close the current model performance is to the upper bound as the squared error of the predictive distribution from the true distribution, which is our definition of epistemic uncertainty. We obtain aleatoric uncertainty as the difference between the mean likelihood of data, which equals the collision probability of predictive distribution, as shown later, and the directly measured aleatoric uncertainty as the test-retest reliability, i.e., kappa statistic. We demonstrate that our formulation can be viewed as a second-order Rényi-entropy version of the conventional maximum-probability-based confidence calibration. The clear interpretability of epistemic uncertainty is of particular importance in human-oriented science, which often relies on squared-error-based hypothesis testing such as analysis of variance (ANOVA) or intraclass correlations. This perspective would be suitable for the relevant interdisciplinary areas, including affective computing. It fills in the gap between AI and human-oriented scientific fields by exploiting kappa-statistic-based (i.e., choice-behavior-based, neither self-reported nor model-predicted) test-retest reliability as a *direct measurement* of aleatoric uncertainty.

In this paper, we provide a mathematical proof of our proposal and evaluate the epistemic uncertainty of a neural model by applying our method to the contents of a facial emotion judgment dataset as targets with high aleatoric uncertainty and simulated data as targets with true confidence distribution. The main contributions of this study can be summarized as follows:

- We propose *collision probability matching (CPM) loss* as a way to accurately measure epistemic uncertainty by taking the difference from aleatoric uncertainty calculated for the data themselves using test-retest reliability.
- We give a proof showing that through CPM, the epistemic uncertainty as the squared error of the predictive distribution from the (unknown) true distribution can be quantified as the difference between the test-retest reliability and mean data likelihood.
- We experimentally verify the effectiveness of the proposed method using a real dataset and simulation [2].

2 Related work

The distinction between the two types of uncertainty, aleatoric and epistemic uncertainties, has been studied in

[1] In terms of whether aleatoric uncertainty is independent of the model, it is not completely model independent [Hüllermeier and Waegeman, 2021]. For example, respondent's affective state (e.g., boredom) is considered as a source of aleatoric uncertainty in the this paper, but it can be partly modeled and removed in our definition.

[2] A sample code of CPM is available at https://github.com/nttcslab/collision-probability-matching.

the machine learning community. Aleatoric uncertainty is regarded as uncertainty due to overlapping class distributions, and it involves quantifying such uncertainties based on binary or multi-class classification within the framework of fuzzy preference relations [Senge et al., 2014, Nguyen et al., 2018]. However, these studies assumed it was possible to obtain the probability of each class label, and thus it is difficult to adapt them to our target, where the ground truth of the confidence is not available.

Other researchers have attempted to quantify aleatoric and epistemic uncertainties based on model output. In the field of Bayesian neural networks [Denker and LeCun, 1990], Gal and Ghahramani claimed that sampling methods with dropout nodes could be used to estimate a model's uncertainty because model uncertainty is regarded as a probability distribution over outputs [Gal and Ghahramani, 2016]. To extend other models, Smith and Gal proposed a method to measure the epistemic uncertainty for detecting adversarial examples by calculating the information gain or mutual information between model and data [Smith and Gal,]. They formulated epistemic entropy as the difference between predictive entropy and aleatoric entropy based on Shannon's entropy. In their study, aleatoric uncertainty was approximated as the mean predictive entropy over Monte-Carlo dropout samples. Shi et al. focused on decomposing the predicted entropy over multiple classes into two distinct sources of uncertainty from the viewpoint of evidence types, and they defined second-order uncertainty using Dirichlet probability distribution functions [Shi et al., 2020]. Jain et al. proposed a method to directly predict intrinsic uncertainty, and they mentioned the special case where the aleatoric uncertainty is constant and can be estimated using the mean squared error of two predicted values as the oracle [Jain et al., 2021]. One advantage of these methods is that no extra data are needed. Of particular note is that in these methods aleatoric uncertainty is *model-dependent*. However, we define aleatoric uncertainty as a *model-independent* criterion, which results in a different formulation.

Predictive uncertainty is closely related to calibration methods [Guo et al., 2017]. Since confidence calibration aims to match the probability of the model's answer, which is usually the maximum probability class, to the representative of the true correctness likelihood, the calibrated confidence can be regarded as the predictive uncertainty. Confidence calibration has been explored in Bayesian binning [Zadrozny and Elkan, 2001] and in an intermediary approach between sigmoid fitting and binning, named isotonic regression [Zadrozny and Elkan, 2002]. With the aim of easily adapting a calibration method to a variety of models, including deep neural network models, Guo et al. proposed a simple approach named temperature scaling [Guo et al., 2017]. Temperature scaling is an extended method of Plat scaling [Niculescu-Mizil and Caruana, 2005]. Since it only increases or decreases the output entropy in post-processing, it does not affect the model's accuracy. Therefore, if the model's accuracy and the output entropy are matched, it can be said that the (distribution of the) confidence or accuracy represents the predictive entropy of the model, which is regarded as the predictive uncertainty. Although confidence calibration focuses only on the predicted class, class-wise and multi-class calibration methods attempt to calibrate all classes. For example, Kull et al. [Kull et al., 2019] proposed a Dirichlet calibration method for multi-class calibration. However, none of the above calibration methods has significantly incorporated aleatoric uncertainty. Therefore, in this paper, to exploit aleatoric uncertainty directly measured from the dataset itself, we reformulate the calibration problem based on second-order Rényi entropy.

3 Collision Probability Matching (CPM)

Here, we clarify the problem setting of this study. We consider multi-class classification, which aims to find a function f, given a data point $x \in \mathbb{R}^d$, that returns a predictive distribution $q \in \Delta^{C-1}$, where Δ^{C-1} is a $(C-1)$-dimensional simplex satisfying $\sum_c q_c = 1$. Predictive class $\hat{y}_{max}$ is assumed to be the class with the maximum probability in the predictive distribution q. If the ground truth *class* is assumed to exist, the classical goal of classification is to find a function that returns $\hat{y}_{max}$ that is as close as possible to the ground truth class. However, the data generation process usually includes uncertainties, such as interpersonal difference or even intrapersonal variability in cognitive judgment, so it is reasonable to assume that there exists no single ground-truth class; instead, it is necessary to find a function that yields predictive distribution q as close as possible to the ground truth *distribution* $p \in \Delta^{C-1}$. Therefore, we attempt to quantify the squared error of the predictive distribution q from the true distribution p as the epistemic (un)certainty of the model, i.e.:

$$\epsilon = \sum_c (p_c - q_c)^2. \tag{1}$$

If it were possible to ask people to judge the same item many times, we could obtain an accurate estimate of the true distribution p, and thus we could directly minimize ϵ. Unfortunately, this is difficult for practical and reliability reasons; for the latter, their judgment may be altered by being exposed to the same item repeatedly. Therefore, we intend to minimize the number of repetitions of rating for each pair of respondent and item, i.e., two. Actually, although not common in the machine learning community, the repetition of rating is a common procedure in human behavior studies to measure test-retest reliability as the probability of the same item receiving the same rating from the same person on two separate occasions. We

Dist. to be used	Prob.-dist.-based measure		Max-prob.-based measure
True distribution p	True collision prob. ($\mathbf{True\text{-}CP}$) $\sum_c p_c^2$ Aleatoric certainty (Test-retest reliability)	$\leq$	True confidence ($\mathbf{TC}$) $p_{y_{max}}$ (hard to obtain)
	$\geq$		$\geq$
True dist. p and predictive dist. q	Cross collision prob. ($\mathbf{Cross\text{-}CP}$) $\sum_c p_c q_c$ Data likelihood	$(\leq)$	Machine accuracy ($\mathbf{MA}$) $p_{\hat{y}_{max}}$
	$=$		$(\simeq)$
Predictive dist. q	Predictive collision prob. ($\mathbf{Pred\text{-}CP}$) $\sum_c q_c^2$ Predictive certainty	$\leq$	Machine confidence ($\mathbf{MC}$) $q_{\hat{y}_{max}}$

Table 1: Important measures for describing our concept. Here, dist. and prob. represent distribution and probability, respectively. The proposed collision probability matching (CPM) matches predictive collision probability $\mathrm{Pred\text{-}CP}$ to cross collision probability $\mathrm{Cross\text{-}CP}$ (indicated by "="). As a side effect, it also makes machine confidence MC close to machine accuracy MA (indicated by "$\simeq$"), which is the goal of confidence calibration. The equations and inequalities in black without parentheses are mathematically valid, but those with parentheses are also approximately achieved by the proposed CPM.

demonstrate that approximated epistemic uncertainty ϵ is measurable using the proposed collision probability matching and the test-retest reliability.

3.1 Measuring epistemic uncertainty ϵ and the proposed collision probability matching loss

Our idea for approximating epistemic (un)certainty ϵ is to use the probability-distribution measures based on second-order Rényi entropy, i.e., collision probabilities (or, more accurately, the exponential of second-order Rényi entropy × -1), listed in the middle column of Table 1: 1) collision probability of the true distribution p, i.e., $\sum_c p_c^2$, which we call true collision probability (True-CP); 2) cross collision probability between the true distribution p and predictive distribution q, i.e., $\sum_c p_c q_c$, which we call cross collision probability (Cross-CP); and 3) collision probability of predictive distribution q, i.e., $\sum_c q_c^2$, which we call predictive collision probability (Pred-CP). Since collision probability is entropy × -1, it is a measure of *certainty*. In what follows, however, we refer to certainty indiscriminately as uncertainty, except where necessary, because uncertainty is used much more frequently in the literature.

Now, let us clarify the main point of the proposed method. In our formulation, when we restrict q to satisfy the condition that its predictive collision probability (Pred-CP) matches to the cross collision probability (Cross-CP), the epistemic uncertainty ϵ is obtained as the difference between the true collision probability (True-CP) and the predictive collision probability (Pred-CP), namely:

$$\epsilon = \mathrm{True\text{-}CP} - \mathrm{Pred\text{-}CP}, \text{ if } \mathrm{Pred\text{-}CP} = \mathrm{Cross\text{-}CP}, \tag{2}$$

which is proven as follows. We call the constraint of $\mathrm{Pred\text{-}CP} = \mathrm{Cross\text{-}CP}$ collision probability matching (CPM), and we call the squared difference of the two terms, i.e. $(\mathrm{Pred\text{-}CP} - \mathrm{Cross\text{-}CP})^2$, *the CPM loss*.

If $\mathrm{Pred\text{-}CP}$ ($\sum_c q_c^2$) is forced to be equal to $\mathrm{Cross\text{-}CP}$ ($\sum_c p_c q_c$), the difference between $\mathrm{True\text{-}CP}$ ($\sum_c p_c^2$) and $\mathrm{Pred\text{-}CP}$ becomes identical to the definition of ϵ in (1), namely the Euclidean distance of q to p:

$$\begin{aligned}
\epsilon &= \sum_c p_c^2 - \sum_c q_c^2 \\
&= \sum_c p_c^2 - \sum_c q_c^2 + 2(\sum_c q_c^2 - \sum_c p_c q_c) \\
&\quad (\because \sum_c q_c^2 - \sum_c p_c q_c = 0) \\
&= \sum_c (p_c^2 - 2p_c q_c + q_c^2) \\
&= \sum_c (p_c - q_c)^2 \\
&\geq 0.
\end{aligned} \tag{3}$$

As explained below, True-CP is obtained as the test-retest reliability, i.e., the percentage of observed agreement of two ratings that the same person gives to the same item at two time points. True-CP is the aleatoric certainty by definition. Cross-CP represents the mean data likelihood, the greater value of which indicates higher performance. Pred-CP is predictive certainty, and it is easy to obtain by definition. Since ϵ is the squared error, it becomes 0 if and only if the predictive distribution q is identical to the true distribution p, and positive otherwise. Therefore, the test-retest reliability is the upper bound of the mean likelihood.

The need for multiple ratings (even though only twice) to measure test-retest reliability is the major limitation of the proposed method. However, the CPM constraint per se needs no secondary labels, and our $\mathrm{Pred\text{-}CP}$ remains an interpretable, *relative* fit measure for model comparison. However, our method is most useful when secondary labels are available making it to be a *fully interpretable, absolute* measure *upper bounded* by test-retest reliability. This is the advantageous property over other related criteria: such as Cohen's kappa, which can be viewed as (True-CP − Cross-CP) / (1 − Cross-CP), and Akaike/Bayesian Information Criteria (AIC/BIC), which are based on mean *log* likelihood, though neither is CPM-constrained nor upper-bounded.

We hereinafter consider the expected Euclidean distance under empirical data distribution as:

$$\begin{aligned} E_{x\sim p_{data}}[\,\epsilon] &= E_{x\sim p_{data}}[\sum_c p_c^2 - \sum_c q_c^2] \\ &= E_{x\sim p_{data}}[\sum_c p_c^2] - E_{x\sim p_{data}}[\sum_c q_c^2]. \end{aligned} \tag{4}$$

3.2 True collision probability $\mathrm{True\text{-}CP}$

We obtain $\mathrm{True\text{-}CP}$ as the expected collision probability of the true distribution p, i.e., the expectation of $\sum_c p_c^2$ under the empirical data distribution. This means the probability that two random samples from the distribution p are identical if they are derived independently. We obtain it as the expected probability that the same respondent chooses the same class twice over multiple items.

Suppose that a target person judges each of M images twice, namely that we have two sets of ratings, $Y = \{y^{(1)}, \cdots, y^{(M)}\}$ and $Y' = \{y'^{(1)}, \cdots, y'^{(M)}\}$, for the first and second ratings, respectively, of M images. The expected collision probability under the empirical distribution given by the data is obtained as

$$\begin{aligned} \mathrm{True\text{-}CP} &:= E_{x\sim p_{data}}[\sum_c p(y=c|x)^2] \\ &= \frac{1}{M}\sum_m\sum_c p(y=c|x^{(m)})^2 && (5) \\ &\approx \frac{1}{M}\sum_m\sum_c 1(y^{(m)}=c)1(y'^{(m)}=c) && (6) \\ &= \frac{1}{M}\sum_m 1(y^{(m)}=y'^{(m)}), \end{aligned}$$

where the conversion from (5) to (6) is based on the assumption that there is a pair of random ratings drawn for $x^{(m)}$. The last form means the frequency that the rating scores in both sequences Y and Y' are identical; this is regarded as the test-retest reliability of data.

Collision probability is a measure of *certainty*, not uncertainty, and $\mathrm{True\text{-}CP}$ is defined solely on the data itself. Therefore, it can be considered the aleatoric certainty. However, if a measure of uncertainty is preferable, we can use $1 - \mathrm{True\text{-}CP}$. Since collision probability ranges from $1/C$ (in the case of flat distribution) to 1 (when one class has the probability of 1 and the rest have 0), it represents an uncertainty ranging from 0 to $1 - 1/C$.

Moreover, the collision probability is equal to or less than True Confidence, as shown in Table 1 and proven in Supplemental A. In addition, to minimize the impact of the secondary ratings, we use Y' only to calculate $\mathrm{True\text{-}CP}$.

3.3 Cross collision probability $\mathrm{Cross\text{-}CP}$

We obtain $\mathrm{Cross\text{-}CP}$ as the expected cross collision probability between the true distribution p and predictive distribution q, i.e., the expectation of $\sum_c p_c q_c$ under the empirical distribution. The expected cross collision probability is expressed as

$$\begin{aligned} \mathrm{Cross\text{-}CP} &:= E_{x\sim p_{data}}[\sum_c p(y=c|x)q(y=c|x)] \\ &= \frac{1}{M}\sum_m\sum_c p(y=c|x^{(m)})q(y=c|x^{(m)}) && (7) \\ &\approx \frac{1}{M}\sum_m\sum_c 1(y^{(m)}=c)q(y=c|x^{(m)}) && (8) \\ &= \frac{1}{M}\sum_m q(y=y^{(m)}|x^{(m)}), \end{aligned}$$

where the conversion from (7) to (8) is based on the assumption that there are N random ratings drawn for $x^{(m)}$ but $N = 1$. Since $q(y = y^{(m)}|x^{(m)})$ is the likelihood of data point $x^{(m)}$ for the given label $y^{(m)}$, the expected cross collision probability is obtained as the mean likelihood of observed dataset Y. Note that the expected cross collision probability consequently gives the *mean* likelihood for all data points, while the ordinary cost function uses the *product* of the likelihood based on the joint probability of the dataset.

3.4 Predictive collision probability $\mathrm{Pred\text{-}CP}$

We obtain $\mathrm{Pred\text{-}CP}$, i.e., the expected predictive distribution q or the expectation of $\sum_c q_c^2$ under the empirical data distribution, in a way similar to $\mathrm{True\text{-}CP}$, i.e.,

$$\begin{aligned} \mathrm{Pred\text{-}CP} &:= E_{x\sim p_{data}}[\sum_c q(y=c|x)^2] \\ &= \frac{1}{M}\sum_m\sum_c q(y=c|x^{(m)})^2. \end{aligned}$$

Here, $\mathrm{Pred\text{-}CP}$ is defined solely on the model, and it may contain both aleatoric and epistemic uncertainties.

3.5 Comparison with conventional definition and confidence calibration

The form of (2) is reasonable when compared with the conventional definition of epistemic uncertainty in the literature. For example in one previous work [Mukhoti et al., 2021, Smith and Gal,], based on Shannon's entropy, i.e., first-order Rényi entropy, epistemic uncertainty (defined as mutual information) is obtained by subtracting the aleatoric uncertainty (entropy) from the predictive uncertainty (entropy). Concisely, this is expressed as epistemic uncertainty = aleatoric uncertainty - predictive uncertainty. By replacing uncertainty with - certainty, we obtain epistemic certainty = aleatoric certainty - predictive certainty. This is the form of (2).

From the perspective of confidence calibration, the three collision probabilities (True-CP, Cross-CP and Pred-CP) correspond to maximum-probability-based measures: true confidence (TC), machine accuracy (MA), and machine confidence (MC), respectively, as listed in the rightmost column of Table 1. MA and MC are set as the targets to match in ordinary confidence calibration [Guo et al., 2017]. Machine accuracy is expressed as $\text{MA} = p_{\hat{y}_{max}}$. When the prediction is $\hat{y}_{max}$ under the true distribution p, this indicates the probability of the prediction being correct, namely the common definition of accuracy. Machine confidence is obtained as $\text{MC} = q_{\hat{y}_{max}}$. This is the probability of the predicted label $\hat{y}_{max}$ in the predictive distribution q. True (human) confidence (TC) is the ground truth of the machine confidence, and it is expressed as $p_{y_{max}}$. Since $p_{y_{max}}$ is the maximum probability of p, it is obvious that $p_{y_{max}} \geq p_{\hat{y}_{max}}$. Therefore, if true confidence is available, we can consider TC as the upper bound of machine accuracy, i.e., aleatoric uncertainty in the data, and the difference between TC and MA as possible room for improvement, i.e., the epistemic uncertainty of the machine. Unfortunately, as explained in Section 1, it is difficult to obtain the true confidence (TC) regarding human cognition. On the other hand, the test-retest reliability is a choice-behavior-based criterion, which can be measured directly from given labels and thus is arguably more reliable.

The predictive collision probability Pred-CP, like machine confidence, may be very high when the machine is overconfident. In the extreme case, when the predictive distribution has a probability of 1 for one class and 0 for the rest, Pred-CP becomes 1. It exceeds True-CP when aleatoric uncertainty exists, i.e., $\text{True-CP} < 1$. The CPM constraint suppresses such overconfidence.

4 Evaluation

To measure the effect of incorporating the proposed method as a cost function for training and that of the model's epistemic uncertainty, we evaluated the change in measures using practical data and simulated data. For the practical data, we used actual human rating data for emotion targets from Dataset-REA, a real dataset using a facial emotion judgment task [Kumano and Nomura, 2019]. In the simulation, we first fitted an item-response model commonly used in the psychological literature on Dataset-REA, and then considered the fitted probability distribution for each respondent to each item as the true distribution p. Then we calculated each measure and evaluated the effects using the dataset.

Note that although there are various affective datasets, including IAPS [Lang et al., 1997] and OASIS [Kurdi et al., 2017], only a very limited number of datasets contain ratings given by *individual* respondents *twice*, while most publicly available datasets include only summary statistics, such as mean and variance, and do not include test-retest reliability, as partly summarized in an earlier work [Zhao et al., 2019a].

Dataset-REA This is the facial emotion judgment dataset. Here, $N = 50$ participants gave 5-point Likert ratings of valence (positive vs. negative) and arousal (high vs. low) emotional dimensions [Russell, 1980] to 120 artificial facial images showing one or two (mixture) of neutral and six basic emotions (i.e., anger, disgust, fear, sadness, surprise, and happiness). For measuring test-retest reliability, each respondent rated a randomly selected 25% portion of the 120 images twice; accordingly, M was 120 for Cross-CP and Pred-CP but 30 for True-CP. True-CP is person-dependent and was calculated as the mean test-retest reliability of the 50 individuals. As the core affect [Russell, 2003], the above two emotional dimensions have frequently been used in the fields of social psychology, metacognition, and affective computing to capture important components of emotions.

Dataset-SIM This dataset includes the distribution of artificially created TC distributions. Since these data are used to verify that our method can accurately calculate the epistemic uncertainty and TC is difficult to obtain, we artificially generated the distribution TC based on Dataset-REA to avoid synthetic data that could not occur in reality. We first set the ground truth distribution p as the predictive distribution of a much simpler model. We used a graded response model, from a family of item-response theory, which has more commonly been used in psychological literature, and fitted the model on Dataset-REA, using the Stan programming language with the Hamiltonian Monte Carlo (HMC). Then, all rating data for $N = 50$ persons and $M = 120$ images were generated by randomly sampling from the fitted categorical distribution p. The test-retest reliability was calculated using the generated second rating for each image in 25% of the images randomly selected from M images for each person.

Training methods. As methods for learning the predictive distribution of personalized cognitive judgment, i.e., those output probability distributions that explain how likely an individual person is to give a particular label to a target, we compared two training losses.

One is the CE Only as a baseline method only using conventional Cross-Entropy loss for the loss function, and the other is our CE+CPM loss, which introduces CPM for CE loss. To compare CE Only and CE+CPM, we built a three-layer neural network. Specifically, the input layer had 85 nodes and took a 35-dimensional vector, including the intensities of 17 AUs and the presence of 18 AUs detected using OpenFace [Baltrusaitis et al., 2018], concatenated with a 50-dimensional one-hot vector, each element of which represents the target person, as $\boldsymbol{x}_i$. The middle layer had 4,608 nodes. The final layer was a softmax layer that outputs a C-d vector, each element of which represents the probability of the corresponding rating score. Here, $C = 5$ in the experiments. The model was trained by minimizing the weighted sum of the standard cross-entropy loss and the respondent-wise squared difference between $\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$ as follows:

$$\begin{aligned}\mathrm{L}_1 = &-\frac{1}{MN}\sum_m\sum_n\sum_c y_c^{(mn)} log(q_c^{(mn)}) \\ &+ w\frac{1}{N}\sum_n(\text{Cross-CP}^{(n)} - \text{Pred-CP}^{(n)})^2,\end{aligned}$$

where n is the index of respondent and w is set to 100 to balance the two losses for CE+CPM. As for the CE Only, we simply cut off the CPM loss by setting w = 0.

The model was trained at the learning rate of the Adam optimizer with 2e-4. The mini-batch size was set to its maximum, i.e., 6,000. These methods were assessed in a leave-one-image-out cross-validation scenario, that is, trained using $M - 1$ images for the entire N persons ($(M - 1) \times N$ samples) and tested on the remaining 1 image for the N persons (N samples) as test data. Although this neural model is relatively small in terms of the number of input dimensions and intermediate layers, it is reasonable for evaluating the proposed method because the method is model-independent. Furthermore, the other reason to use a simple neural model is that its well-studied features are known in dealing with the target data, and they also perform reasonably well.

To calculate the machine accuracy and confidence, the predicted labels were determined to be the maximum probability class. We implemented these methods ourselves due to their simplicity. The lack of existing deep neural networks for personalized cognitive judgment makes it difficult to compare the proposed method with previous ones, such as those using dropout samples. Moreover, the requirement for additional data, i.e., secondary ratings, would make the proposed method less suitable for test data. Therefore, we consider the proposed method to be more meaningful for use on training data, and we performed the evaluation using the entire dataset (6,000 samples) both for training and test sets.

Results-REA We first verified the effectiveness of CPM using CE+CPM loss. Figure 1 shows how the model was trained when the proposed loss function CE+CPM was used for training. All scores that changed by training increased as the number of epochs increased on both valence and arousal data. Furthermore, the results show that $\mathrm{True\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$ got closer as training progressed, and MA and MC also got closer as a corollary. This indicates that the CPM produced the effect of confidence calibration. In addition, the machine accuracy MA ($\sum_c p_c q_c = p_{\hat{y}_{max}}$) exceeded the score of the test-retest reliability $\mathrm{True\text{-}CP}$ on the arousal dataset at around 600 epochs, although the model used in this evaluation is a simple model without much elaboration. However, the effect of CPM brought MA and MC closer together, resulting in MA converging around the test-retest reliability score.

Then, we compared the measures obtained with the CE Only and CE+CPM losses, which represent the conditions before and after the introduction of the proposed CPM to CE loss (Table 2). The results show that both $\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$ are lower than $\mathrm{True\text{-}CP}$ in our CE+CPM, while they are equal or higher than CP in CE Only. These results indicate that the relationships between the three collision probabilities shown in Table 1 were satisfied in the experiment for both valence and arousal dimensions. In addition, the epistemic uncertainty of the methods, which can be calculated as the difference between the test-retest reliability ($\mathrm{True\text{-}CP}$) and the matched $\mathrm{True\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$, was 0.14 $(= 0.61 - 0.47)$ for valence and 0.11 $(= 0.50-0.39)$ for arousal. These measures are verified in the simulation results and further discussed in Section 5.

Results-SIM To verify the correctness of epistemic uncertainty calculated by our method, we compared the ground truth epistemic uncertainty using Dataset-SIM. To see the impact of the number of items (images), we randomly selected n images from Dataset-SIM. Sampling was done 10 times with different random seeds for each n, then training and testing were also done 10 times respectively. Figure 2 shows the results of ϵ error, calculated by mean value of the ground truth ϵ minus estimated $\hat{\epsilon}$, $|\hat{\epsilon} - \epsilon|$, of the 50 persons, where the number of items increases. The results indicate that the error mostly converges and approaches zero as the number of data increases. These results support the relationship in (3), indicating that our method correctly estimates the true epistemic uncertainty. (See Supplemental B for other comparison results of ϵ and $\hat{\epsilon}$ according to w.)

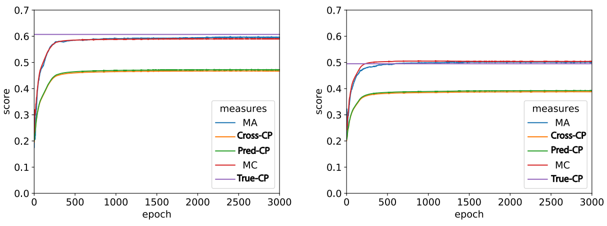
CE+CPM		CE Only	
Full prob.-dist.	Max-prob.	Full prob.-dist.	Max-prob.
True-CP: 0.61	TC: –	True-CP: 0.61	TC: –
$\geq$		$\leq$	
Cross-CP: 0.47	MA: 0.60	Cross-CP: 0.61	MA: 0.62
$\parallel$		$\nparallel$	
Pred-CP: 0.47	MC: 0.59	Pred-CP: 0.94	MC: 0.96

(a) CE+CPM vs. CE Mode in Valence dataset.

CE+CPM		CE Only	
Full prob.-dist.	Max-prob.	Full prob.-dist.	Max-prob.
True-CP: 0.50	TC: –	True-CP: 0.53	TC: –
$\geq$		$\leq$	
Cross-CP: 0.39	MA: 0.50	Cross-CP: 0.55	MA: 0.54
$\parallel$		$\nparallel$	
Pred-CP: 0.39	MC: 0.50	Pred-CP: 0.90	MC: 0.93

(b) CE+CPM vs. CE Mode in Arousal dataset.

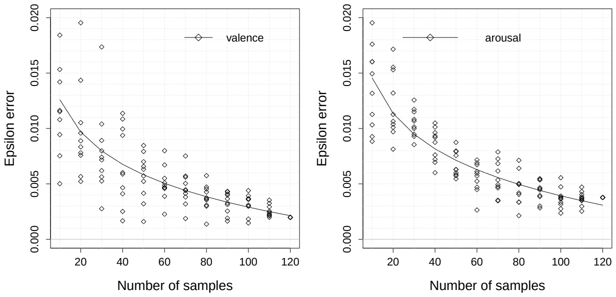
Table 2: Results when the model was trained using our CE+CPM loss and those with CE Only. For CE Only, Cross-CP and Pred-CP were equal to or greater than True-CP, and it was clearly overconfident. On the other hand, CE+CPM successfully satisfied that Cross-CP and Pred-CP do not exceed True-CP. As a side effect, MC was also well calibrated, and this resulted in MA=MC.



(a) Valence

(b) Arousal

Figure 1: Training progress of CE+CPM with respect to the five performance measures listed in Table 1, excluding the unknown true confidence (TC). For both emotion dimensions, the training progressed rapidly at first (< 300 epochs) but then suddenly slowed. These values nearly converged at an epoch between 2,000 and 3,000; at the final epoch, Cross-CP $= 0.47$ vs. Pred-CP $= 0.47$ and MA $=$ 0.60 vs. MC $= 0.59$ for valence, and Cross-CP $= 0.39$ vs. Pred-CP $= 0.39$ and MA $= 0.50$ vs. MC $= 0.50$ for arousal. Note that although we only brought Cross-CP (orange) close to Pred-CP (green), confidence calibration, namely MA (blue) = MC (red), was almost achieved as well.



(a) Valence

(b) Arousal

Figure 2: The epsilon error is the difference between the estimated epistemic uncertainty and true epistemic uncertainty calculated as the mean of $|\hat{\epsilon} - \epsilon|$ of the 50 persons, where CE+CPM loss was used. x-axis represents the number of images used. Each plot shows the epsilon error of 10 trials. The lines show the regression curves obtained by a logarithmic approximation. For both valence and arousal, the error tended to converge and approach zero gradually as the number of data increased.

5 Discussion

Reliability of True-CP. True-CP is person-dependent and it was calculated as the mean test-retest reliability of N individuals, depending on the total number of secondary labels (i.e. $\propto MN$). We also calculated the test-retest reliability error as the difference between the True-CP, and the mean proportion of matched pairs between two randomly selected labels for each p. The mean absolute error was as small as $0.0044{\pm}0.0036$ (SD). This suggests that True-CP is reliable.

Independence of test-retest pairs. The test-retest pairs may not be fully independent, but their impact looks limited, as demonstrated in Fig. 2. Memory is a major factor of their dependence [Müller et al., 2012], but the influence has been shown to be limited in visual tasks with some intra-pair intervals [McKelvie, 1992]. Further, in order to minimize their influence on evaluation, we used the secondary labels only to determine test-retest reliability and not to train the model.

Handling small data. We defined the problem based on the expected values of both true and predictive collision probabilities as well as cross collision probability. This is a natural choice for extending the binary observation, i.e., match or mismatch, for each data point to a continuous measure (test-retest reliability). However, this means that the limited sample size may bias the test-retest reliability (True-CP) and the cross collision probability Cross-CP. Therefore, for small sizes, a Bayesian approach such as hierarchical models, in which respondents (or items) are handled as a random factor, would be useful for measuring the test-retest reliability.

Necessity of the CPM during training. In terms of methodology, the proposed CPM, which brings $\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$ closer together, can also be achieved post-hoc, e.g. by using the temperature scaling, with decoupling from model training. However, in our preliminary experiment, the temperature scaling was not able to estimate epistemic uncertainty ϵ accurately (see Supplemental C for details). This suggests that the CPM is needed during training.

Difficulties in comparison with existing uncertainty methods. This paper focuses on showing the proof of the proposed CPM and its experimental verification, while comparisons of uncertainty values, which are calculated by conventional methods, are beyond the scope of this paper. Direct comparison with popular methods is hard due to the different definition of aleatoric uncertainty. In MC-dropout literature [Smith and Gal,], aleatoric uncertainty is defined as $\mathbb{E}_{p(\omega|D)}H[p(y|x,\omega)]$, where the expectation is approximated using dropout samples from $p(\omega|D)$; p is a standard notation here. This means that the uncertainty depends on model parameter ω (as in ensemble and variational inference methods), unlike ours. In fact, when we used the code of [Gal and Ghahramani, 2016] for our data, the uncertainty decreased (though slightly) as the dropout rate is lower (Pearson's r=.66), as expected. We expect various models to be evaluated with our method in the future, and this will contribute to improving model performance not only apparently but also truly.

Advantages compared to direct confidence rating approaches. In terms of both data collection and modeling, our approach is better suited for larger data and thus for deep learning than confidence annotations. Our choice-based approach provides unbiased measures based on central limit theory, but needs some interval between first and second labeling (though the pairs and their intervals can be overlapped with each other) to minimize the memory effect. Confidence assessment works even for small datasets, but suffer from people's overconfidence or bias [Lichtenstein et al., 1982].

Applicability to binary and multi labels. Since the proposed method is in a general mathematical formulation, another important direction would be to validate its effectiveness on other various uncertain/difficult tasks, such as human-preference estimation, aesthetic/ethical/moral judgment, and medical diagnosis. For example, the proposed method would also be applicable for modeling binary responses as well as multiclass responses, and those of each individual (y_n) and their mean ($\hat{y}$) (by ignoring n and considering each individual's response as a single measurement of a randomly selected individual from the population). This would work if individual difference is small enough compared to item (image) variability. While human judgment covers a wide range of applications, such as classification, prediction, and decision making, examples outside this range would be death/survival of living organisms (e.g., under chemical dose, stress) for binary cases and DNA replication (normal, insertion, or deletion) for multiclass cases. See Supplemental Material D for more details.

The CPM also may work for multilabel approach when each multilabel is normalized to be unit-sum. However, it appears to be problematic in measuring $\mathrm{True\text{-}CP}$. It can be assumed that each class label follows an independent Bernoulli distribution and that reliability=1 if the multilabel has a single true label, and reliability=0 otherwise. For binary classes, the reliability equals the $\mathrm{True\text{-}CP}$. But, for $C > 2$ classes, they matched neither mathematically nor experimentally in our preliminary simulation analysis.

Potential insights that our method may provide. Curiously, the proposed method revealed a similarity between valence and arousal dimensions in the epistemic uncertainty of the models used in this study, although the literature has repeatedly reported different performances between the two dimensions on various models without distinguishing the two types of uncertainties [Gunes and Pantic, 2010]. This suggests that their difference in predictive performance may stem from their difference in aleatoric uncertainty in the data. This would provide some insight for future studies in affective computing and related areas, such as social/cognitive psychology.

6 Conclusion

In this paper, we proposed collision probability matching (CPM) loss, which integrates the problems of epistemic uncertainty measurement and calibration into a unified framework based on second-order Rényi entropy. We defined epistemic uncertainty as the squared error of predictive distribution from the true distribution and obtained it as the difference between aleatoric uncertainty and predictive uncertainty. In our formulation, aleatoric uncertainty is equal to the standard definition of test-retest reliability, namely the probability of the same item receiving the same rating from the same person on two separate occasions, whereas predictive uncertainty is constrained such that the cross collision probability between the true and predictive distributions matches the collision probability of predictive distribution. We provided a proof of the methodology as well as supporting results on real and simulated dataset.

References

[Abdar et al., 2021] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76:243–297.

[Baltrusaitis et al., 2018] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L. (2018). Openface 2.0: Facial behavior analysis toolkit. In 13th IEEE International Conference on Automatic Face Gesture Recognition, pages 59–66.

[Denker and LeCun, 1990] Denker, J. S. and LeCun, Y. (1990). Transforming neural-net output levels to probability distributions. In Proceedings of the 3rd International Conference on Neural Information Processing Systems, pages 853–859.

[Flexer and Lallai, 2019] Flexer, A. and Lallai, T. (2019). Can we increase inter-and intra-rater agreement in modeling general music similarity?. In Proceedings of the 20th annual conference of the International Society for Music Information Retrieval, pages 494–500.

[Gal and Ghahramani, 2016] Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning, pages 1050–1059.

[Ghahramani, 2015] Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. Nature, 521(7553):452–459.

[Gunes and Pantic, 2010] Gunes, H. and Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions, 1(1):68–99.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, pages 1321–1330.

[Hüllermeier and Waegeman, 2021] Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 110(3):457–506.

[Hüllermeier and Waegeman, 2021] Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn., 110(3):457–506.

[Jain et al., 2021] Jain, M., Lahlou, S., Nekoei, H., Butoi, V., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. (2021). DEUP: direct epistemic uncertainty prediction. arXiv preprint, arXiv:2102.08501.

[Jogan and Stocker, 2014] Jogan, M. and Stocker, A. A. (2014). A new two-alternative forced choice method for the unbiased characterization of perceptual bias and discriminability. Journal of Vision, 14(3):20–20.

[Koo and Li, 2016] Koo, T. and Li, M. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2):155–63.

[Kramer et al., 2018] Kramer, R. S. S., Mileva, M., and Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. PLOS ONE, 13(8):1–17.

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

[Kumano and Nomura, 2019] Kumano, S. and Nomura, K. (2019). Multitask item response models for response bias removal from affective ratings. In Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction, pages 1–7.

[Kurdi et al., 2017] Kurdi, B., Lozano, S., and Banaji, M. R. (2017). Introducing the open affective standardized image set (OASIS). Behavior Research Methods, 49(2):457–470.

[Lang et al., 1997] Lang, P. J., Bradley, M. M., Cuthbert, B. N., et al. (1997). International affective picture system (IAPS): Technical manual and affective ratings. NIMH Center for the Study of Emotion and Attention, 1(3):39–58.

[Li and Ma, 2020] Li, H.-H. and Ma, W. J. (2020). Confidence reports in decision-making with multiple alternatives violate the bayesian confidence hypothesis. Nature Communications, 11(1):2041–1723.

[Lichtenstein et al., 1982] Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In Judgment Under Uncertainty: Heuristics and Biases., pages 306–334.

[Marmpena et al., 2018] Marmpena, M., Lim, A., and Dahl, T. S. (2018). How does the robot feel? perception of valence and arousal in emotional body language. Paladyn, Journal of Behavioral Robotics, 9(1):168–182.

[McKelvie, 1992] McKelvie, S. J. (1992). Does memory contaminate test-retest reliability? The Journal of General Psychology, 119(1):59–72.

[Mukhoti et al., 2021] Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., and Gal, Y. (2021). Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. arXiv preprint, arXiv:2102.11582.

[Müller et al., 2012] Müller, U., Kerns, K. A., and Konkin, K. (2012). Test–retest reliability and practice effects of executive function tasks in preschool children. The Clinical Neuropsychologist, 26(2):271–287.

[Nguyen et al., 2018] Nguyen, V.-L., Destercke, S., Masson, M.-H., and Hüllermeier, E. (2018). Reliable multi-class classification based on pairwise epistemic and aleatoric uncertainty. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, pages 5089–5095.

[Niculescu-Mizil and Caruana, 2005] Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning, pages 625–632.

[Ovadia et al., 2019] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 32.

[Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161.

[Russell, 2003] Russell, J. A. (2003). Core affect and the psychological construction of emotion. Psychological Review, 110(1):145–172.

[Senge et al., 2014] Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., and Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Information Sciences, 255:16–29.

[Shi et al., 2020] Shi, W., Zhao, X., Chen, F., and Yu, Q. (2020). Multifaceted uncertainty estimation for label-efficient deep learning. Advances in Neural Information Processing Systems, 33.

[Sim and Wright, 2005] Sim, J. and Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Physical Therapy, 85(3):257–268.

[Smith and Gal,] Smith, L. and Gal, Y. Understanding measures of uncertainty for adversarial example detection. In Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence, pages 560–569.

[Thulasidasan et al., 2019] Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. (2019). On mixup training: Improved calibration and predictive uncertainty for deep neural networks. Advances in Neural Information Processing Systems, 32.

[Truong et al., 2009] Truong, K. P., Van Leeuwen, D. A., Neerincx, M. A., and Jong, F. (2009). Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion. In Proceedings of Interspeech 2009, pages 2027–2030.

[Zadrozny and Elkan, 2001] Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In Proceedings of the 18th International Conference on Machine Learning, volume 1, pages 609–616.

[Zadrozny and Elkan, 2002] Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 694–699.

[Zhang et al., 2017] Zhang, L., Tan, J., Han, D., and Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discovery Today, 22(11):1680–1685.

[Zhao et al., 2019a] Zhao, S., Wang, S., Soleymani, M., Joshi, D., and Ji, Q. (2019a). Affective computing for large-scale heterogeneous multimedia data: A survey. ACM Transactions on Multimedia Computing, Communications, and Applications, 15(3s):1–32.

[Zhao et al., 2019b] Zhao, Z.-Q., Zheng, P., Xu, S.-t., and Wu, X. (2019b). Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(11):3212–3232.

[Zhou et al., 2021] Zhou, Y., Ishigaki, T., and Kumano, S. (2021). Deep explanatory polytomous item-response model for predicting idiosyncratic affective ratings. In Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction, pages 1–8.

Supplementary Material: Collision Probability Matching Loss for Disentangling Epistemic Uncertainty from Aleatoric Uncertainty

A Confidence upper-bounds collision probability

Confidence upper-bounds collision probability, namely:

$$\sum_c p_c^2 \leq p_{y_{max}}.$$

Confidence $p_{y_{max}}$ is the maximum probability of p, and thus $p_c \leq p_{y_{max}}$. Since $p_c \geq 0$ as a probability, it follows $p_c^2 \leq p_c \cdot p_{y_{max}}$. Therefore,

$$\begin{aligned}\sum_c p_c^2 &\leq \sum_c p_c \cdot p_{y_{max}} \\ &= 1 \cdot p_{y_{max}} \\ &= p_{y_{max}},\end{aligned}$$

where the first equation uses $\sum_c p_c = 1$.

B Impact of the weight of CPM loss w on the estimation of aleatoric uncertainty ϵ

To evaluate how estimated ϵ is close to the ground truth with difference w, we performed a simulation study in which label data was pseudo-generated. In the simulation, we first determined the ground truth of target distribution p. Therefore, we can directly measure ϵ by definition, i.e., $\sum_c (p_c - q_c)^2$. It makes possible to evaluate how the expected ϵ value under empirical data distribution ((4) in the body) was close to the ground truth ϵ. To obtain the true ϵ, we determine the ground truth distribution of human rating data by using a simple model.

We set the ground truth distribution p to be the predictive distribution of a graded response model, a family of item-response models, which has more commonly been used in psychological literature. We fitted the model using Stan programming language with the Hamiltonian Monte Carlo (HMC) to the original human response data (Dataset-REA) described in Sec.4. We next re-generated rating data for $N = 50$ persons and $M = 120$ images randomly sampled from a categorical distribution with p. Then, we randomly selected 25% of the images from M images for each person, and generated second rating for each image to calculate their test-retest reliability. We fitted the same three-layer neural network, which was used in Sec. 4, using the same CE+CPM loss. Specifically, the input layer had 85 nodes and took a 35-dimensional vector, including the intensities of 17 AUs and the presence of 18 AUs detected using OpenFace [Baltrusaitis et al., 2018], concatenated with a 50-dimensional one-hot vector, each element of which represents the target image, as $\boldsymbol{x}^{(m)}$. The middle layer had 4,608 nodes. The final layer was a softmax layer that outputs a $(C - 1)$-dimensional simplex, each element of which represents the probability of the corresponding label.

The model was trained by minimizing CE+CPM loss, i.e. the weighted sum of the standard cross-entropy loss and the respondent-wise squared difference between Cross-CP and Pred-CP; namely:

$$\mathrm{L}_1 = -\frac{1}{MN}\sum_m \sum_n \sum_c y_c^{(mn)} log(q_c^{(mn)}) + w\frac{1}{N}\sum_n (\text{Cross-CP}^{(n)} - \text{Pred-CP}^{(n)})^2, \tag{1}$$

where n is the index of respondent. We compared the differences due to the strength of the restraint conditions with $w = 10$ and $w = 10,000$, respectively.

The model was trained through the learning rate of the Adam optimizer with 2e-4. The mini-batch size was set to be maximum, i.e. 6,000.

Table 1 shows the results when $w = 10$ and $w = 10,000$. When $w = 10$, $\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$ are not close, whereas when $w = 10,000$, $\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$ are matched because of the strong constraints that bring $\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$ closer together. It was observed that MC was calibrated and get closer to MC when $\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$ were brought closer to each other. Here we compared the both ϵ values of the true distribution with the approximate ϵ which is the difference of $\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$. At $w = 10$, ϵ was 0.13 and approximate ϵ was 0.05, whereas at $w = 10,000$, ϵ and approximate ϵ were close, with values of 0.19 and 0.19, respectively. From the results, we showed that ϵ, or epistemic uncertainty, can be calculated with high accuracy by bringing $\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$ closer together.

C Post-hoc approach of collision probability matching using temperature scaling

C.1 Method

As an alternative method that can be applied to models already trained, we also propose post-hoc calibration method and evaluated if the method named CE→TS satisfies the CPM constraint post-hoc via the temperature scaling, which is commonly used for confidence matching [Guo et al., 2017]. This method consists of two stages. In the first stage, it was trained with only the cross-entropy loss, i.e., the first term of (1). In the second stage, the input of the softmax layer, $z \in \mathbb{R}^C$, was fed to the temperature scaling, $q = \sigma_{SM}(z/t)$, where σ_{SM} is the softmax function. The temperature that satisfies $\mathrm{Cross\text{-}CP}^{(n)} = \mathrm{Pred\text{-}CP}^{(n)}$ was found for each respondent n using a grid search.

C.2 Results

We observed that the collision probability of CPM-constrained predictive collision probability ($\mathrm{Pred\text{-}CP}$) did not actually exceed the test-retest reliability using CE→TS method on Dataset-SIM. Table 2 shows the comparison results of measures before and after CPM using CE→TS method and the results of CE+CPM method as reference. The measures obtained using the original predictive distribution of the prediction model ($\hat{q}$) and the collision-probability-matched one (q) on the left and right-hand sides of arrows in the results of CE→TS method.

CE→TS method actually yielded the corrected collision probability (0.41) below the test-retest reliability (0.61). Therefore, the epistemic uncertainty could be meaningfully measured as the difference between the test-retest reliability and the expected likelihood, i.e., 0.20 ($= 0.61 - 0.41$). Similarly, for arousal, the epistemic uncertainty was 0.20 ($= 0.55 - 0.35$). The uncorrected machine accuracy for arousal (0.54) was lower than that for valence (0.62) in line with affective computing literature (e.g. [Gunes and Pantic, 2010]), in which the two types of uncertainty have hardly been distinguished. Curiously, on the other hand, their epistemic uncertainties turned out to be comparable. This suggests that their difference in accuracy stemmed mainly from the aleatoric uncertainty, but not from the model itself. From another point of view, this result is reasonable since the structure of the model used to estimate both is the same.

w=10		w=10,000	
Full prob.-dist.	Max-prob.	Full prob.-dist.	Max-prob.
True-CP	TC	True-CP	TC
0.56	0.56	0.56	0.56
Cross-CP	MA	Cross-CP	MA
0.49	0.47	0.37	0.36
Pred-CP	MC	Pred-CP	MC
0.52	0.52	0.37	0.37

Table 1: Results of collision probability matching using simulated data depends on the difference of w, i.e., constraint strength for collision probability matching.

CE→TS method	
Full prob.-dist.	Max-prob.
True-CP	TC
0.61	
Cross-CP	MA
0.61 → **0.41**	0.62
Pred-CP	MC
0.94 → **0.41**	0.96 → 0.61

(a) Valence

CE→TS method	
Full prob.-dist.	Max-prob.
True-CP	TC
0.53	
Cross-CP	MA
0.55 → **0.35**	0.54
Pred-CP	MC
0.90 → **0.35**	0.93 → 0.53

(b) Arousal

Table 2: Results of collision probability matching: Uncorrected (left side of arrow) and corrected, in which $\text{Cross-CP} = \text{Pred-CP}$ (right side of arrow) in CE→TS method.

Valence		Arousal	
Full prob.-dist.	Max-prob.	Full prob.-dist.	Max-prob.
True-CP	TC	True-CP	TC
0.47	0.47	0.37	0.37
Cross-CP	MA	Cross-CP	MA
0.43	0.54	0.36	0.47
Pred-CP	MC	Pred-CP	MC
0.43	0.55	0.36	0.47

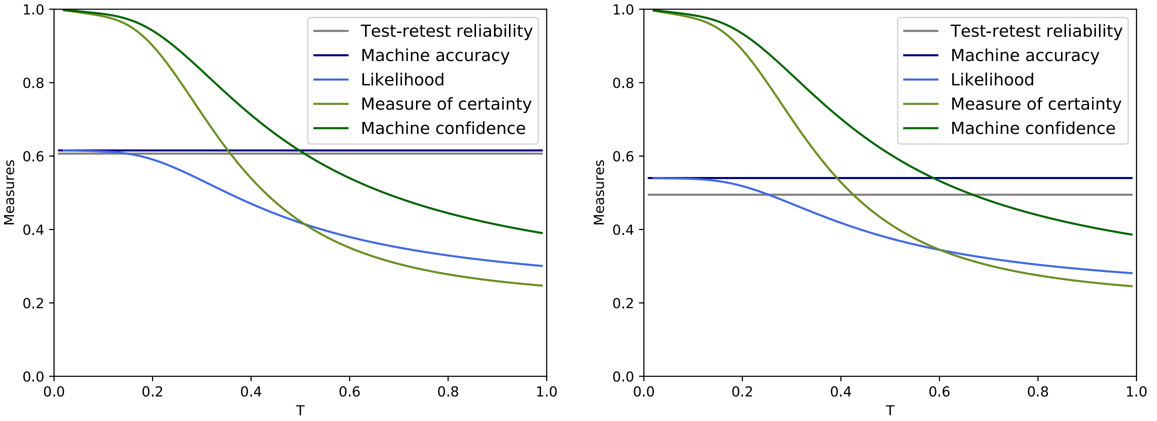
Table 3: Results of post-hoc collision probability matching using simulated data. The left two columns show the valence and the right shows the arousal results.

Cross-CP matched Pred-CP at $t = 0.51$ with the value of 0.41 in valence and at $t = 0.6$ with 0.35 in arousal, respectively (see Supplemental B). At the same temperature, MC was also calibrated and close MA for both dimensions. This seems reasonable, although not necessary, when its probability distribution variant, the cross-collision probability, is already calibrated. Notice that True-CP (test-retest reliability) is constant since it is model-independent, and the temperature scaling does not affect the machine accuracy [Guo et al., 2017].

Furthermore, for valence, the accuracy of 0.62 was higher than the test-retest reliability of 0.61, i.e., $\text{MA} \geq \text{True-CP}$. This could occur because $\text{TC} \geq \text{True-CP}$, as already proven in Section 3.1 and Supplemental A, and $\text{TC} \geq \text{MA}$ as explained in Section 3.5.

C.3 Problems of post-hoc calibration by comparing the ground truth ϵ

We also evaluate how close the estimated ϵ to the ground truth by pseudo-generated dataset as described in Section B. The estimated ϵ, ie, $\hat{\epsilon}$, is 0.058 while the ground truth ϵ is 0.039 for valence. The difference is 0.019 and it is significantly larger than the CPM Loss results. As well in arousal results, $\hat{\epsilon}$ is 0.020 while the ground truth ϵ is 0.039, and the difference (0.019) is also significant larger than the CPM Loss results. Therefore, this method is not an accurate measure of epistemic uncertainty.



(a) Valence

(b) Arousal

Figure 1: Five measures (in y axis) versus temperature t (in x axis) using CE→TS method. (a, b) on the *facial emotion judgement dataset*. For all emotional dimensions, i.e., valence, arousal, and emotional similarity, the proposed CPM and confidence calibration were achieved at $t = 0.51$ in valence and $t = 0.6$ in arousal namely, the light blue and green curves ($\mathrm{Cross\text{-}CP}$ and $\mathrm{Pred\text{-}CP}$) intersected (i.e., collision-probability-matched), and the dark blue line (MA) and the dark green curve (MC) intersected (i.e., confidence-matched). More importantly, the crossing point of the light blue and green curves was below the gray line showing the test-retest reliability ($\mathrm{True\text{-}CP}$), and their vertical difference, i.e., 0.20 $(= 0.61 - 0.41)$ for valence, and 0.20 $(= 0.55 - 0.35)$ for arousal, is considered the epistemic uncertainty of the model on the test data.

Figure 1 shows the five measures versus temperature T using CE→TS Model. In both valence and arousal dataset, although test-retest reliability score exceeds machine accuracy, we can see that the measure where measure of certainty and likelihood match is lower than the test-retest reliability score.

D Applicability to multi-labels

Some readers may think that multilabel approach will also work for the CPM when each multilabel is normalized to a unit sum vector. However, there appear to be problems in measuring the $\mathrm{True\text{-}CP}$. Suppose that the label for each class follows an independent Bernoulli distribution and that reliability=1 if the multilabel y has a single true label, and reliability$= 0$ otherwise. For binary classes, this definition makes it easy to prove that reliability equals $\mathrm{True\text{-}CP}$. However, for multiple classes ($C > 2$), they do not agree both mathematically and experimentally by simulation.

Verification: As long as the correct answer is a distribution, we can say that the discretization error goes up (information drops) in the order of multi-label and single-label [Jia et al., 2018]. But with multi-labels, it seems difficult to get the correct $\mathrm{True\text{-}CP}$ when there are more than 3 classes. If the generation of multi-labels from the true distribution follows a class-by-class independent Bernoulli process, multi-labels also have some properties similar to single labels with an increased number of observations. For example, a normalized histogram of an infinite number of samples will match the true distribution. However, it seems difficult to obtain test-retest reliability ($\mathrm{True\text{-}CP}$) from a single multi-label observation for more than two classes ($C > 2$). First, we would consider test-retest reliability $= 1$ when only one class is labeled, since it is consistent, and test-retest reliability $= 0$ otherwise. In the case of binary classes, $\mathrm{True\text{-}CP}$ can be obtained correctly with infinite samples (assuming $p = (p_1, 1 - p_1)$, we can represent probability of a label being $(1, 0) = p_1^2$ and probability of a label being $(0, 1) = (1 - p_1)^2$, meaning the sum of these probabilities, i.e. $p_1^2 + (1 - p_1)^2$, equal to $\sum_c p_c^2$, the definition of $\mathrm{True\text{-}CP}$. However, when $C > 2$, such calculation does not necessarily match $\sum_c p_c^2$. To validate it in analytically, we ran a simple simulation with a large number of true distributions p generated from Dirichlet distribution. We observed that this method is often overvalued for three or more classes. After creating 100 Dirichlet distributions and calculating the test-retest reliability, the mean absolute error for the binary class was small (0.004 ± 0.004 (SD)), while the mean absolute error for the multi-class was much larger (0.08 ± 0.06).

References

[Baltrusaitis et al., 2018] Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L. (2018). Openface 2.0: Facial behavior analysis toolkit. In 13th IEEE International Conference on Automatic Face Gesture Recognition, pages 59–66.

[Gunes and Pantic, 2010] Gunes, H. and Pantic, M. (2010). Automatic, dimensional and continuous emotion recognition. International Journal of Synthetic Emotions, 1(1):68–99.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning, pages 1321–1330.

[Jia et al., 2018] Jia, X., Li, W., Liu, J., and Zhang, Y. (2018). Label distribution learning by exploiting label correlations. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32.