
Legitimate ground-truth-free metrics for deep uncertainty classification scoring

Arthur Pignet
Owkin Inc.

Chiara Regniet
Owkin Inc.

John Klein
Owkin Inc.

Abstract

Despite the increasing demand for safer machine learning practices, the use of Uncertainty Quantification (UQ) methods in production remains limited. This limitation is exacerbated by the challenge of validating UQ methods in absence of UQ ground truth. In classification tasks, when only a usual set of test data is at hand, several authors suggested different metrics that can be computed from such test points while assessing the quality of quantified uncertainties. This paper investigates such metrics and proves that they are theoretically well-behaved and actually tied to some uncertainty ground truth which is easily interpretable in terms of model prediction trustworthiness ranking. Equipped with those new results, and given the applicability of those metrics in the usual supervised paradigm, we argue that our contributions will help promoting a broader use of UQ in deep learning.

1 INTRODUCTION

Uncertainty quantification (UQ) is a research topic in machine learning that is gaining traction due to the increasing need for alignment with reliability requirements arising from safety critical application fields such as healthcare or autonomous driving. However, it is often overlooked and even seldom used in production. We argue that one of the main reasons behind this disparity is the lack of connections between uncertainty performance metrics and interpretable uncertainty "ground truth". Filling this gap would allow machine learning practitioners to better know to what extent the pre-

dicted uncertainty is actionable and this paper is an attempt to pave the way toward this goal.

To meet this ambition, a first impediment is that there is no universally agreed upon definition of uncertainty (Lahlou et al., 2021) which impairs the formalization of uncertainty quantification ground truth. This work aligns with Kirchhof et al. (2023) and regards uncertainty quantification as an estimation problem of the trustworthiness $s(\mathbf{x})$ of a model's prediction for any input $\mathbf{x} \in \mathcal{X}$. The function s is referred to as the (uncertainty) scoring function in the sequel.

Equipped with this vision, uncertainty quantification ground truth can be regarded as a function s^* that is interpretable and objectively informs on the correctness of the model predictions. Let (X, Y) denote an input-target pair of random variables governed by the data generative distribution $p_{X,Y}$. A realization of (X, Y) is denoted by $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$. For convenience, we will use interchangeably $P(X = \mathbf{x}, Y = y)$ and $p_{X,Y}(\mathbf{x}, y)$. We focus on binary classification tasks and $\mathcal{Y} = \{0, 1\}$.

For classification problems, the minimizer of the posterior expected 0-1 loss is the Bayes classifier:

$$\hat{y}_{\text{bay}}(\mathbf{x}) = \arg \max_{y \in \mathcal{Y}} P(Y = y | X = \mathbf{x}), \forall \mathbf{x} \in \mathcal{X}. \quad (1)$$

If a training algorithm would succeed in learning the true conditional distribution of class labels given inputs, one would automatically gain access to an objective way to quantify how certain is the correctness of the model prediction $\hat{y}_{\text{bay}}(\mathbf{x})$. Indeed, we have

$$\text{err}(\mathbf{x}; \hat{y}_{\text{bay}}, \boldsymbol{\pi}) = P(Y \neq \hat{y}_{\text{bay}} | X = \mathbf{x}) \quad (2)$$

$$= 1 - \pi_{\hat{y}_{\text{bay}}}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}, \quad (3)$$

where $\boldsymbol{\pi}(\mathbf{x})$ is such that $Y|X = \mathbf{x} \sim \text{Mult}(\boldsymbol{\pi}(\mathbf{x}))$.

The above scoring function is often referred to as the Bayes error and captures aleatoric uncertainty, i.e. the irreducible uncertainty carried by $p_{X,Y}$. As such, it is a legitimate ground truth s^* to quantify the uncertainty

of the Bayes classifier. Note that the UQ question is trivially solved in this ideal case since the Bayes classifier leverages the very probabilities $P(Y = y|X = \mathbf{x})$, from which this ground truth is obtained.

Deep learning leverages probabilistic classifiers $\hat{\pi}$, where $\hat{\pi}(\mathbf{x})$ typically represents a vector of probabilities generated by a softmax layer, serving as an approximation of $\pi(\mathbf{x})$. For input \mathbf{x} , the predicted class label is $\hat{y} = \arg \max_{y \in \mathcal{Y}} \hat{\pi}_y(\mathbf{x})$. In practice, due to epistemic uncertainty, $\hat{\pi}$ deviates from π . The gap between $\hat{\pi}(\mathbf{x})$ and the conditional distribution $p_{Y|\mathbf{x}}$ can be substantial depending on the amount of training data and modeling hypotheses. Based on this observation, two simple candidates for s^* arise from equation (3) in the context of UQ for a trained model $\hat{\pi}$:

- By substituting classifier \hat{y}_{bay} with classifier \hat{y} , we obtain the probability of misclassification

$$\phi(\mathbf{x}) := \text{err}(\mathbf{x}; \hat{y}, \pi) \quad (4)$$

$$= P(Y \neq \hat{y} | X = \mathbf{x}) \quad (5)$$

$$= 1 - \pi_{\hat{y}}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}. \quad (6)$$

- By substituting the actual probabilities with softmax entries, we obtain an assessment of the commitment of the model to non-optimal decisions

$$\varphi(\mathbf{x}) := \text{err}(\mathbf{x}; \hat{y}_{\text{bay}}, \hat{\pi}) \quad (7)$$

$$= 1 - \hat{\pi}_{\hat{y}_{\text{bay}}}(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}. \quad (8)$$

Even endowed with definitions (6) and (8), a major difficulty in UQ is the impossibility to evaluate those in real-world datasets or uncertainty ground truth in general (Kong et al., 2023). In the supervised learning paradigm, only (\mathbf{x}, y) pairs are available, leaving the practitioner with no choice but to compute metrics using that test data. This paper focuses on the following angle in UQ evaluation: a purely metric based assessment of a UQ scoring function is valid if the metrics are maximized by one of the ground truth definitions (6) and (8). We prove that one widely used metric in UQ benchmark is maximized by (6). We also introduce a new metric that is maximized by (8).

To sum up, our contributions are the following:

- We introduce a new metric, UQ-C-INDEX, which is maximized by a different scoring ground truth and offers an alternative standpoint for UQ evaluation.
- We prove that UQ-AUC and UQ-C-INDEX are actionable tools to identify subsets of data points for which the model reliability is higher.
- We provide an empirical evaluation of the suggested metrics on synthetic data with known scoring ground truth.
- We demonstrate the flexibility and minimalism of the setting in which those metrics can be used in production on real datasets.

The investigated metrics can be generalized to the multi-class case through the usual one-versus-rest extensions.

1.1 Scope

The literature around UQ in deep learning is vast and the following paragraphs explain which part of this literature do not fall within the scope of the present paper.

This paper is not a UQ benchmark Since many UQ techniques were introduced (see Section 6 for an overview), several studies were conducted to benchmark them (Ovadia et al., 2019; Galil et al., 2023b,a; Kirchhof et al., 2023). Those benchmarks are insightful as they not only illustrate the limitations of UQ techniques but also offer guidance to more optimally leverage UQ techniques. We clarify that the present paper is not a benchmark. Our objective is to investigate uncertainty estimator metrics, some of which are actually used in the above cited benchmarks, from a theoretical standpoint. By legitimizing a widely used metric and introducing a new one, our contributions reinforce the conclusions drawn in past benchmarks and provide insights for future ones.

This paper does not deal with OOD data There are two dominant use cases showcased in deep learning UQ articles: (i) ranking inputs based on classification error risk (ii) identifying out-of-domain (OOD) inputs. The former is a legitimate expectation from a well-behaved quantification, grading each model prediction likeliness based on patterns identified in inputs. The rationale behind the latter is that OOD data exhibit peculiar patterns compared to training data and have low likelihood under $p_{X,Y}$. Consequently, data uncertainty, i.e. epistemic uncertainty arising from limited training sample size, should be high in the vicinity of OOD data.

- We provide theoretically grounded evidence that UQ-AUC - a metric used in most UQ benchmarks - is a legitimate way to assess the correctness of the quantified uncertainties even in absence of identified uncertain subsets of data points or distributional hypotheses.
- We shed light on important connections between UQ-AUC and other related prior arts.

Input ranking is the focus of this paper so as to fit to the classical ML setting where test data are iid with respect to the training data distribution $p_{X,Y}$. Indeed, the UQ metrics that we examine in this paper only require samples from the data generative distribution to be computed, which is always possible on real data by resorting to cross-validation. Conversely, the OOD detection use case requires extra OOD data that is not always at hand. However, if some are, elements of the proposed metric-based evaluation scheme remain applicable and this is discussed in Section 16.

2 SHEDDING THEORETICAL LIGHT ON METRICS FOR UQ EVALUATION

The present section first provides arguments further legitimizing functions ϕ and φ as relevant uncertainty scoring ground truths. It then examines a known scoring function metric from the literature and sheds light on its ties to ground truth ϕ in spite of requiring only access to test data points for computation. It also introduces a new metric enjoying similar properties, but this time being tied to ground truth φ . Note that while Wimmer et al. (2023) examined the properties of scoring functions themselves, we examine the properties of metrics used to evaluate scoring functions.

The metrics analyzed in this section are meant to evaluate the ability of a scoring s to rank inputs \mathbf{x} from least to most likely to lead to a deviation of $\hat{y}(\mathbf{x})$ from $\{Y|X = \mathbf{x}\}$ or $\hat{y}_{\text{bay}}(\mathbf{x})$.

To investigate the goodness of ranking, it is useful to define, within the set \mathcal{M}_X of p_X -measurable functions, the equivalence class \mathcal{E}_s of scorings that almost surely rank inputs the same way as s :

$$\mathcal{E}_s := \{s' \in \mathcal{M}_X | P(\{s(X) \geq s(X')\} \cap \{s'(X) < s'(X')\}) = P(\{s(X) < s(X')\} \cap \{s'(X) \geq s'(X')\}) = 0\}. \quad (9)$$

2.1 Validity of ϕ and φ to filter inputs based on uncertainty risks

Before unveiling how UQ metrics can assess the closeness between a scoring function s and either ground truth ϕ or φ , let us first legitimize those ground truths in the context of input scoring based on predictive uncertainty. We propose to meet this objective by proving that sub-level sets of functions ϕ and φ can gradually filter out inputs mapped to incorrect class labels (in some sense to be formalized below).

In other words, we check if the investigated ground truth are actionable to define input space regions in which risk can be controlled.

Let $\mathcal{C}_{\text{mce}}(\mathcal{E}_s)$ denote the set of values achieved by the **misclassification error** $\text{MCE}_s(\beta) = P(Y \neq \hat{Y} | s(X) \leq \beta)$, $\forall \beta \in \mathbb{R}$ and $\mathcal{B}_{\text{mce}}(s) = \{\min(\text{MCE}_s^{-1}(c)) | c \in \mathcal{C}_{\text{mce}}(\mathcal{E}_s)\}$ the set of minima of pre-images of MCE_s . The following result provides insights on the capability of sets to filter out inputs for which model predictions are unreliable.

Theorem 2.1. $\forall s \in \mathcal{E}_\phi, \exists \psi_s$ a strictly increasing mapping from $\mathcal{C}_{\text{mce}}(\mathcal{E}_\phi)$ to $\mathcal{B}_{\text{mce}}(s)$ such that

$$P(Y \neq \hat{Y} | s(X) \leq \psi_s(\gamma)) = \gamma, \quad (10)$$

$$\forall \gamma \in \mathcal{C}(\mathcal{E}_\phi) \subset (0; 1 - \text{ACC}].$$

In the above theorem ACC stands for the accuracy of predictor \hat{y} . A proof of this result is given in Section 9. A similar result for ground truth φ follows. Let us introduce $\mathcal{C}_{\text{mbc}}(\mathcal{E}_s)$ the set of values achieved by the **misalignment with the Bayes classifier** $\text{MBC}_s(\beta) = P(\hat{Y}_{\text{bay}} \neq \hat{Y} | s(X) \leq \beta)$, $\forall \beta \in \mathbb{R}$ and $\mathcal{B}_{\text{mbc}}(s) = \{\min(\text{MBC}_s^{-1}(c)) | c \in \mathcal{C}_{\text{mbc}}(\mathcal{E}_s)\}$.

Theorem 2.2. $\forall s \in \mathcal{E}_\varphi, \exists \psi_s$ a strictly increasing mapping from $\mathcal{C}_{\text{mbc}}(\mathcal{E}_\varphi)$ to $\mathcal{B}_{\text{mbc}}(s)$ such that

$$P(\hat{Y}_{\text{bay}} \neq \hat{Y} | s(X) \leq \psi_s(\gamma)) = \gamma, \quad (11)$$

$$\forall \gamma \in \mathcal{C}_{\text{mbc}}(\mathcal{E}_\varphi) \subset (0; P(\hat{Y}_{\text{bay}} \neq \hat{Y}))].$$

A proof of this result is given in Section 10.

Simply put, these theorems state that ground truth scoring functions can span input space regions where the misclassification/misalignment probabilities can be confined to a prescribed value. This is arguably one of the most desirable outcome of UQ. The input space regions at hand are the sub-level sets $\mathcal{L}_\beta(s) := \{\mathbf{x} \in \mathcal{X} | s(\mathbf{x}) \leq \beta\}$ of a scoring function s . To better understand the implication of Theorem 2.1, one can also make connections with set-valued classifiers such as those issued by the conformal prediction framework (Lei and Wasserman, 2015; Vovk et al., 2005). Indeed, let us introduce the following set-valued classifier

$$\mathcal{C}_\beta(\mathbf{x}; s) = \begin{cases} \hat{y}(\mathbf{x}) & \text{if } \mathbf{x} \in \mathcal{L}_\beta(s) \\ \mathcal{Y} & \text{otherwise} \end{cases}. \quad (12)$$

Then Theorem 2.1 says that, for any $s \in \mathcal{E}_\phi$, the set-valued classifier $\mathcal{C}_\beta(\cdot; s)$ can achieve any coverage $\gamma \in \mathcal{C}_{\text{mce}}(\mathcal{E}_s)$ by varying β , i.e. $P(Y \in \mathcal{C}_\beta(X; s)) = \gamma$. This is also true for φ but with a different flavor of coverage defined as $P(\hat{Y}_{\text{bay}} \in \mathcal{C}_\beta(X; s))$. Just like for conformalized predictors, coverage is not enough and one must check that, for different values β , sub-level

sets $\mathcal{L}_\beta(s)$ capture a large amount of inputs which is equivalent to have set-valued predictors $\mathcal{C}_\beta(\cdot; s)$ with small expected cardinalities. This point will be addressed for each ground truth in the next subsections.

It can also be noted that (10) is reminiscent of confidence calibration. An important difference is that calibration is less flexible in the sense that it holds for a unique scoring function whereas (10) holds for infinite families of scorings. Moreover, the calibration property involves a conditioning on level sets of some the confidence score. Note that one could build a nested family of sets from a calibrated classifier to achieve (10) at the unavoidable cost of re-indexing.

2.2 Checking that scoring functions match misclassification probabilities

In practice, one has only access to a labeled (in-distribution) test dataset $\mathcal{D}_{\text{test}}$ of size n_{test} sampled from the i.i.d. random variables $(Y^{(i)}, X^{(i)})_{1 \leq i \leq n_{\text{test}}}$. To evaluate the ability of a scoring function $s(\mathbf{x})$ to align with misclassification probability $\phi(\mathbf{x})$, it would be instrumental to leverage misclassification probabilities $\phi(\mathbf{x}^{(i)})$ of the test points but those probabilities cannot be observed. One can thus wonder to what extent misclassification errors are a good proxy of $\phi(\mathbf{x}^{(i)})$. Let us denote the misclassification function as follows:

$$\text{mis}(\mathbf{x}^{(i)}, y^{(i)}; \hat{y}) = \mathbb{1}_{y^{(i)} \neq \hat{y}(\mathbf{x}^{(i)})}. \quad (13)$$

One can then leverage misclassification errors to check if they match score s by checking that,,

$$\begin{aligned} \text{mis}(\mathbf{x}^{(i)}, y^{(i)}; \hat{y}) &\leq \text{mis}(\mathbf{x}^{(j)}, y^{(j)}; \hat{y}) \\ \Leftrightarrow s(\mathbf{x}^{(i)}) &\leq s(\mathbf{x}^{(j)}), \forall i \neq j. \end{aligned} \quad (14)$$

Since misclassification errors are binary values, it is only worth checking the above equivalence for (i, j) pairs such that $\text{mis}(\mathbf{x}^{(i)}, y^{(i)}; \hat{y}) \neq \text{mis}(\mathbf{x}^{(j)}, y^{(j)}; \hat{y})$. In essence, this evaluation calculates the empirical probability that a correctly predicted instance has a lower uncertainty score than an incorrectly predicted one when both are chosen at random. ROC-AUC computes exactly empirical probabilities of this kind. We thus call this UQ metric the uncertainty quantification AUC and when $n_{\text{test}} \rightarrow \infty$, it writes:

$$\begin{aligned} \text{UQ-AUC}(s) &:= P\left(s(X^{(i)}) \leq s(X^{(j)}) \mid \text{mis}(X^{(i)}, Y^{(i)}; \hat{y}) \leq \text{mis}(X^{(j)}, Y^{(j)}; \hat{y})\right). \end{aligned} \quad (15)$$

This metric is precisely the one used in (Gal and Ghahramani, 2016; Hendrycks and Gimpel, 2016). A

first important theoretical result is that UQ-AUC is maximized for any score s in the equivalence class of the misclassification probability.

Theorem 2.3 (Maximizing UQ-AUC). *A scoring function s^* is a maximizer of UQ-AUC iff s^* belongs to the equivalence class \mathcal{E}_ϕ of ground truth ϕ .*

A proof of this result, which builds upon (Cl  men  on et al., 2008, Proposition 1), is given in Section 11. What is particularly interesting with this result is that it holds for the actual misclassification probability ground truth without requiring access to misclassification probabilities $\phi(\mathbf{x}^{(i)})$ for its computation.

As mentioned in the previous subsection, the legitimacy of ϕ as an uncertainty scoring ground truth also relies on its ability to produce sub-level sets $(\mathcal{L}_\beta(\phi))_{\beta \in (0;1]}$ with a large probability to contain realizations of X . To be able to compare scoring functions in this regard, let us formalize this notion into the following preorder \prec_{mce} : let s and s' be two scoring functions for which (10) holds,

$$\begin{aligned} s \prec_{\text{mce}} s' &\Leftrightarrow \mathcal{L}_{\psi_s(\gamma)}(s) \subseteq \mathcal{L}_{\psi_{s'}(\gamma)}(s'), \forall \gamma \\ &\text{and } \mu(\mathcal{R}_{ss'}) > 0 \end{aligned} \quad (16)$$

where $\mathcal{R}_{ss'} = \{\gamma \in \mathcal{C} \mid P(X \in \mathcal{L}_{\psi_s(\gamma)}(s) \setminus \mathcal{L}_{\psi_{s'}(\gamma)}(s')) \mid Y = \hat{Y}\} > 0\}$ and μ is the Lebesgue measure. It can be proved that no other scoring s for which the condition of Theorem 2.1 hold can uniformly dominate a member s^* of \mathcal{E}_ϕ .

Theorem 2.4. *For any $s^* \in \mathcal{E}_\phi$ and any other scoring s for which (10) holds we have $s^* \not\prec_{\text{mce}} s$.*

A proof of this result is given in Section 12 which was the missing piece to fully establish the legitimacy of ϕ and as an uncertainty scoring ground truth.

2.3 Checking that scoring functions catch misalignment with labels predicted by the Bayes classifier

In the following paragraphs, we introduce a new metric for scoring functions which is meant to catch inputs for which high softmax scores were assigned to non-optimal class labels, i.e. different from \hat{y}_{bay} . This new metric relies on a misclassification gap defined as

$$\Delta(\hat{\pi}(\mathbf{x}), y) := 1 - \hat{\pi}_y(\mathbf{x}). \quad (17)$$

By substituting function mis with function Δ in (15),

a new metric called UQ-C-index¹ is obtained:

$$\text{UQ-C-index}(s) := P\left(s\left(X^{(i)}\right) \leq s\left(X^{(j)}\right) \middle| \Delta\left(\hat{\pi}\left(X^{(i)}\right), Y^{(i)}\right) \leq \Delta\left(\hat{\pi}\left(X^{(j)}\right), Y^{(j)}\right)\right). \quad (18)$$

Interestingly, this newly introduced metric is maximized by any member of the equivalence class of ground truth φ as stated in the following theorem.

Theorem 2.5 (Maximizing **UQ-C-index**). *Suppose $p_{\hat{\Pi}} = \hat{\pi} \# p_X$ has no atom, where $\hat{\pi} \# p_X$ is the push-forward of p_X by $\hat{\pi}$. Then, a scoring function s^* is a maximizer of UQ-C-index iff s^* belong to the equivalence class \mathcal{E}_φ of ground truth φ .*

See Section 13 for a proof. This result shows that maximizing UQ-C-index is focused on checking the alignment between model probabilistic predictions with the predictions of the (optimal) Bayes classifier.

Similarly as for the previous metric, let us introduce preorder \prec_{mbc} : let s and s' be two scoring functions for which (11) holds,

$$s \prec_{\text{mbc}} s' \Leftrightarrow \mathcal{L}_{\psi_s(\gamma)}(s) \subseteq \mathcal{L}_{\psi_{s'}(\gamma)}(s'), \forall \gamma \text{ and } \mu(\mathcal{R}_{ss'}) > 0. \quad (19)$$

Observe that although the definition of $\mathcal{R}_{ss'}$ is the same as for pre-order \prec_{mce} , the two pre-orders are different because the functions ψ_s and $\psi_{s'}$ involved in their respective definitions arise from property (11) instead of (10). It can be proved that the sub-level sets spanned by members of \mathcal{E}_φ exhibit a form of optimality stated in the following theorem.

Theorem 2.6. *Suppose $p_{s^*, \Delta} = (s^*, \Delta) \# p_{X,Y}$ and $p_{s, \Delta} = (s, \Delta) \# p_{X,Y}$ have no atom. For any $s^* \in \mathcal{E}_\varphi$ and any other scoring s for which (11) holds we have $s^* \not\prec_{\text{mbc}} s$.*

A proof of theorem Theorem 2.6 can be found in Section 14. This theorem was the missing piece to fully establish the legitimacy of φ and as an uncertainty scoring ground truth alternative.

3 RELATED WORK

If the role of an uncertainty scoring function is to provide insights regarding a model’s trustworthiness, how can one be sure that scoring functions themselves are

trustworthy? ML practitioners would typically investigate this question on a held-out dataset as part of an evaluation phase. Unfortunately, assessing the quality of quantified uncertainties appears to be more challenging than other ML predictions as there is no genuine ground truth in the form of (6) or (8) which is mainly why no consensus on UQ metrics has emerged in the literature (Ovadia et al., 2019). This performance evaluation challenge was addressed in prior arts (sometimes depending on the sought use case: input ranking or OOD detection). The following paragraphs provide an overview of these related works.

One line of UQ validation leverages a companion task that aims at leveraging meaningful signals in an uncertainty score. One such task is uncertainty sampling for active learning (Gal et al., 2017; Nguyen et al., 2022). If a UQ method correctly selects data points with high predictive uncertainty, revealing their labels should provide a strong supervisory signal and allow to maximize accuracy increase. At each active learning round, Gal et al. (2017) select the input that maximizes an uncertainty scoring function, add it to the training set and re-train the model. When using the total entropy, mutual information, or variation ratio function, the test accuracy increases at a pace significantly faster compared to random selection of the next input to insert in the training set. Nguyen et al. (2022) conducted a broader range of similar experiments and show that, depending on the model class, uncertainty scoring functions capturing epistemic uncertainty rather than aleatoric uncertainty can be far better acquisition functions. In a similar spirit, UQ scoring can be used as feedback for the initial task. Mortier et al. (2023) build weighted ensembles where the weights of the ensemble members are computed from the scoring issued by the UQ technique. If a UQ technique is well behaved, its corresponding ensemble should promote reliable classifiers and achieve better classification performances. However, evaluation in this setting requires the uncertainty scores to be well calibrated. In contrast, this paper addresses the evaluation of scores regardless of their calibration performances. Reinforcement learning can also typically benefit from feedback and is used for validation in (Gal and Ghahramani, 2016). Specifically, in situations where the Q-value function is approximated by a neural network, predictive uncertainty can be leveraged to perform Thomson sampling in lieu of an ϵ -greedy search. The higher the expected cumulative reward, the better the uncertainty estimates.

A rigorous way to check if quantified uncertainties align with errors is to rely on controlled experiments that typically involve a physical system (Caldeira and Nord, 2020). Depending on the system in question, sources of aleatoric and/or epistemic uncertainty can

¹In general, the concordance index also takes into account censorship but in the current use case there is no such censorship. This terminology is meant to emphasize that the metric relies on continuous functions.

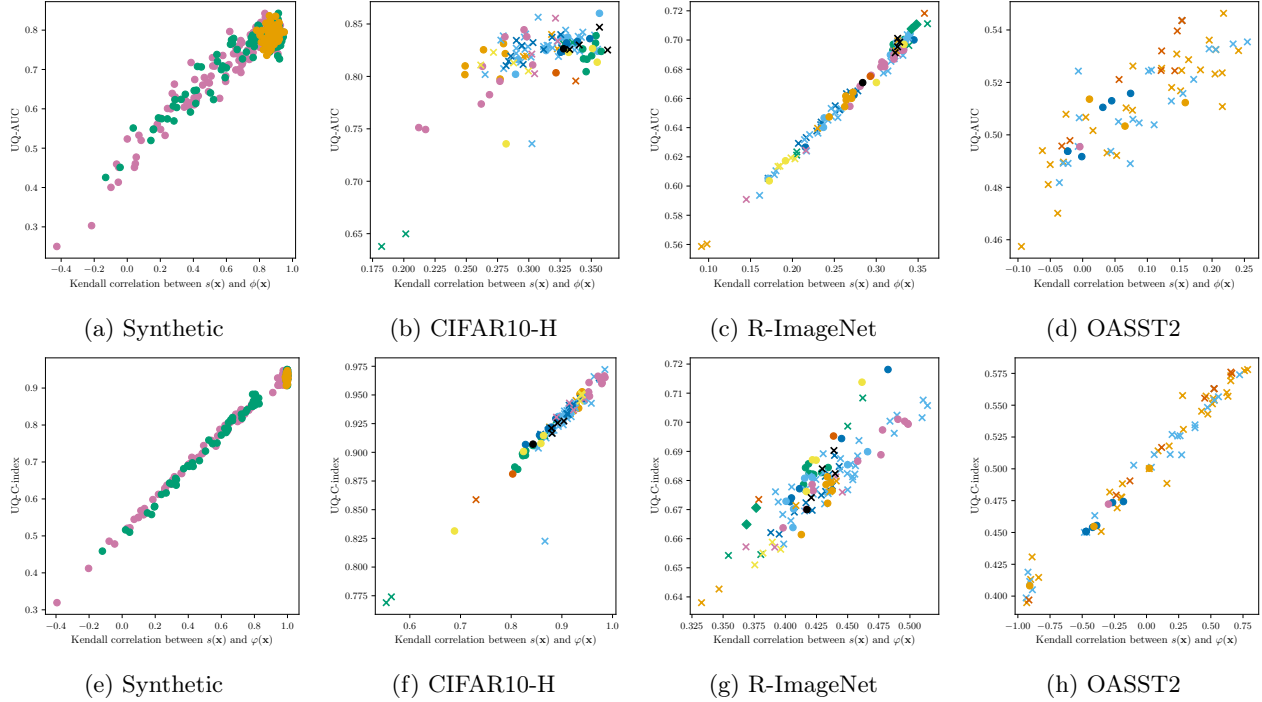


Figure 1: Synthetic dataset and real-world datasets with human annotations. Uncertainty ground truth-free metric UQ-AUC (Top row, resp. UQ-C-index in the bottom row) versus ground truth dependent Kendall correlation between $s(\mathbf{x})$ and $\phi(\mathbf{x})$, (resp. between $s(\mathbf{x})$ and $\varphi(\mathbf{x})$) for several UQ algorithms and hyperparameter configurations. All metrics are computed on a held-out test set. Note that Kendall correlations are intractable in practice, contrary to the UQ-AUC or UQ-C-index..

Models used for Synth: Entropy \bullet , Deep Ensemble \bullet , Monte-Carlo Dropout \bullet ,

Models used for CIFAR10 and R-ImageNet: GoogleNet \bullet , AlexNet \times , ResNet \bullet , RegNet \times , Swin \bullet , ResNeXt \times , MobileNet \bullet , EfficientNet \times , ViT \bullet , Wide ResNet \times , MnasNet \bullet , ConvNeXt \times , Inception \bullet , DenseNet \times , VGG \bullet , ShuffleNet \times ,

Models used for OASST2: BART \times , DistilBERT \bullet , DeBERTa \times , RoBERTa \bullet , BERT \times , XLM \bullet , MiniLM \times , E5 \bullet , mContriever \bullet , Other \times

Table 1: Pearson correlations between UQ metrics and ground-truth dependent Kendall coefficients κ . Each correlation value is followed by lower and upper 95% confidence intervals, obtained using Fisher’s z-transformation. Significantly better correlation is shown in bold letters.

Dataset	UQ metric	Correl. with $\kappa(s; \phi)$	Correl. with $\kappa(s; \varphi)$
Synth.	UQ-AUC(s)	94.56% (93.36;95.55)	93.44% (92.00;94.3)
	UQ-C-index(s)	98.81% (98.54;99.03)	99.62% (99.53;99.69)
CIFAR10-H	UQ-AUC(s)	65.18% (53.99;74.75)	54.01% (39.33;66.00)
	UQ-C-index(s)	-0.04% (-22.41;14.83)	95.61% (93.67;96.97)
R-ImageNet	UQ-AUC(s)	99.56% (99.36;99.69)	39.42% (23.47;53.75)
	UQ-C-index(s)	40.20% (23.56;54.56)	88.07% (83.17;91.63)
OASST2	UQ-AUC(s)	82.50% (73.10; 88.82)	71.69% (57.84; 81.53)
	UQ-C-index(s)	74.01% (61.04; 83.12)	98.88% (98.19; 99.31)

be varied which will impact the learnt model emulating the physical system. In a sensitivity analysis spirit, one may check that a scoring function varies monotonically with each source and compute a correlation metric. However this objective evaluation remains grounded with a single regression task and does not reach the genericity objective to evaluate quantified uncertainties on arbitrary data and classification tasks.

In spite of their appeal, the above cited elements of the literature do not possess the sought level of genericity allowing for uncertainty estimator evaluation solely based on (\mathbf{x}, y) pairs. Indeed all these methods are indirect measures of uncertainty scores, strongly dependent on the chosen backbone algorithm. In the next paragraph, we present metric based evaluation prior arts that comply with this requirement.

A popular way to check the legitimacy of predicted probabilities is to resort to calibration metrics (Nemani et al., 2023; Wimmer et al., 2023; Jospin et al., 2022; Henne et al., 2020; Ovadia et al., 2019). By definition, obtaining calibrated probabilities is a much stronger requirement than obtaining a score that correctly ranks unreliable predictions. Indeed, suppose one wishes to rank unreliable predictions thanks to calibrated softmax probabilities, then any composition of those probability with an increasing function will rank inputs in the same way. This invariance property of the investigated metrics compared to the expected calibration error is further highlighted in Section 8.

When assessing whether quantified uncertainties align with classification errors, many authors converged to a similar setting: leverage a scoring function and evaluate the latter as if it was meant to solve another binary classification task in which inputs are mapped to either $\{Y = \hat{y}|\mathbf{x}\}$ or $\{Y \neq \hat{y}|\mathbf{x}\}$ (Henne et al., 2020; Gal and Ghahramani, 2016; Hendrycks and Gimpel, 2016). Given a held-out test dataset it is possible to construct those misclassification labels. Finally, ROC-AUC can readily be used, thereby yielding the UQ-AUC metric formally defined in the previous section. Similarly, in the OOD detection use case, provided that some OOD inputs can be obtained, ROC-AUC of a UQ scoring with respect to in-distribution versus OOD labels can be computed (Gal and Ghahramani, 2016; Hendrycks and Gimpel, 2016; Lakshminarayanan et al., 2017)

Finally, the closest elements of the literature compared to this paper are (Geifman et al., 2019) and (Henne et al., 2020), each of them introducing a UQ metric. The metric proposed in (Geifman et al., 2019) uses a threshold on scoring functions to filter out low score data and build a subset of higher score data. For each such subset, one can compute its size fraction (compared to the entire subset) and an accuracy. By

browsing several threshold values, a curve is built and the metric value is the AUC of this curve. In the sequel, we not only prove that, for any fixed predictor \hat{y} , this metric is in bijection with UQ-AUC (thereby essentially grasping the same information as the latter) but also that it benefits from the same results as UQ-AUC. A similar remark holds for the work of Henne et al. (2020) who compare UQ techniques based on a curve whose AUC is proportional to the UQ-AUC.

3.1 Metric comparison synthesis

In the following paragraphs, the main differences between metrics investigated in the previous subsections and other closely related works are highlighted. We denote the metric from (Geifman et al., 2019) G-AUC and the one arising from the work of (Henne et al., 2020) H-AUC. Let us first formally establish the bonds of these AUCs with UQ-AUC.

Lemma 3.1. *For any scoring function s and fixed predictor \hat{y} with accuracy ACC , we have*

$$G-AUC(s) = (1 - ACC)^2 + 2(1 - UQ-AUC(s))(1 - ACC)ACC, \quad (20)$$

$$H-AUC(s) = UQ-AUC(s)(1 - ACC)ACC. \quad (21)$$

The above lemma clarifies important connections between prior arts. It is clear that UQ benchmarks relying on either UQ-AUC, G-AUC or H-AUC will lead to identical conclusions. This also confirms the novelty of metric UQ-C-index compared to prior arts. It can be argued that the influence of the accuracy of \hat{y} in the definitions of the G-AUC and H-AUC metrics make them harder to interpret than UQ-AUC. For example, the maximal value of those two metrics also depends on ACC and thus cannot be used as universal reference.

Another important point to highlight are the different shades of scoring performance that UQ-AUC and UQ-C-index promote respectively. Those are in fact rooted in the definitions of the uncertainty ground truth to which each metric is tied and the different validity properties they enjoy (Theorem 2.1 and Theorem 2.2 respectively). While UQ-AUC promotes scoring functions allowing to secure input space regions where the trained model issues correct class predictions, UQ-C-index promotes scoring functions allowing to secure input space regions where the trained model issues optimal class predictions. The difference between the two is rather subtle but it can be argued that UQ-C-index has the advantage to reward scoring functions that rank highly inputs that the model learnt to map to the most probable class labels and filters out correct labeling arising from lucky outcomes where $Y \neq \hat{y}_{\text{bay}}$ while having $Y = \hat{y}$. On a same note, when the aleatoric uncertainty vanishes, then properties Theorem 2.1 and Theorem 2.2 become

equivalent and the two metrics will promote similar scoring functions.

4 EXPERIMENTS

This section is dedicated to experimentally illustrating the soundness of the investigated UQ metrics². In this setting, we emphasize that the proposed metrics are ground-truth free and can be used to evaluate UQ methods solely based on (\mathbf{x}, y) pairs. However for the sake of comparison, we need a ground-truth dependent metric, namely the Kendall correlation $\kappa(s; \text{gt})$ between a scoring s and ground truth $\text{gt} \in \{\phi, \varphi\}$. We conducted our experiment first on synthetic data, where the uncertainty ground-truth is tractable by design. Then we scale our experiment on three real-world datasets, CIFAR10, R-ImageNet and OASST2, and leverage extra human annotations as a proxy for UQ ground-truth. More details on the datasets and the models used are provided in Section 17. Extra experiments notably illustrating some of our theoretical findings are presented in Section 17.

In the supervised learning setting, the ground truths ϕ and φ cannot be obtained from (\mathbf{x}, y) pairs which is why this kind of evaluation can only be carried out (with maximal reliability) using synthetic data where the data generating distribution is known. To obtain closed forms for ϕ and φ , we design a toy binary classification dataset which is a mixture of two overlapping Gaussian distributions in input space, one for each label. The dataset is described at length in Section 17.

We trained several UQ aware models, and for each UQ method, we vary the hyper-parameters and evaluate the performances of the obtained scoring functions using UQ-AUC, UQ-C-index and the Kendall correlation between the scoring function and either ϕ or φ . Scatter plots are reported in Figures 1a and 1b. In those plots, each dot corresponds to one function $s(\mathbf{x})$. An excellent match with the ground-truth dependent coefficients can be observed which confirms that both the UQ-AUC and UQ-C-index are well-behaved metrics to evaluate the quality of the uncertainty scoring functions. This is confirmed by Table 1 which shows Pearson correlations between the two metrics with the Kendall coefficients depending on any of the ground truth. It is remarkable that very good correlations are obtained for each metric even with the ground truth it is not tied to and even if aleatoric uncertainty is significant in the data.

An important message from those experiments is that the asymptotical qualities of the UQ-AUC and UQ-C-index metrics established by theoretical results from

the previous section empirically propagate to the finite sample case. Besides, Figures 1a and 1b highlight that the choice of hyper-parameters (either of the model, training procedure or UQ method) is crucial to obtain meaningful scores. Indeed, several runs exhibit a negative Kendall correlation. The UQ-AUC and UQ-C-index experimentally have the desirable property that for negative Kendall correlation, their values are below 0.5 most of the time.

For the experiment on real-world image datasets, we leverage 116 models trained on ImageNet. For CIFAR10-H, we fine-tuned those models on the ultimate layer for 10 epochs. Following the work of Peterson et al. (2019), we leverage extra human annotations as a proxy for uncertainty ground truth, and evaluate the correlation between the scoring functions and the aforementioned human uncertainty. This is meant to check if the investigated metrics, without having access to ground truth, are still aligned with human judgments. We thus compare the investigated metrics UQ-AUC and UQ-C-index to Kendall correlations (between $s(\mathbf{x})$ and ground truths computed from $P(Y_{\text{human}}|\mathbf{x})$). For the natural language dataset OASST2, 70 pre-trained models were leveraged. Several human annotations per text are also available to assess the text content as positive or negative thus enabling to build an uncertainty ground truth. Additional information on the real-world datasets and the leveraged models are provided in Section 17.

Figure 1c to Figure 1g show the correlation of the investigated metrics with the uncertainty ground-truths they are tied to. Similarly as for synthetic data, high levels of correlations are observed between the Kendall coefficients of a ground truth and the metric maximized by this ground truth. However, looking at Table 1 significantly lower correlations are observed when Kendall coefficients of a ground truth are compared to the metric not maximized by them. This shows that the two metrics do not catch the same signals within scoring functions and offer different standpoints as to the performances of the those functions. We believe that this contrast is rooted in the very definitions of the ground truths, one seeking rewarding correct predictions, the other one rewarding correct decisions.

5 CONCLUSION

One of the most popular outcomes of deep uncertainty methods are scoring functions which are meant to rank inputs based on their likeliness to be mapped to correct class labels by a trained model. A legitimate expectation of some end-user of such scoring functions is to be able to check if the issued ranking are effectively allowing to secure input space regions in which the

²All code is available in the following repository: <https://github.com/owkin/legitimate-uq-metrics>.

model exhibits higher trustworthiness. Unfortunately, this ambition is impaired by the absence of uncertainty ground truth to perform this evaluation. This paper examines two types of uncertainty ground truth. We establish the validity of those in terms of their ability to yield input space regions where a classifier can be safely used at a level that can be chosen by the end-user. In addition, we investigate two metrics that can be computed solely using a classical test set while being tied to each identified ground truth scoring function. We prove that the class of functions that almost surely rank inputs in the same way each ground truth are the maximizers of one of the metrics. Finally, we also prove that the safer input space regions spanned by those maximizers exhibit a form of optimality in the sense that no other scoring with identical properties can encompass a larger volume of correct predictions. Our claims are also supported by empirical evidence in various scenarios. We hope that our contributions will unlock a wider use of UQ methods for safer machine learning practices.

A potential misuse of those metrics would arise from not taking into account finite sample randomness in the reported metric values. Note that since the investigated metrics rely on well studied performance measures (ROC-AUC and C-index), statistical tests designed to assess the significance of performance gaps can be readily used for their uncertainty analogues.

Acknowledgements

The authors would like to thank Geneviève Robin, Jean-Philippe Vert and Alex Nowak for the valuable discussions and insightful feedback that helped shaping the ideas presented in this work. We are also thankful to Thomas Chaigneau for its valuable help on the OASST2’s experiments.

References

- Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *CoRR*, abs/2006.07159, 2020. URL <https://arxiv.org/abs/2006.07159>.
- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.
- João Caldeira and Brian Nord. Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms. *Machine Learning: Science and Technology*, 2(1):015002, 2020.
- Stéphan Cléménçon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008. ISSN 00905364. URL <http://www.jstor.org/stable/25464648>.
- Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.
- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.
- Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers. *ICLR*, 2023a.
- Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. *ICLR*, 2023b.
- Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. *ICLR*, 2019.
- Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.
- Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.
- Maximilian Henne, Adrian Schwaiger, Karsten Roscher, and Gereon Weiss. Benchmarking uncertainty estimation methods for deep learning with safety-related metrics. In *SafeAI@ AAAI*, pages 83–90, 2020.
- Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bannamoun. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.
- Michael Kirchhof, Bálint Mucsányi, Seong Joon Oh, and Dr Enkelejda Kasneci. Url: A representation learning benchmark for transferable uncertainty estimates. *Advances in Neural Information Processing Systems*, 36:13956–13980, 2023.

- Lingkai Kong, Harshavardhan Kamarthi, Peng Chen, B. Aditya Prakash, and Chao Zhang. Uncertainty quantification in deep learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 5809–5810, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599577.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. Openassistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=VSJotgbPHF>.
- Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*, 2021.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
- Jing Lei and Larry Wasserman. Distribution-free prediction bands for nonparametric regression. *Quality control and applied statistics*, 60(1):109–110, 2015.
- Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512, 2020a.
- Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475, 2020b.
- Thomas Mortier, Viktor Bengs, Eyke Hüllermeier, Stijn Luca, and Willem Waegeman. On the calibration of probabilistic classifier sets. In *International Conference on Artificial Intelligence and Statistics*, pages 8857–8870. PMLR, 2023.
- Venkat Nemani, Luca Biggio, Xun Huan, Zhen Hu, Olga Fink, Anh Tran, Yan Wang, Xiaoge Zhang, and Chao Hu. Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial. *Mechanical Systems and Signal Processing*, 205:110796, 2023.
- Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.
- Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626, 2019.
- Xin Qiu and Risto Miikkulainen. Detecting misclassification errors in neural networks with a gaussian process model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8017–8027, 2022.
- Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, et al. Is one annotation enough?-a data-centric image classification benchmark for noisy and ambiguous label estimation. *Advances in Neural Information Processing Systems*, 35:33215–33232, 2022.
- Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15 (1):1929–1958, 2014.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.
- Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.
- Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence*, pages 2282–2292. PMLR, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
 - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes Applicable]
 - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
 - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
 - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
 - (b) Complete proofs of all theoretical results. [Yes]
 - (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
 - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
 - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
 - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
 - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - (a) Citations of the creator If your work uses existing assets. [Not Applicable]
 - (b) The license information of the assets, if applicable. [Not Applicable]
 - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
 - (d) Information about consent from data providers/curators. [Not Applicable]
 - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
 - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
 - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
 - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Legitimate ground-truth-free metrics for deep uncertainty classification scoring: Supplementary Materials

6 UQ METHOD OVERVIEW

One of the goals of the present paper is to shed light on well behaved metrics to compare UQ methods. As such, this appendix reviews some of the most popular methods in this field. Since $\hat{\pi}$ is a softmax vector of probabilities, the neural network can be regarded as an uncertainty representation of its own. UQ scorings computed from softmax probabilities constitute a baseline studied in (Hendrycks and Gimpel, 2016). Those scoring functions are presented in Section 7.

The softmax baseline offers very limited additional insights compared to the network output. In binary classification, $\hat{\pi}_1$ is the predicted probability for class 1 given \mathbf{x} and is a sufficient statistic of the corresponding Bernoulli distribution. Scoring functions depending on $\hat{\pi}_1$ thus capture the same kind of information with respect to the uncertainty carried by that Bernoulli distribution. To gain additional insights, authors rapidly moved to a more flexible uncertainty representation involving a hierarchical model sometimes also referred to as second order probabilities. In classification, the set of first order distributions are multinomial/categorical distributions and live in the standard ℓ -dimensional simplex where ℓ is the cardinal of \mathcal{Y} . A second order distribution should allocate probabilities to each element of the simplex which is precisely what the Dirichlet distribution does.

Sensoy et al. (2018) introduced a model called evidential network (EvN) whose output is a vector $\mathbf{e}(\mathbf{x}) \in \mathbb{R}^{\ell}$ such that $\hat{\Pi}|\mathbf{x} \sim \text{Dir}(\mathbf{e} + \mathbf{1})$, with a specific architecture and objective loss in the form of an integrated risk with respect to $\hat{\pi}$. Dirichlet/Multinomial hierarchical models is a typical case where Bayesian statistics offer the possibility to compute the posterior distribution $p_{\hat{\Pi}|\mathbf{x}, \mathcal{D}}$ where \mathcal{D} are observed training data. Usually, since $\hat{\pi}(\mathbf{x})$ is the first level multinomial distribution, the randomness in $\hat{\Pi}(\mathbf{x}, \mathcal{D})$ can be viewed as arising from the randomness of $\Theta|\mathcal{D}$, where Θ denotes the model parameters. If training data would allow to find out that $p_{\Theta|\mathcal{D}} = \delta_{\theta}$ then the (epistemic) uncertainty in $\hat{\Pi}$ would be wiped out. Inferring the posterior parameter distribution is precisely what Bayesian neural networks (BNNs) (Blundell et al., 2015; Graves, 2011) propose to do.

Another very popular UQ technique is called deep ensembles (DEs) and was introduced by (Lakshminarayanan et al., 2017). In this case, the idea is to use multiple instances $\left(\hat{\pi}^{(u)}\right)_{u=1}^{n_{\text{ens}}}$ of a single network architecture. The finite set of learnt functions are different due to random sampling of the initial network parameters, shuffling during stochastic gradient descent and adversarial perturbations. DEs share some similarity with BNNs as, in practice, samples from the model posterior are used which de facto constitute an ensemble of models.

Another way to introduce a distribution over the network parameters is to view each of them as a Bernoulli variable allowing each parameter to switch between its learnt value and zero, i.e. through dropout (Srivastava et al., 2014). By sampling from these Bernoulli variables, we obtain MC Dropout (Gal and Ghahramani, 2016) and a set of predictive probabilities is also issued.

For end-users interested in the OOD detection use case, it is very instrumental to leverage input distances as shown by Liu et al. (2020a). The authors use a Gaussian process on logits whose kernel brings distance awareness and induces second order probabilities through the softmax function.

Instead of sampling by using multiple models, i.e. through the distribution of Θ , another possibility is to do this through the input distribution by leveraging augmentation mechanisms. Wang et al. (2019) propose to do so at test time, hence the name of this method called test time augmentation (TTA). The data augmentation being a random process, it induces a distribution on $\hat{\Pi}|\mathbf{x}$.

Finally, it should also be mentioned that another branch of the literature is dedicated to UQs learnt from a held-out set of data iid under $p_{X,Y}$. One can typically learn a “meta-classifier” to map inputs to their likeliness of being incorrectly classified by $\hat{\pi}$, see (Qiu and Miikkulainen, 2022) and references therein for examples of such approaches. Here again, the meta-classifier can be chosen to output first or second order types of representation.

7 SCORING FUNCTIONS

As evoked in Section 6, in binary classification, when quantifying uncertainty using first order probabilities, the scoring functions one can craft all carry roughly the same information. Let $\tilde{Y}|\mathbf{x}$ denotes a random variable with distribution $\text{Ber}(\hat{\pi}_1)$. Possible choices for a scoring $s(\mathbf{x})$ are typically $\text{Var}[\tilde{Y}|X = \mathbf{x}]$, the entropy $H(\tilde{Y}|X = \mathbf{x})$ or the gap $1 - \hat{\pi}_{\hat{y}}$.

For uncertainty representation leveraging second order probabilities, there is increased flexibility in the design of scoring functions. Concerning BNNs, DEs and MCD, one has access to a set of predicted softmax vectors $(\hat{\pi}^{(u)})_{u=1}^{n_{\text{ens}}}$. Usual scoring functions are defined as follows:

- Variation Ratio $\text{VR} = 1 - \hat{\pi}_{\hat{y}}$,
- total entropy $\text{TE} = H(\tilde{Y}^{(\text{ens})}|X = \mathbf{x})$ where $\tilde{Y}^{(\text{ens})}|X = \mathbf{x}$ denotes a random variable whose distribution is $\hat{\pi}^{(\text{ens})} = \frac{1}{n_{\text{ens}}} \sum_{u=1}^{n_{\text{ens}}} \hat{\pi}^{(u)}$,
- aleatoric entropy $\text{AE} = \frac{1}{n_{\text{ens}}} \sum_{u=1}^{n_{\text{ens}}} H(\tilde{Y}^{(u)}|X = \mathbf{x})$ where $\tilde{Y}^{(u)}|X = \mathbf{x}$ denotes a random variable whose distribution is $\hat{\pi}^{(u)}$,
- mutual information MI which by definition is equal to $\text{TE} - \text{AE}$,
- and free energy $-T \cdot \log \sum_{y \in \mathcal{Y}} e^{c_y/T}$ where \mathbf{c} is the vector of logits (pre-softmax output) and T is a temperature parameter (Liu et al., 2020b).

As can be understood from the above nomenclature, ensemble based UQ allows finer grained scoring which can disambiguate aleatoric from epistemic uncertainty. Concerning TTA, one uncertainty score proposed by the authors is obtained by computing the entropy of the empirical multinomial distribution obtained from the set of predicted labels to which augmented inputs were mapped. According to the authors, this score should align with aleatoric uncertainty as the process simulates outcomes in the vicinity of an input. For epistemic uncertainty, the authors use MCD without TTA.

8 METRICS ASSESSING RANKING PERFORMANCE VERSUS METRICS ASSESSING CALIBRATION PERFORMANCE

The scope of this paper is input ranking based on classification error risk. Equipped with a well-behaved ranking, one can secure regions of the input space where the model, whose uncertainty is being quantified, can be used safely. As an illustration of the discrepancy between this scope and calibration, we designed an illustrative example leveraging Temperature Scaling (Guo et al., 2017). If the Temperature Scaling transform is applied to the softmax probabilities issued by $\hat{\pi}$ with different scaling factors, then different ECE values are observed. If inputs are ranked using the softmax scoring function, then applying Temperature Scaling will not modify the ranking. In this case, metrics such as UQ-AUC or UQ-C-index are invariant to this transform as shown in the following figure obtained using the synthetic dataset as can be seen on Figure 2.

This simple example shows that very high UQ-AUC or UQ-C-index values can be achieved by the softmax uncertainty scoring (thereby allowing to correctly identify risk-prone inputs) while the underlying softmax probabilities are not necessarily well calibrated.

To further emphasize that calibration do not capture the ground truth signal sought in this paper, one can check the correlation of ECE and $\kappa(s, \phi)$ or $\kappa(s, \varphi)$. On the synthetic dataset, the achieved correlations are respectively equal to 0.09% and -10.57%.

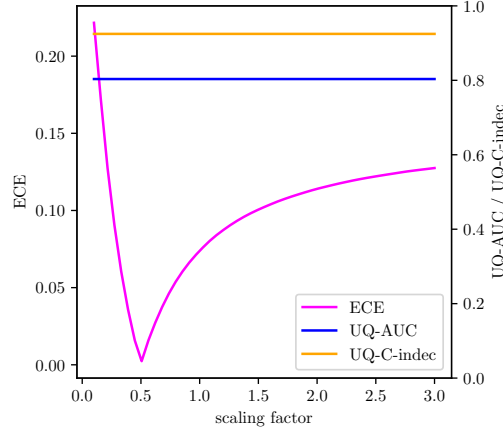


Figure 2: ECE, UQ-AUC and UQ-C-index as functions of the scaling factor of Temperature Scaling.

Besides, in Section 2.1, we explain that (10) holds for infinite families of scorings while calibration holds for a unique score. This statement refers to (Guo et al., 2017, eq. (1)) defining perfect calibration of a confidence estimate taking values in the full range of the $[0; 1]$ interval. Two estimates achieving perfect calibration in this sense need to be equal almost everywhere. This is in sharp contrast to the elements of \mathcal{E}_ϕ which do not have to be equal almost everywhere while all maximizing UQ-AUC. This is why we argue that ECE and its variants solve a different evaluation problem compared to the metrics investigated in this paper.

9 PROOF OF Theorem 2.1

Theorem 2.1 shows that sub-level sets $\mathcal{L}_\beta(s) := \{\mathbf{x} \in \mathcal{X} | s(\mathbf{x}) \leq \beta\}$ of any scoring in the equivalence class of ground truth ϕ yield a family of nested sets in which the accuracy of the predictor \hat{y} decreases with β . To prove this result, we will need to leverage an intermediate result which is first proved in the following paragraphs.

9.1 Intermediate result

Theorem 9.1. *For any pair of scoring function s_1 and $s_2 \in \mathcal{E}_s$ and any $\beta \in \mathbb{R}$, $\exists \gamma \in \mathbb{R}$ such that*

$$P(X \in \mathcal{L}_\beta(s_1) \setminus \mathcal{L}_\gamma(s_2) \cup \mathcal{L}_\gamma(s_2) \setminus \mathcal{L}_\beta(s_1)) = 0. \quad (22)$$

By definition of \mathcal{E}_s , we have

$$\Leftrightarrow \begin{cases} P(\{s(X) \geq s(X')\} \cap \{s_1(X) < s_1(X')\}) = 0 \\ P(\{s(X) < s(X')\} \cap \{s_1(X) \geq s_1(X')\}) = 0 \end{cases} \quad (23)$$

If s is constant p_X -almost everywhere, then $P(X \in \mathcal{L}_\beta(s)) = 0$ or 1. This implies that $P(s(X) < s(X')) = 0$. Leveraging the top equation in (23), one can write

$$P(\{s(X) \geq s(X')\} \cap \{s_1(X) < s_1(X')\}) = 0, \quad (24)$$

$$\Leftrightarrow P(\{s(X) \geq s(X')\} \cap \{s_1(X) < s_1(X')\}) + P(s(X) < s(X')) = 0, \quad (25)$$

$$\Leftrightarrow P(\{s(X) \geq s(X')\} \cap \{s_1(X) < s_1(X')\}) + P(\{s(X) < s(X')\} \cap \{s_1(X) < s_1(X')\}) = 0, \quad (26)$$

$$\Leftrightarrow P(s_1(X) < s_1(X')) = 0. \quad (27)$$

This means that s_1 is also constant p_X -almost everywhere. If $P(X \in \mathcal{L}_\beta(s)) = 0$, it suffices to choose γ sufficiently small so that $P(X \in \mathcal{L}_\gamma(s_1)) = 0$ as well and (22) holds for the scorings s and s_1 . Likewise, if $P(X \in \mathcal{L}_\beta(s)) = 1$, it suffices to choose γ sufficiently high so that $P(X \in \mathcal{L}_\gamma(s_1)) = 1$ as well and (22) holds for the scorings s and s_1 .

We now proceed with a scoring s that is not constant p_X -almost everywhere and thus $P(s(X) < s(X')) > 0$. Based on the top equation in (23), this implies that

$$P(s_1(X) \geq s_1(X') | s(X) < s(X')) = 0, \quad (28)$$

$$\Leftrightarrow P(s_1(X) < s_1(X') | s(X) < s(X')) = 1. \quad (29)$$

Now suppose $X \in \mathcal{L}_\beta(s)$ and $X' \notin \mathcal{L}_\beta(s)$. Since this implies that $s(X) < s(X')$, we deduce

$$P(s_1(X) < s_1(X') | X \in \mathcal{L}_\beta(s), X' \notin \mathcal{L}_\beta(s)) = 1. \quad (30)$$

The above equation means that $\mathcal{L}_\beta(s)$ and its complementary set define a partition of the input space such that s_1 almost surely takes values smaller in $\mathcal{L}_\beta(s)$ than in its complementary set. Define γ as the following essential supremum:

$$\gamma := \inf \{M \in \mathbb{R} | P(X \in \mathcal{L}_\beta(s) | X \notin \mathcal{L}_M(s_1)) = 0\} \quad (31)$$

By construction, we have $P(\mathcal{L}_\beta(s) \setminus \mathcal{L}_\gamma(s_1)) = 0$. Moreover, thanks to (30), we know that $P(\mathcal{L}_\gamma(s_1) \setminus \mathcal{L}_\beta(s)) = 0$ otherwise there would be a subset $U \subset \mathcal{L}_\gamma(s_1)$ in the complementary set of $\mathcal{L}_\beta(s)$ such that $P(X \in U) > 0$ and one could find a smaller scalar than γ allowing to exclude U while keeping $P(X \in \mathcal{L}_\beta(s) | X \notin \mathcal{L}_\gamma(s_1) \setminus U) = 0$. Relation (22) thus holds for scoring s and s_1 .

Since Relation (22) is symmetric and also holds for scoring s and s_2 , by transitivity, it holds as well for s_1 and s_2 .

9.2 Main proof

In this proof, we are interested in the following quantities

$$E(\beta; s) := P(Y \neq \hat{Y} | X \in \mathcal{L}_\beta(s)). \quad (32)$$

Remembering that $\phi(\mathbf{x}) = P(Y \neq \hat{Y} | X = \mathbf{x})$, one can write

$$E(\beta; \phi) = \frac{\int_{\mathcal{L}_\beta(\phi)} P(Y \neq \hat{Y} | X = \mathbf{x}) p_X(\mathbf{x}) d\mathbf{x}}{P(X \in \mathcal{L}_\beta(\phi))}, \quad (33)$$

$$\leq \frac{\int_{\mathcal{L}_\beta(\phi)} \beta p_X(\mathbf{x}) d\mathbf{x}}{P(X \in \mathcal{L}_\beta(\phi))}, \quad (34)$$

$$\leq \beta. \quad (35)$$

Besides, if ϕ is almost surely constant, then relation (10) is trivially met³. Let us thus now suppose that this co-domain contains at least two distinct elements β and $\beta + h$ such that $\beta < 1$ and where h is a positive scalar, smaller than $1 - \beta$. One can write

$$E(\beta + h; \phi) = \frac{P(Y \neq \hat{Y}, X \in \mathcal{L}_{\beta+h}(\phi))}{\underbrace{P(X \in \mathcal{L}_\beta(\phi))}_b + \underbrace{P(X \in \mathcal{L}_{\beta+h}(\phi) \setminus \mathcal{L}_\beta(\phi))}_a}, \quad (36)$$

where a and b are positive scalars.

The numerator of $E(\beta + h; \phi)$ can be written as

$$P(Y \neq \hat{Y}, X \in \mathcal{L}_{\beta+h}(\phi)) = \int_{\mathcal{L}_{\beta+h}(\phi)} P(Y \neq \hat{Y} | X = \mathbf{x}) p_X(\mathbf{x}) d\mathbf{x}, \quad (37)$$

$$= bE(\beta) + \int_{\mathcal{L}_{\beta+h}(\phi) \setminus \mathcal{L}_\beta(\phi)} P(Y \neq \hat{Y} | X = \mathbf{x}) p_X(\mathbf{x}) d\mathbf{x}. \quad (38)$$

³In this case, the sub-level sets $\mathcal{L}_\beta(\phi)$ have a measure either equal to 0 or 1. Only those with a strictly positive measure are involved in (10) and then $E(\beta; \phi) = P(Y \neq \hat{Y}) = 1 - \text{ACC}$.

We also know that for any $\mathbf{x} \in \mathcal{L}_{\beta+h}(\phi) \setminus \mathcal{L}_{\beta}(\phi)$, $P(Y \neq \hat{Y}|\mathbf{x}) > \beta$. We thus can write

$$E(\beta + h; \phi) > \frac{bE(\beta; \phi) + a\beta}{b + a} \quad (39)$$

We can now leverage relation (35) stating that $E(\beta; \phi) \leq \beta$ to obtain

$$E(\beta + h; \phi) > \frac{bE(\beta; \phi) + aE(\beta; \phi)}{b + a}, \quad (40)$$

$$> E(\beta; \phi). \quad (41)$$

We hereby conclude that $E(., \phi)$ is a strictly increasing function of β . An illustration compliant with claim is shown in Figure 3d. By observing that $E(\beta = 0^+; \phi) = 0$ and $E(\beta = 1; \phi) = P(Y \neq \hat{Y}) = 1 - \text{ACC}$, we know that the set $\mathcal{C}_{\text{mce}}(\mathcal{E}_{\phi})$ of achievable values of $E(., \phi)$ is a subset of $]0; 1 - \text{ACC}]$. Let $\psi_{\phi} = E(., \phi)^{-1}$, then obviously $E(\psi_{\phi}(\gamma); \phi) = P(Y \neq \hat{Y} | X \in \mathcal{L}_{\psi_{\phi}(\gamma)}(s)) = \gamma$, $\forall \gamma \in \mathcal{C}_{\text{mce}}(\mathcal{E}_{\phi})$ and (10) holds for ground truth scoring ϕ .

Moreover, thanks to Theorem 9.1, we know that $\forall s \in \mathcal{E}_{\phi}$ there is bijection ψ_s such that $E(\psi_s(\gamma), s) = E(\gamma, \psi_{\phi}^{-1} \circ \phi) = \gamma$. As ψ_s is clearly a strictly increasing function of γ , then (10) holds for s as well which concludes the proof.

10 PROOF OF Theorem 2.2

Theorem 2.2 shows that sub-level sets $\mathcal{L}_{\beta}(s)$ of any scoring in the equivalence class of ground truth φ yield a family of nested sets in which the probability of predicting the same class label as the Bayes classifier decreases with β . As such, we are interested in this proof in the following quantities

$$E(\gamma; s) := P(\hat{Y}_{\text{bay}} \neq \hat{Y} | X \in \mathcal{L}_{\gamma}(s)). \quad (42)$$

Remembering that $\varphi(\mathbf{x}) = \frac{1}{2} \|f_{\theta}(\mathbf{x}) - \delta_{\hat{y}_{\text{bay}}}\|_1 = 1 - \hat{\pi}_{\hat{y}_{\text{bay}}}$, one can write

$$P(\hat{Y}_{\text{bay}} \neq \hat{Y}, X \in \mathcal{L}_{\beta}(\varphi)) = \int_{\mathcal{L}_{\beta}(\varphi)} P(\hat{Y}_{\text{bay}} \neq \hat{Y} | X = \mathbf{x}) p_X(\mathbf{x}) d\mathbf{x}, \quad (43)$$

$$= \int_{\mathcal{L}_{\beta}(\varphi)} \mathbb{1}_{\varphi(\mathbf{x}) \geq 1/2} p_X(\mathbf{x}) d\mathbf{x}, \quad (44)$$

$$= \begin{cases} 0 & \text{if } \beta \leq 1/2, \\ \int_{\mathcal{L}_{\beta}(\varphi) \setminus \mathcal{L}_{1/2}(\varphi)} p_X(\mathbf{x}) d\mathbf{x} & \text{otherwise} \end{cases}, \quad (45)$$

$$= \max(0, P(X \in \mathcal{L}_{\beta}(\varphi)) - P(X \in \mathcal{L}_{1/2}(\varphi))). \quad (46)$$

We deduce that $E(\beta; \varphi) = P(\hat{Y}_{\text{bay}} \neq \hat{Y} | X \in \mathcal{L}_{\beta}(\varphi)) = \max\left(0, 1 - \frac{P(X \in \mathcal{L}_{1/2}(\varphi))}{P(X \in \mathcal{L}_{\beta}(\varphi))}\right)$ which implies that $E(., \varphi)$ is null on $]0; 1/2]$ and increasing on $]1/2; 1]$. An illustration compliant with claim is shown in Figure 3d. In particular, $E(1; \varphi) = 1 - P(X \in \mathcal{L}_{1/2}(\varphi)) = P(\hat{Y}_{\text{bay}} \neq \hat{Y})$. Consequently, we know that the set $\mathcal{C}_{\text{mbc}}(\mathcal{E}_{\varphi})$ of achievable values of $E(., \varphi)$ is a subset of $]0; P(\hat{Y}_{\text{bay}} \neq \hat{Y})]$. Let $\psi_{\varphi} = E(., \varphi)^{-1}$, then obviously $E(\psi_{\varphi}(\gamma); \varphi) = P(\hat{Y}_{\text{bay}} \neq \hat{Y} | X \in \mathcal{L}_{\psi_{\varphi}(\gamma)}(s)) = \gamma$, $\forall \gamma \in \mathcal{C}_{\text{mbc}}(\mathcal{E}_{\varphi})$ and (11) holds for ground truth scoring φ .

Moreover, thanks to Theorem 9.1, we know that $\forall s \in \mathcal{E}_{\varphi}$ there is bijection ψ_s such that $E(\psi_s(\gamma), s) = E(\gamma, \psi_{\varphi}^{-1} \circ \varphi) = \gamma$. As ψ_s is clearly a strictly increasing function of γ , then (11) holds for s as well which concludes the proof.

11 PROOF OF Theorem 2.3

This appendix contains a proof that s is a maximizer of UQ-AUC iff $s \in \mathcal{E}_\phi$. The fact that ϕ and any increasing measurable function of ϕ is a maximizer of UQ-AUC is a direct consequence of (Cl  men  on et al., 2008, Proposition 1) and this part of the proof is thus omitted. It remains to prove that other elements in \mathcal{E}_ϕ are also maximizers and that all maximizers belong to \mathcal{E}_ϕ .

We start by showing that any $s \in \mathcal{E}_\phi$ is a maximizer of UQ-AUC. To comply with notations in (Cl  men  on et al., 2008), let us introduce the following binary relation

$$r_s(\mathbf{x}, \mathbf{x}') = \begin{cases} 1 & \text{if } s(\mathbf{x}) \geq s(\mathbf{x}') \\ -1 & \text{otherwise} \end{cases}. \quad (47)$$

Let us also define the following random variables $V = \mathbb{1}_{Y \neq \hat{Y}|X}$, $V' = \mathbb{1}_{Y' \neq \hat{Y}'|X'}$ and $Z = V - V'$. (Cl  men  on et al., 2008) proved that s is a maximizer of the UQ-AUC iff it minimizes

$$P(Z \cdot r_s(X, X') < 0) = P(\{Z \cdot r_s(X, X') < 0\} \cap \{Z \cdot r_\phi(X, X') < 0\}) \quad (48)$$

By definition of \mathcal{E}_ϕ , we also have $P(r_\phi(X, X') \neq r_s(X, X')) = 0$ and events $\{Z \cdot r_s(X, X') < 0\}$ and $\{Z \cdot r_\phi(X, X') < 0\}$ have the same probability therefore s is also a maximizer of UQ-AUC.

We now move to the proof that any maximizer of UQ-AUC is in \mathcal{E}_ϕ . Let s denote a maximizer of UQ-AUC such that $\exists \mathcal{A} \subset \mathcal{X}^2$ with $\lambda = P((X, X') \in \mathcal{A}) > 0$ and $P(r_\phi(X, X') \cdot r_s(X, X') < 0 | (X, X') \in \mathcal{A}) = 1$.

Let us then define the following binary relation

$$\tilde{r}(\mathbf{x}, \mathbf{x}') = \begin{cases} \max\{r_s(\mathbf{x}, \mathbf{x}'); r_\phi(\mathbf{x}, \mathbf{x}')\} & \text{if } Z > 0 \text{ and } (\mathbf{x}, \mathbf{x}') \in \mathcal{A} \\ \min\{r_s(\mathbf{x}, \mathbf{x}'); r_\phi(\mathbf{x}, \mathbf{x}')\} & \text{if } Z < 0 \text{ and } (\mathbf{x}, \mathbf{x}') \in \mathcal{A} \\ r_\phi(\mathbf{x}, \mathbf{x}') & \text{otherwise} \end{cases}. \quad (49)$$

One then has

$$P(Z \cdot \tilde{r}(X, X') < 0) = P(Z \cdot \tilde{r}(X, X') < 0, (\mathbf{x}, \mathbf{x}') \in \mathcal{A}) + P(Z \cdot r_\phi(X, X') < 0, (\mathbf{x}, \mathbf{x}') \notin \mathcal{A}). \quad (50)$$

Because inside \mathcal{A} we know that $r_s(X, X') = -r_\phi(X, X')$, by construction of \tilde{r} , we have $P(Z \cdot \tilde{r}(X, X') < 0, (\mathbf{x}, \mathbf{x}') \in \mathcal{A}) = 0$. This implies that $P(Z \cdot \tilde{r}(X, X') < 0) = P(Z \cdot r_\phi(X, X') < 0, (\mathbf{x}, \mathbf{x}') \notin \mathcal{A}) < P(Z \cdot r_\phi(X, X') < 0)$. This is in contraction with ϕ being a maximizer of UQ-AUC which concludes the proof.

12 PROOF OF Theorem 2.4

In this appendix, one aims at showing that a scoring s for which (10) holds is not uniformly better than $s^* \in \mathcal{E}_s$. In this proof, we will use the following random variables $V(\mathbf{x}) = \mathbb{1}_{Y \neq \hat{Y}|\mathbf{x}}$ and $V'(\mathbf{x}') = \mathbb{1}_{Y' \neq \hat{Y}'|\mathbf{x}'}$. For the sake concision, we also define $V = V(X)$, $V' = V'(X')$, $S = s(X)$ and $S' = s(X')$.

We will also be need to introduce the scoring $\tilde{s} = \psi_s^{-1} \circ s$ where ψ_s is the strictly increasing function allowing to re-index the sub-level set of s to achieve (10). For any $\gamma \in \mathcal{C}_{\text{mbc}}(\mathcal{E}_s)$, we thus have

$$\mathbf{x} \in \mathcal{L}_\gamma(\tilde{s}) \Leftrightarrow \tilde{s}(\mathbf{x}) \leq \gamma \quad (51)$$

$$\Leftrightarrow \psi_s(\tilde{s}(\mathbf{x})) \leq \psi_s(\gamma) \quad (52)$$

$$\Leftrightarrow s(\mathbf{x}) \leq \psi_s(\gamma) \quad (53)$$

$$\Leftrightarrow \mathbf{x} \in \mathcal{L}_{\psi_s(\gamma)}(s). \quad (54)$$

Since ψ_s^{-1} is an increasing function, $\tilde{s} \in \mathcal{E}_s$ and (10) also holds for this scoring. This means that the bijection to re-index sub-level sets of \tilde{s} as part of property (10) is the identity function. Likewise, we define \tilde{s}^* from s^* in the same way.

Suppose $s^* \prec_{\text{mce}} s \Leftrightarrow \tilde{s}^* \prec_{\text{mce}} \tilde{s}$. The UQ-AUC achieved by some scoring s then writes

$$\text{UQ-AUC}(s) = P(S \leq S' | V < V'), \quad (55)$$

$$= P(\tilde{S} \leq \tilde{S}' | V < V'), \quad (56)$$

$$= \int P(\tilde{S} \leq \tilde{S}', \tilde{S}' = \gamma | V < V') d\gamma, \quad (57)$$

$$= \int P(\tilde{S} \leq \gamma | V < V') d\gamma, \quad (58)$$

$$= \int P(X \in \mathcal{L}_\gamma(\tilde{s}) | V < V') d\gamma, \quad (59)$$

$$= \int P(X \in \mathcal{L}_\gamma(\tilde{s}) | V = 0, V' = 1) d\gamma, \quad (60)$$

$$= \int P(X \in \mathcal{L}_\gamma(\tilde{s}) | V = 0) d\gamma, \quad (61)$$

$$= \int P(X \in \mathcal{L}_\gamma(\tilde{s}) | Y = \hat{Y}) d\gamma, \quad (62)$$

$$> \int P(X \in \mathcal{L}_\gamma(\tilde{s}^*) | Y = \hat{Y}) d\gamma. \quad (63)$$

We now can conclude that if s was dominating this s^* in the sense defined in Theorem 2.4, then it would achieve a larger UQ-AUC than the maximizers of the latter which is not possible.

13 PROOF OF Theorem 2.5

This appendix addresses the question of which family of functions maximizes the uncertainty concordance index. We address binary classification only, i.e. $\mathcal{Y} = \{0, 1\}$. We remind that $\text{UQ-C-index}(s)$ refers to the asymptotical definition of this metric which reads

$$\text{UQ-C-index}(s) = P(s(X) \leq s(X') | \Delta(\hat{\pi}(X), Y) \leq \Delta(\hat{\pi}(X'), Y')). \quad (64)$$

Based on the work of (Cl  men  on et al., 2008), we know that an optimal scoring s^* is spanned by checking if $\rho_+(\mathbf{x}, \mathbf{x}')$ is greater (or not) than $\rho_-(\mathbf{x}, \mathbf{x}')$ where those functions are defined as follows:

$$\rho_+(\mathbf{x}, \mathbf{x}') = P(\Delta(\hat{\pi}(\mathbf{x}), Y) - \Delta(\hat{\pi}(\mathbf{x}'), Y') > 0 | X = \mathbf{x}, X' = \mathbf{x}'), \quad (65)$$

$$\rho_-(\mathbf{x}, \mathbf{x}') = P(\Delta(\hat{\pi}(\mathbf{x}), Y) - \Delta(\hat{\pi}(\mathbf{x}'), Y') < 0 | X = \mathbf{x}, X' = \mathbf{x}'). \quad (66)$$

Using (17), we thus have

$$\rho_+(\mathbf{x}, \mathbf{x}') = P(\hat{\pi}_{Y'}(\mathbf{x}') > \hat{\pi}_Y(\mathbf{x}) | X = \mathbf{x}, X' = \mathbf{x}'), \quad (67)$$

$$= \sum_y \sum_{y'} P(\hat{\pi}_{y'}(\mathbf{x}') > \hat{\pi}_y(\mathbf{x}) | X = \mathbf{x}, X' = \mathbf{x}', Y = y, Y' = y') \quad (68)$$

$$= \sum_y \sum_{y'} \mathbb{1}_{\hat{\pi}_{y'}(\mathbf{x}') > \hat{\pi}_y(\mathbf{x})} \nu_y(\mathbf{x}) \nu_{y'}(\mathbf{x}'), \quad (69)$$

where $\nu_y(\mathbf{x}) := P(Y = y | X = \mathbf{x})$ and $\nu_{y'}(\mathbf{x}') := P(Y' = y' | X' = \mathbf{x}')$ are notations introduced for improved readability. As not all indicator functions in (69) can be equal to one simultaneously, a case by case examination leads to the following expression of ρ_+ :

$$\rho_+(\mathbf{x}, \mathbf{x}') = \sum_{y', y \text{ s.t. } \hat{\pi}_{y'}(\mathbf{x}') > \hat{\pi}_y(\mathbf{x})} \nu_y(\mathbf{x}) \nu_{y'}(\mathbf{x}'). \quad (70)$$

Since $p_{\hat{\Pi}}$ has no atom, we know that $>$ is a total order for the entries of $\hat{\pi}(\mathbf{x})$ and $\hat{\pi}(\mathbf{x}')$, i.e. there is no tie in the sorting and $\hat{\pi}_y(\mathbf{x}) \neq \hat{\pi}_{y'}(\mathbf{x}'), \forall y, y'$.

Let us break down the different sorting possibilities for a fixed (arbitrary) $(\mathbf{x}, \mathbf{x}')$. For the sake of readability, we omit dependencies over \mathbf{x} and \mathbf{x}' in most of the sequel of this proof and we use $\hat{\pi}'_y$ to denote $\hat{\pi}_y(\mathbf{x}')$ or \hat{y}' to denote $\hat{y}(\mathbf{x}')$. Let $\iota = \min(\hat{\pi}_0, \hat{\pi}_1, \hat{\pi}'_0, \hat{\pi}'_1)$.

As $\mathcal{Y} = \{0; 1\}$, we investigate the following cases:

- $\iota = \hat{\pi}_{y_0}$, i.e. it is an entry of $\hat{\pi}$ and class label with minimal probability is y_0 . This implies that $\max(\hat{\pi}_0, \hat{\pi}_1, \hat{\pi}'_0, \hat{\pi}'_1) = \hat{\pi}_{1-y_0}$. Consequently, for any y' the condition under the sum in (70) is always checked for $y = y_0$ and never checked for $y \neq y_0$. We obtain that

$$\rho_+(\mathbf{x}, \mathbf{x}') = P(Y = y_0 | X = \mathbf{x}). \quad (71)$$

- $\iota = \hat{\pi}'_{y'_0}$, i.e. it is an entry of $\hat{\pi}'$ and class label with minimal probability is y'_0 . This implies that $\max(\hat{\pi}_0, \hat{\pi}_1, \hat{\pi}'_0, \hat{\pi}'_1) = \hat{\pi}'_{1-y'_0}$. Consequently, for any y the condition under the sum in (70) is always checked for $y' \neq y'_0$ and never checked for $y' = y'_0$. We obtain that

$$\rho_+(\mathbf{x}, \mathbf{x}') = 1 - P(Y' = y'_0 | X' = \mathbf{x}'). \quad (72)$$

Checking to which distribution ι belongs to also corresponds to having either a more confident prediction for \mathbf{x} or \mathbf{x}' . They thus can also be determined in terms the entropies of $\hat{\pi}$ and $\hat{\pi}'$.

Since $Z = \Delta(\hat{\pi}(\mathbf{x}), Y) - \Delta(\hat{\pi}(\mathbf{x}'), Y')$ is a continuous random variable, checking that $\rho_+ > \rho_-$ amounts to checking that $\rho_+ > \frac{1}{2}$. We can thus now deduce an optimal UQ-C-index ranking rule

$$r^*(\mathbf{x}, \mathbf{x}') = \omega \mathbb{1}_{p_{Y|X=\mathbf{x}}(y_0) > \frac{1}{2}} + (1 - \omega) \mathbb{1}_{1-p_{Y'|X'=\mathbf{x}'}(y'_0) > \frac{1}{2}}, \quad (73)$$

$$= \omega \mathbb{1}_{\hat{y} \neq \hat{y}_{\text{bay}}} + (1 - \omega) \mathbb{1}_{\hat{y}' \neq \hat{y}'_{\text{bay}}}, \quad (74)$$

where $\omega = \mathbb{1}_{H(\hat{\pi}) < H(\hat{\pi}')} \in \{0; 1\}$ is a binary variable representing which input (from \mathbf{x} and \mathbf{x}') has the highest predicted entropy.

In order to further simplify the above, let us consider the $\omega = 1$ case and assume:

$$\hat{y} \neq \hat{y}_{\text{bay}} \quad (75)$$

$$\text{and } H(\hat{\pi}) < H(\hat{\pi}'). \quad (76)$$

We now investigate the following sub-cases:

- $\hat{y}' = \hat{y}'_{\text{bay}}$: In this case, $\|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 < 1$ and $\|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1 > 1$ where δ_y denotes the one-hot encoded vector corresponding to distribution δ_y .
- $\hat{y}' \neq \hat{y}'_{\text{bay}}$: As both predictions are in disagreement with the Bayes classifier, we know that 0.5 is between $\hat{\pi}'_{\hat{y}'_{\text{bay}}}$ (resp. $\hat{\pi}_{\hat{y}_{\text{bay}}}$) and \hat{y}'_{bay} (resp. \hat{y}_{bay}). In addition, $H(\hat{\pi}) < H(\hat{\pi}') \Rightarrow |\hat{\pi}'_{\hat{y}'_{\text{bay}}} - 0.5| < |\hat{\pi}_{\hat{y}_{\text{bay}}} - 0.5|$. Combining the two pieces of information, we have $0.5 - \hat{\pi}'_{\hat{y}'_{\text{bay}}} < 0.5 - \hat{\pi}_{\hat{y}_{\text{bay}}}$ which also means $\|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 < \|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1$.

Next, we consider the $\omega = 0$ case and assume:

$$\hat{y}' = \hat{y}'_{\text{bay}} \quad (77)$$

$$\text{and } H(\hat{\pi}) \geq H(\hat{\pi}'). \quad (78)$$

We investigate the following sub-cases:

- $\hat{y} = \hat{y}_{\text{bay}}$: As both predictions are correct, we know that $\hat{\pi}'_{\hat{y}'_{\text{bay}}}$ (resp. $\hat{\pi}_{\hat{y}_{\text{bay}}}$) is between 0.5 and \hat{y}'_{bay} (resp. \hat{y}_{bay}). In addition, $H(\hat{\pi}) \geq H(\hat{\pi}') \Rightarrow |\hat{\pi}'_{\hat{y}'_{\text{bay}}} - 0.5| \geq |\hat{\pi}_{\hat{y}_{\text{bay}}} - 0.5|$. Combining the two pieces of information, we have $\hat{\pi}'_{\hat{y}'_{\text{bay}}} - 0.5 \geq \hat{\pi}_{\hat{y}_{\text{bay}}} - 0.5$ which also means $\|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 \leq \|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1$.

- $\hat{y} \neq \hat{y}_{\text{bay}}$: In this case, $\|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1 > 1$ and $\|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 < 1$.

In all cases, we have $\|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 \leq \|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1$. We are thus provided with the following implication:

$$r^*(\mathbf{x}, \mathbf{x}') = 1 \Rightarrow \|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 \leq \|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1 \quad (79)$$

To prove the contraposition, assume $\|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 \leq \|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1$

We investigate the following cases:

$$\begin{cases} \|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1 < 1 & \text{and} & \|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 < 1 \Rightarrow \hat{y}' = \hat{y}'_{\text{bay}} \\ \|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1 \geq 1 & \text{and} & \|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 \geq 1 \Rightarrow \hat{y} \neq \hat{y}_{\text{bay}} \\ \|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1 \geq 1 & \text{and} & \|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 < 1 \Rightarrow \hat{y} \neq \hat{y}_{\text{bay}} \quad \text{and} \quad \hat{y}' = \hat{y}'_{\text{bay}} \end{cases} \quad (80)$$

Note that the last case, $\|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1 < 1$ and $\|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 > 1$ is not compatible under the initial assumption.

- For the first case, using the symmetry of H , we have that the closer the prediction to the Bayes prediction is, the lower is the entropy, hence $H(\hat{\pi}) \geq H(\hat{\pi}')$.
- For the second case, the closer the prediction from the Bayes prediction is, the higher is the entropy, hence $H(\hat{\pi}) < H(\hat{\pi}')$.
- For the last case, either we have $H(\hat{\pi}) < H(\hat{\pi}')$, and $\hat{y} \neq \hat{y}_{\text{bay}}$, or we have $H(\hat{\pi}) \geq H(\hat{\pi}')$ and $\hat{y}' = \hat{y}'_{\text{bay}}$.

In all the cases, we have $\|\hat{\pi}' - \delta_{\hat{y}'_{\text{bay}}}\|_1 \leq \|\hat{\pi} - \delta_{\hat{y}_{\text{bay}}}\|_1 \Rightarrow r^*(\mathbf{x}, \mathbf{x}') = 1$.

We can conclude that an optimal ranking rule for the UQ-C-index is the one that checks if the prediction is closer to the Bayes prediction or not.

$$r^*(\mathbf{x}, \mathbf{x}') = 1 \Leftrightarrow \frac{1}{2} \|\hat{\pi}(\mathbf{x}) - \delta_{\hat{y}_{\text{bay}}(\mathbf{x})}\|_1 \leq \frac{1}{2} \|\hat{\pi}(\mathbf{x}') - \delta_{\hat{y}_{\text{bay}}(\mathbf{x}')}\|_1 \quad (81)$$

And an associated scoring function is

$$s^*(\mathbf{x}) = \varphi(\mathbf{x}) = \frac{1}{2} \|\hat{\pi}(\mathbf{x}) - \delta_{\hat{y}_{\text{bay}}(\mathbf{x})}\|_1. \quad (82)$$

This proves that φ is a maximizer of the UQ-C-index. Using the same arguments as in the proof of Theorem 2.3 (see Section 11), we can directly conclude that s^* is a maximizer of UQ-C-index iff $s^* \in \mathcal{E}_\varphi$.

14 PROOF OF Theorem 2.6

In this appendix, one aims at showing that a scoring s for which (11) holds is not uniformly better than $s^* \in \mathcal{E}_s$. For the sake concision, we define $\Delta = \Delta(\hat{\pi}(X), Y)$, $\Delta' = \Delta(\hat{\pi}(X'), Y')$, $S = s(X)$ and $S' = s(X')$. Similarly as in Section 12, we denote $\tilde{s} = \psi_s^{-1} \circ s$ where ψ_s is the strictly increasing function involved in (11) and one than has $P(\hat{Y}_{\text{bay}} \neq \hat{Y} | X \in \mathcal{L}_\gamma(\tilde{s})) = \gamma, \forall \gamma \in \mathcal{C}_{\text{mbc}}(\mathcal{E}_\varphi)$. Likewise, we define \tilde{s}^* from s^* in the same way.

The UQ-C-index achieved by some scoring s then writes

$$\text{UQ-C-index}(s) = P(S \leq S' | \Delta < \Delta'), \quad (83)$$

$$= P(\tilde{S} \leq \tilde{S}' | \Delta < \Delta'), \quad (84)$$

$$= \int P(\tilde{S} \leq \tilde{S}', \tilde{S}' = \gamma | \Delta < \Delta') d\gamma, \quad (85)$$

$$= \int P(\tilde{S} \leq \gamma | \Delta < \Delta') d\gamma, \quad (86)$$

$$= \int \int P(\tilde{S} \leq \gamma, \Delta' = \alpha | \Delta < \Delta') d\gamma d\alpha, \quad (87)$$

$$= \int \int P(\tilde{S} \leq \gamma | \Delta < \alpha) P(\Delta' = \alpha | \Delta < \Delta') d\gamma d\alpha, \quad (88)$$

$$= \int \int P(\tilde{S} \leq \gamma | \Delta < \alpha) \frac{P(\Delta' = \alpha, \Delta < \Delta')}{P(\Delta < \Delta')} d\gamma d\alpha, \quad (89)$$

$$= 2 \int \int P(\tilde{S} \leq \gamma | \Delta < \alpha) P(\Delta < \alpha) d\gamma d\alpha, \quad (90)$$

$$= 2 \int \int P(X \in \mathcal{L}_\gamma(\tilde{s}) | \Delta < \alpha) P(\Delta < \alpha) d\gamma d\alpha, \quad (91)$$

$$= 2 \int \underbrace{\int P(X \in \mathcal{L}_\gamma(\tilde{s}) | \Delta < \alpha) d\gamma}_{J(\alpha)} P(\Delta < \alpha) d\alpha, \quad (92)$$

Define $J^*(\alpha) = \int P(X \in \mathcal{L}_\gamma(\tilde{s}^*) | \Delta < \alpha) d\gamma$. We have $J^*(\alpha) \leq J(\alpha)$ for any α as $\mathcal{L}_\gamma(\tilde{s}^*) \subset \mathcal{L}_\gamma(\tilde{s})$. In the remainder of this proof, s^* and all other $*$ -marked notations refer to this particular UQ-C-index maximizer.

Let us also remark that function J can be re-written as follows

$$J(\alpha) = \int \frac{P(\tilde{S} \leq \gamma, \Delta < \alpha)}{P(\tilde{S} \leq \infty, \Delta < \alpha)} d\gamma. \quad (93)$$

We can check that the cdf of distribution $p_{\tilde{S}, \Delta}$ is the main building brick in the definition of function J . Since $p_{S, \Delta}$ has no atom its cdf is continuous and so are the cdf of $p_{\tilde{S}, \Delta}$ and function J . Likewise, since $p_{S^*, \Delta}$ has no atom, J^* is continuous as well. Now we see that if there is an α_0 for which $J^*(\alpha_0) < J(\alpha_0)$ then the UQ-C-index of s would be larger than the UQ-C-index of s^* as the strict inequality holds at least for a small neighborhood (of positive Lebesgue measure) around α_0 . The next paragraphs will show that $J^*(\frac{1}{2}) < J(\frac{1}{2})$.

We have

$$J(\alpha) = \int_0^1 \int_{\mathcal{L}_\gamma(\tilde{s})} p_{X | \Delta < \alpha}(\mathbf{x}) d\mathbf{x} d\gamma, \quad (94)$$

$$= \frac{1}{P(\Delta < \alpha)} \int_0^1 \int_{\mathcal{L}_\gamma(\tilde{s})} P(\Delta < \alpha | X = \mathbf{x}) p_X(\mathbf{x}) d\mathbf{x} d\gamma. \quad (95)$$

We also have

$$\Delta(\hat{\pi}(\mathbf{x}), Y) < \frac{1}{2} \Leftrightarrow 1 - \hat{\pi}_Y < \frac{1}{2}, \quad (96)$$

$$\Leftrightarrow \hat{\pi}_Y > \frac{1}{2}, \quad (97)$$

$$\Leftrightarrow Y = \hat{y}. \quad (98)$$

This implies that

$$J\left(\frac{1}{2}\right) = \frac{1}{P(\Delta < \frac{1}{2})} \int P(Y = \hat{Y}, X \in \mathcal{L}_\gamma(\tilde{s})) d\gamma. \quad (99)$$

Because $P(\Delta < \frac{1}{2}) = P(Y = \hat{Y})$, we obtain

$$J\left(\frac{1}{2}\right) = \int P(X \in \mathcal{L}_\gamma(\tilde{s}) | Y = \hat{Y}) d\gamma, \quad (100)$$

$$> \int P(X \in \mathcal{L}_\gamma(\tilde{s}^*) | Y = \hat{Y}) d\gamma, \quad (101)$$

$$> J^*\left(\frac{1}{2}\right). \quad (102)$$

The penultimate inequality is obtained by leveraging the hypothesis that $s^* \prec_{\text{mbc}} s \Leftrightarrow \tilde{s}^* \prec_{\text{mbc}} \tilde{s}$ which involves the existence of a region $\mathcal{R}_{\tilde{s}^* \tilde{s}}$ with positive Lebesgue measure where $P(X \in \mathcal{L}_\gamma(\tilde{s}) | \hat{Y}_{\text{bay}} = \hat{Y}) > P(X \in \mathcal{L}_\gamma(\tilde{s}^*) | \hat{Y}_{\text{bay}} = \hat{Y})$. We now can conclude that, under this hypothesis, \tilde{s} would achieve a larger UQ-C-index than one maximizer of the latter which is not possible.

15 PROOF OF Theorem 3.1

Similarly as in Section 13, for (X, Y) and (X, Y') two random variable pairs i.i.d with respect to the data distribution $p_{X, Y}$, we denote $V = \mathbb{1}_{Y \neq \hat{Y} | X}$, $S = s(X)$ and $S' = s(X')$. Starting from the definition of G-AUC, and for any fixed scoring function s , one can write:

$$\text{G-AUC}(s) = P(V = 1 | S \leq S'). \quad (103)$$

Since for any i.i.d. couple of random variables, one has $P(S \leq S') = P(S' \leq S) = \frac{1}{2}$, we thus have

$$\text{AUC}(s) = 2P(V = 1, S \leq S'), \quad (104)$$

$$= 2P(V = 1, V' = 0, S \leq S') + 2P(V = 1, V' = 1, S \leq S'). \quad (105)$$

We can remark that

$$P(V = 1, V' = 1, S \leq S') = P(S \leq S' | V = 1, V' = 1) P(V = 1, V' = 1). \quad (106)$$

Similarly as stated above, since $(S | V = 1)$ and $(S' | V' = 1)$ are i.i.d, then $P(S \leq S' | V = 1, V' = 1) = P(S' \leq S | V = 1, V' = 1) = \frac{1}{2}$. In addition, we also have independence of V and V' and by definition of accuracy (for a fixed predictor \hat{y}), we thus obtain

$$P(V = 1, V' = 1, S \leq S') = \frac{1}{2} (1 - \text{ACC})^2. \quad (107)$$

Besides, one can also remark that

$$P(V = 1, V' = 0, S \leq S') = P(V' < V | S \leq S'), \quad (108)$$

$$= P(S \leq S' | V' < V) P(V' < V), \quad (109)$$

$$= \left(1 - \underbrace{P(S' < S | V' < V)}_{=1 - \text{UQ-AUC}(s)} \right) P(V' < V) \quad (110)$$

Moreover, we have

$$P(V' < V) = P(V' < 1 | V = 1) P(V = 1) + \underbrace{P(V' < 0 | V = 0)}_{=0} P(V = 0), \quad (111)$$

$$= P(V' = 0 | V = 1) P(V = 1), \quad (112)$$

$$= P(V' = 0) P(V = 1), \quad (113)$$

$$= \text{ACC}(1 - \text{ACC}). \quad (114)$$

Finally, by substituting (114) in (110) and then (110) and (107) in (105), one obtains (20).

Concerning H-AUC, by definition, one has

$$\text{G-AUC} = \int_0^{1-\text{ACC}} P\left(S \leq F_{S|V=1}^{-1}\left(\frac{t}{1-\text{ACC}}\right), V=0\right) dt, \quad (115)$$

where $F_{S|V=1}^{-1}$ is the inverse function of the cdf of $S|V=1$. Let us use the following change of variable under the integral:

$$u := F_{S|V=1}^{-1}\left(\frac{t}{1-\text{ACC}}\right), \quad (116)$$

$$\Leftrightarrow t = F_{S|V=1}(u)(1-\text{ACC}), \quad (117)$$

$$\text{and } dt = p_{S|V=1}(u)(1-\text{ACC}) du. \quad (118)$$

We therefore obtain

$$\text{G-AUC} = \int_{s_{\min}}^{s_{\max}} P(S \leq u, V=0) p_{S|V=1}(u)(1-\text{ACC}) du, \quad (119)$$

$$= (1-\text{ACC}) \text{ACC} \int_{s_{\min}}^{s_{\max}} F_{S|V=0}(u) p_{S|V=1}(u) du, \quad (120)$$

$$(121)$$

where s_{\min} and s_{\max} are respectively the minimal and maximal values taken by S . Finally, since UQ-AUC is also defined as $\int_{s_{\min}}^{s_{\max}} F_{S|V=0}(u) p_{S|V=1}(u) du$, one obtains (21).

16 ADJUSTED METRICS FOR OOD DETECTION

In this appendix, we examine to what extent some of the results of this paper for the use case of misclassification error prone input detection can be transposed to the OOD input detection use case.

We start with a situation in which a set of OOD data points is at hand. Since one can hold out a fraction of in-distribution (InD) points for evaluation purposes, it is thus possible to build a test dataset for which the label InD/OOD is available. Based on these labels, one can easily build an OOD analogue of ground truth ϕ and of UQ-AUC as done in prior arts. They will enjoy similar theoretical properties as those of Theorem 2.1, Theorem 2.3 and Theorem 2.4.

The definition of an OOD analogue of the UQ-C-index appears to be less obvious (if not ill-advised) as one needs to define a meaningful gap function $\Delta(\cdot)$ based on the network output $\hat{\pi}(\mathbf{x})$. One cannot directly compare softmax vectors with one-hot-encoding of the InD/OOD label. Consequently, it is necessary to first plug softmax vector into an OOD sensitive scoring function such as the free energy but this has the disadvantage of making the evaluation dependent of this choice. Another possibility is to train a companion model to discriminate InD from OOD data. Here again, the evaluation becomes dependent on modeling choices and training procedure.

In the more challenging situation where no OOD is available to carry out the evaluation of the UQ score, it can envisioned to rely on distance informed versions of some metrics. For example, a C-index could be defined based the distance $\Delta(\mathbf{x}) = \min_{\mathbf{x}' \in \mathcal{I}_{\text{test}} \setminus \{\mathbf{x}\}} d(\mathbf{x}, \mathbf{x}')$ where d is a distance on the input space and $\mathcal{I}_{\text{test}}$ is the set of inputs from the test set. For this approach to catch meaningful signals, the distance d should intuitively rely on the data geometry otherwise the curse of dimensionality will come into play. Moreover, such a C-index will not directly inherit from the results of Section 2.3 as those rely on the definition of the gap function Δ .

17 EXPERIMENTAL DETAILS

17.1 Toy dataset

Our experiments were carried out in around 10 hours of compute time on a 16-CPU and 256GB RAM machine where parallelization across CPUs was used. For the toy dataset, we sample scalars μ_0 and μ_1 from a 2-dimensional

Gaussian distribution $\mathcal{N}(0, \tau I_2)$ where I_2 is the 2-by-2 identity matrix, with τ is a positive scalar controlling the overlap between $\mathcal{N}(\mu_1, \sigma)$ and $\mathcal{N}(\mu_0, \sigma)$. Then for each data point $i \leq n$, we sample the label $y^{(i)}$ from a Bernoulli distribution of parameter p . Then we sample the input feature from the corresponding Gaussian, $X^{(i)} \sim p_{X|Y=y^{(i)}} = \mathcal{N}(\mu_{y^{(i)}}, \sigma)$.

This construction allows us to compute the exact input-conditional distribution:

$$\pi(\mathbf{x}) = P(Y = y|X = \mathbf{x}) = \frac{p_y p_{X|Y=y}(\mathbf{x})}{\sum_{y' \in \mathcal{Y}} p_{y'} p_{X|Y=y'}(\mathbf{x})}. \quad (122)$$

In the above, $p_1 = p$ and $p_0 = 1 - p$. Distribution $\pi(\mathbf{x})$ allows us to compute the ground truth $\phi(\mathbf{x})$ and $\varphi(\mathbf{x})$. We use $n = 1000$, $\tau = \sigma = 1$ and $p_0 = p_1 = 0.5$. We use 600 samples for training, and 400 for testing, stratified by label. The dataset is displayed in Figure 3a along with examples of ground truth functions computed for an arbitrary predictor in Figures 3b and 3c. This predictor uses $x_2^{(i)} - \frac{1}{n} \sum_{j=1}^n x_2^{(j)}$ as logits and its decision frontier is thus a horizontal line crossing the middle of the training sample. Finally, Figure 3d shows the compliance of those ground truths with Theorems 2.1 and 2.2.

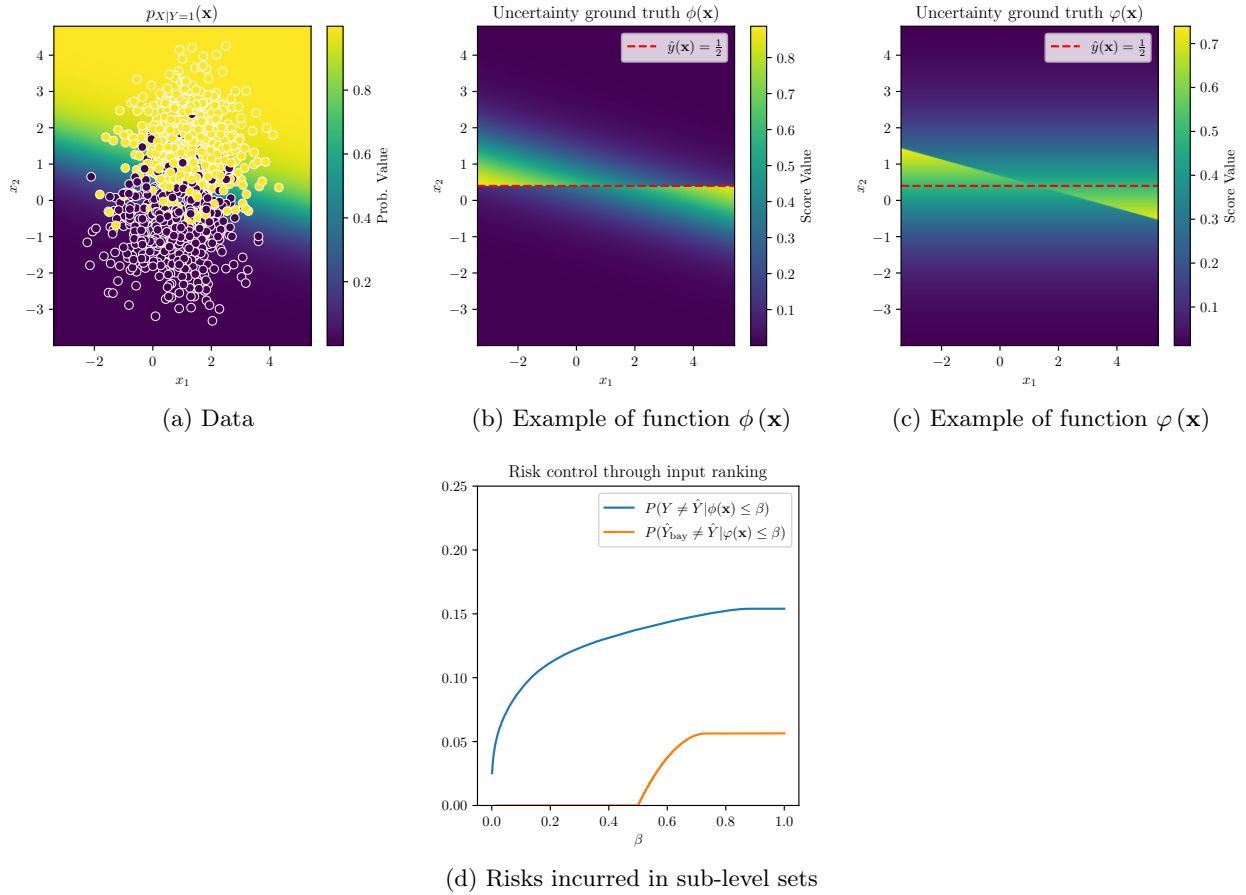


Figure 3: Synthetic dataset illustrations. (a) Each dot represents a sample, whose color indicates the label. The background color indicates the probability $P(Y = 1|X = \mathbf{x})$. (b) & (c) Examples of ground truth scoring ϕ and φ for an arbitrary predictor whose frontier decision is the dashed red line. (d) Different types of risk incurred in sub-level sets $\mathcal{L}_\beta(\phi)$ and $\mathcal{L}_\beta(\varphi)$ respectively. As implied by Theorems 2.1 and 2.2 those risks can be controlled by lowering the value of β .

Once the dataset has been sampled, we train an algorithm by minimizing the cross entropy, using hyper parameters described below.

Each algorithm is trained for 50 epochs, with 504 samples per batch. We used Adam optimizer, with learning

rate set to either 0.005, 0.025 or 0.05. The models are MLPs with one or two hidden layers, of size 64, 32 or 16. The tested combinations are [64, 32], [32, 32], [32, 16] and [64]. For deep ensemble we used 5 or 10 models, and for MC-Dropout we sample either 10, 50 or 100 predictions. We used a dropout rate set to either 0.1, 0.3 or 0.5.

Lastly, we vary the scoring function among those described in Section 7. For DEs and MCD we used the entropy, the aleatoric entropy, the free energy, the mutual information, and the variance of the predicted class. For the softmax baseline, we used only the entropy, the free energy and the variance of the prediction.

Once a model is trained, we compute the UQ score $s(\mathbf{x}^{(i)})$ and the ground truths $\phi(\mathbf{x}^{(i)})$ and $\varphi(\mathbf{x}^{(i)})$ for all samples in the test set, and then the Kendall correlation between those. Figure 4 and Figure 5 illustrate the distribution of the UQ scores versus each ground truth for several randomly-picked models. Figure 6 shows the same results as in Figures 1a and 1e in different subplots (one for each UQ algorithm and UQ metric).

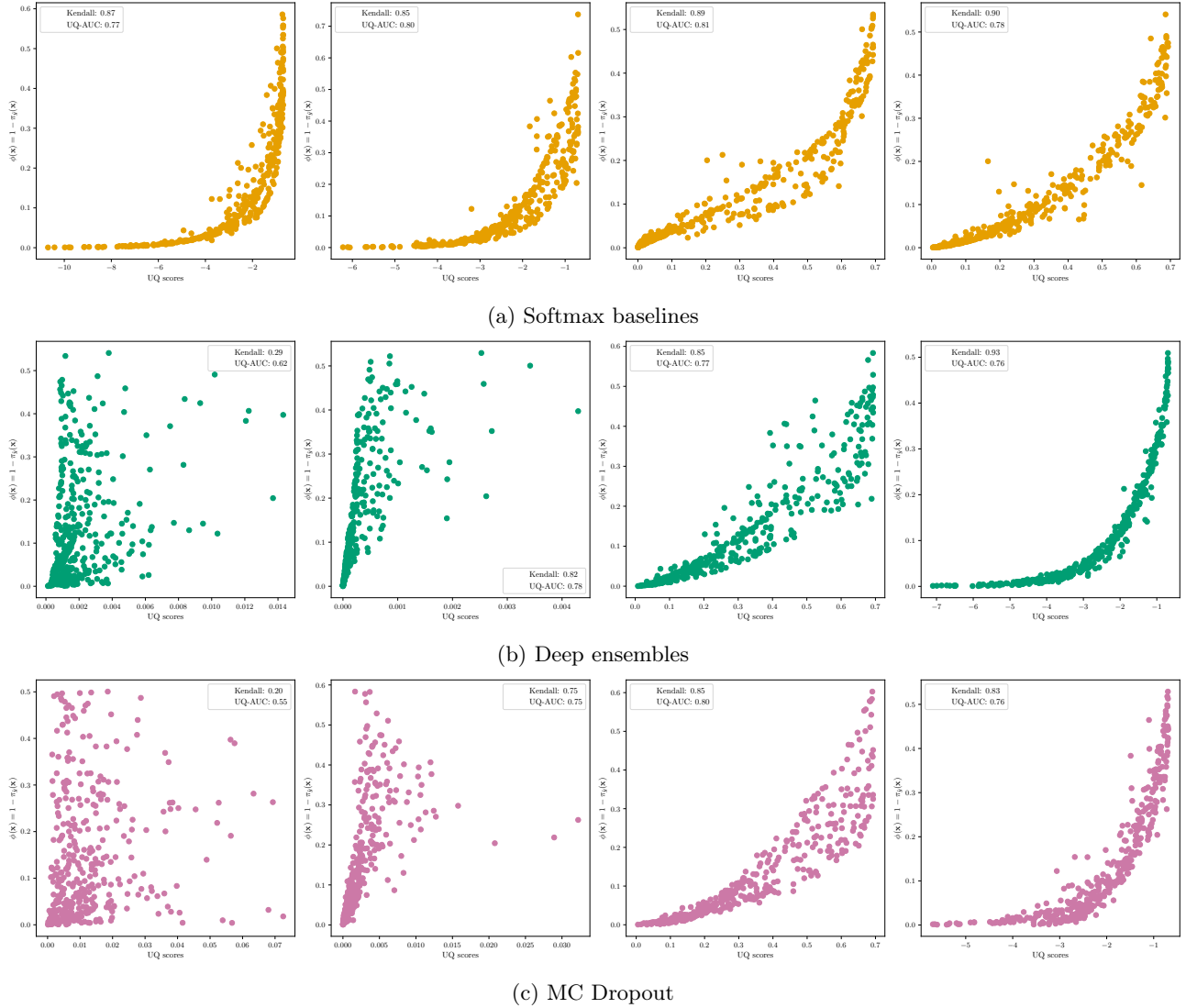


Figure 4: Examples of scatter plots of UQ scores $s(\mathbf{x})$ versus ground truth $\phi(\mathbf{x})$ for randomly-picked trained UQ models. First row corresponds to the SFT baseline, second one are Deep Ensemble algorithms and the last one MC dropout algorithms. Each plot corresponds to one point in Figure 1e

17.2 Real-world datasets

To scale our experiments to real-world datasets, we leverage two classical computer vision datasets with human annotations of uncertainty: CIFAR10-H (Peterson et al., 2019) and ReaL-ImageNet (Beyer et al., 2020). We also

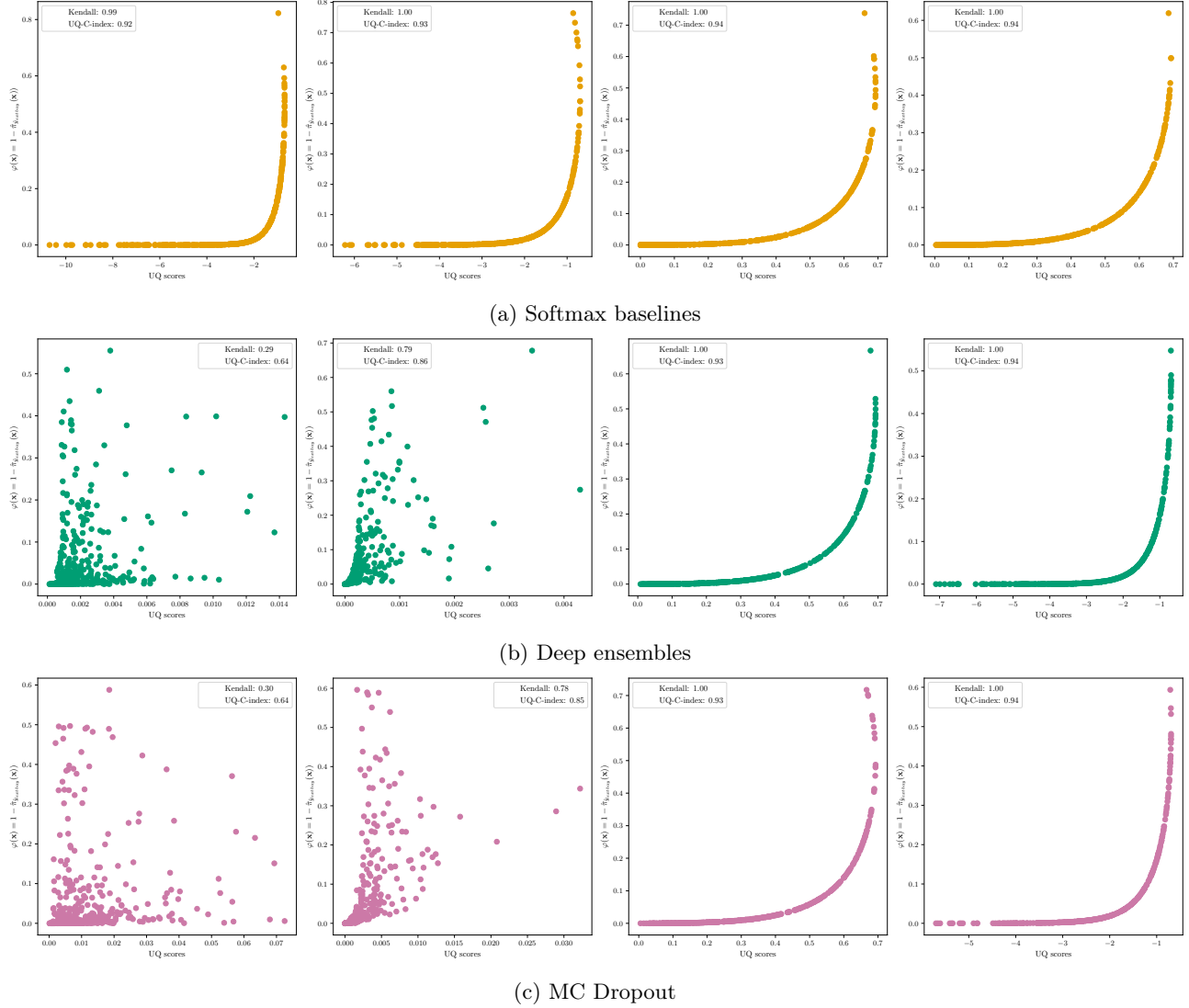


Figure 5: Examples of scatter plots of UQ scores $s(\mathbf{x})$ versus ground truth $\varphi(\mathbf{x})$ for randomly-picked trained UQ models. First row corresponds to the SFT baseline, second one are Deep Ensemble algorithms and the last one MC dropout algorithms. Each plot corresponds to one point in Figure 1a

leverage a large-scale dataset of text-to-text predictions with human judgments of uncertainty: OASST2 (Köpf et al., 2023). For CIFAR10-H, we use the same train/test split as in (Schmarje et al., 2022). The extra human annotations are obtained utilizing over 500k crowd-sourced human categorization judgments. We use the soft label distribution provided in (Schmarje et al., 2022). For the models, we leverage pre-trained ImageNet models available through the torchvision package. The complete list of models is provided in Section 17.2.1. To adapt these models to CIFAR10-H, we train linear probes on top of the penultimate layer of each network and report metrics computed on the validation folds. The training is conducted using SGD with a learning rate of 0.002 and momentum of 0.9 for 10 epochs with a batch size of 32. Probabilities $\hat{\pi}(\mathbf{x})$ are computed using the softmax function.

For ReaL-ImageNet, we use the pre-trained models without any fine-tuning and report metrics computed on the validation set from the original paper. For the soft label distribution, we utilize the provided scores for each rated (image, label) pair, as computed by the Dawid and Skene (1979) algorithm on the crowd-sourced votes. We normalize these scores to sum to one across labels, yielding the soft label distribution. For more details on the data collection process and the computation of these scores, see (Beyer et al., 2020).

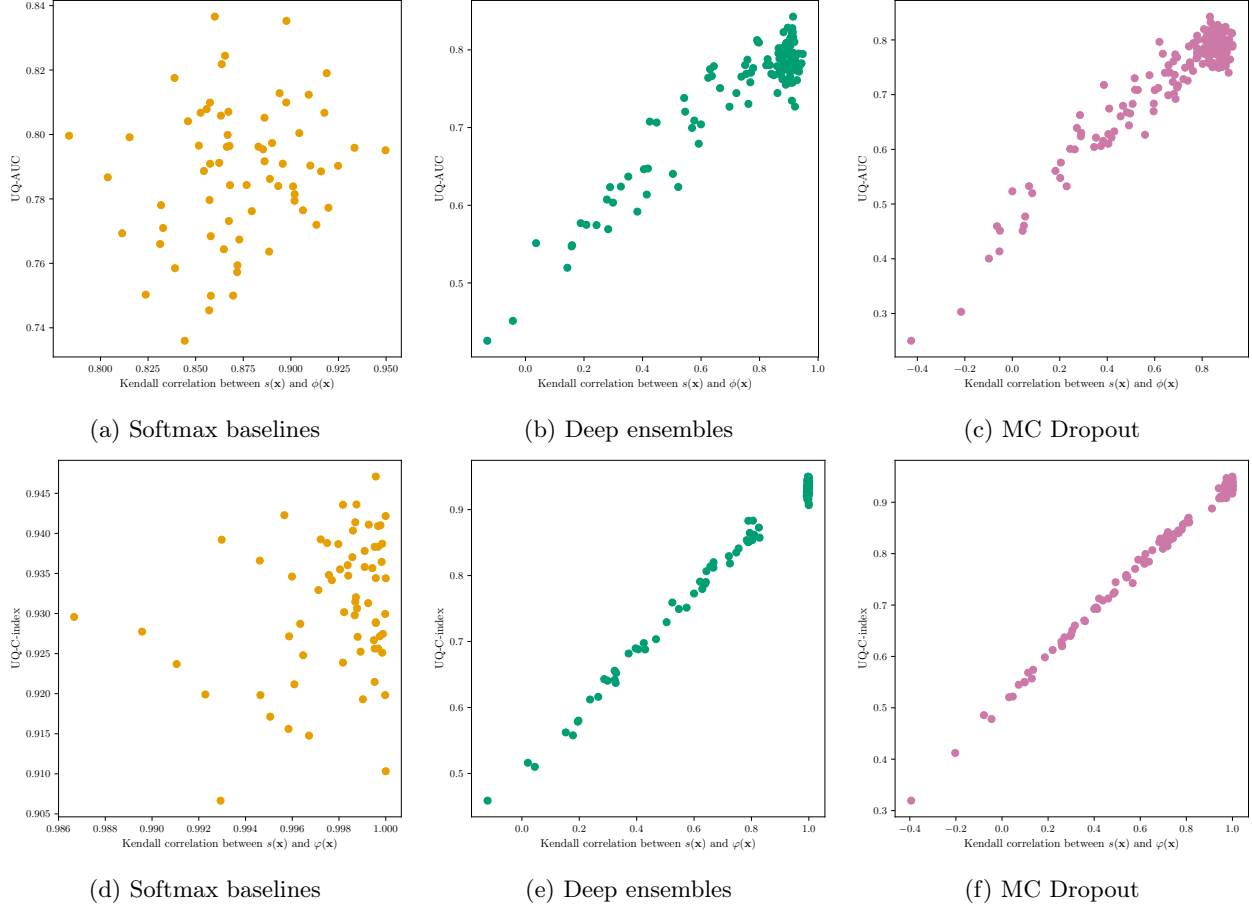


Figure 6: UQ-AUC (resp. UQ-C-index) versus Kendall correlation between $s(\mathbf{x})$ and $\phi(\mathbf{x})$ (resp. $\varphi(\mathbf{x})$) for several UQ algorithms. All metrics are computed on a held-out test set. Note that Kendall correlations are intractable in practice, contrary to the UQ-AUC or UQ-C-index.

The OASST2 dataset includes 6.6K text inputs for validation in various languages with several human annotations across 15 text content categories. Those categories were sorted into negative and positive labels to create a binary classification task. We performed inference using 70 pre-trained models available in the HuggingFace hub for zero-shot classification on the 15 categories. Likewise, model predictions were aggregated into softmax probabilities to address the binary classification task. For each model, we filtered out the sentences that were not in a language supported by the model.

As for the UQ scoring functions, we only consider the Entropy of the softmax output.

Computation were done on a single T4 GPU for ImageNet as only inference is needed, taking approximately 1 hour. For CIFAR10-H, we run each experiment on 15 T4 GPUs, taking approximately 15 hours. For OASST2, we run our experiments parallelly on 8 V100 GPUs, taking approximately 10 hours. Thanks to zero-shot classification, no training is needed.

17.2.1 List of models

Here are the list of models used in our experiments on CIFAR10-H and R-ImageNet. We used the torchvision implementation of those models, with the available pretrained weights on ImageNet.

- ResNet family: ResNet18, ResNet34, ResNet50, ResNet101, ResNet152
- ResNeXt family: ResNeXt50_32x4d, ResNeXt101_32x8d, ResNeXt101_64x4d
- Wide ResNet family: Wide_ResNet50_2, Wide_ResNet101_2

- VGG family: VGG11, VGG11_bn, VGG13, VGG13_bn, VGG16, VGG16_bn, VGG19, VGG19_bn
- EfficientNet family: EfficientNet_B0 to B7, EfficientNet_V2_S, EfficientNet_V2_M, EfficientNet_V2_L
- RegNet family: RegNet_X (400MF, 800MF, 1.6GF, 3.2GF, 8GF, 16GF, 32GF), RegNet_Y (400MF, 800MF, 1.6GF, 3.2GF, 8GF, 16GF, 32GF, 128GF)
- MobileNet family: MobileNet_V2, MobileNet_V3_Small, MobileNet_V3_Large
- ShuffleNet family: ShuffleNet_V2_x0.5, ShuffleNet_V2_x1.0, ShuffleNet_V2_x1.5, ShuffleNet_V2_x2.0
- DenseNet family: DenseNet121, DenseNet161, DenseNet169, DenseNet201
- Vision Transformer (ViT) family: ViT_B_16, ViT_B_32, ViT_L_16, ViT_L_32, ViT_H_14
- Swin Transformer family: Swin_T, Swin_S, Swin_B, Swin_V2_T, Swin_V2_S, Swin_V2_B
- ConvNeXt family: ConvNeXt_Tiny, ConvNeXt_Small, ConvNeXt_Base, ConvNeXt_Large
- MNASNet family: MNASNet0.5, MNASNet0.75, MNASNet1.0, MNASNet1.3
- Other models: AlexNet, GoogLeNet, Inception_V3

As for the models used for OASST2, we used the following models:

- BART family: BART-large-mnli, DistilBART-mnli-12-1, DistilBART-mnli-12-3, DistilBART-mnli-12-6, DistilBART-mnli-12-9
- BERT family: BERT-base-multilingual-cased, BERT-base-turkish-cased, BERT-base-spanish-wwm-cased, NB-BERT-base, MobileBERT-uncased
- RoBERTa family: RoBERTa-base, RoBERTa-large, DistilRoBERTa-base, XLM-RoBERTa-base, XLM-RoBERTa-large
- DeBERTa family: DeBERTa-v3-base, DeBERTa-v3-large, DeBERTa-v3-xsmall, mDeBERTa-v3-base
- DistilBERT family: DistilBERT-base-uncased, DistilBERT-base-multilingual-cased
- MiniLM family: MiniLM2-L6-H768, MiniLMv2-L6, MiniLMv2-L12
- E5 family: E5-base, E5-large, E5-large-instruct
- mContriever family: mContriever-msmarco
- Other models: FLAN-T5-base, CamemBERT-base, RuBERT-base, Megatron-BERT-large