Predictive Uncertainties Based on Proper Scoring Rules

Nikita Kotelevskii [1 2] **Maxim Panov** [2]

Abstract

This paper presents a theoretical framework for understanding uncertainty through the lens of statistical risks. It introduces a method to differentiate between aleatoric uncertainty, which is related to inherent data variability, and epistemic uncertainty, which is linked to lacking of best model parameters knowledge. We explain how pointwise risk can be decomposed into Bayes risk and Excess risk, showing that Excess risk, linked to epistemic uncertainty, corresponds to Bregman divergences. To convert these theoretical risk measures into practical uncertainty estimates, we propose using a Bayesian approach, approximating the risks through posterior distributions. We validate our method on image datasets, assessing its capability to identify out-of-distribution and misclassified data using the AUROC metric. Our findings demonstrate the efficacy of this approach and provide practical insights for estimating uncertainty in real-world scenarios.

1. Introduction

In the modern world, predictive models are applied in a variety of fields requiring high-risk decisions such as medical diagnosis (Shen et al., 2017; Litjens et al., 2017), finance (Ozbayoglu et al., 2020; Heaton et al., 2017), autonomous driving (Grigorescu et al., 2020; Mozaffari et al., 2020) and others. A careful analysis of model predictions is required to mitigate the risks. Hence, it is of high importance to evaluate the predictive uncertainty of the models. In recent years, a variety of approaches to quantify predictive uncertainty have been proposed (Kotelevskii et al., 2022; Kendall & Gal, 2017; Van Amersfoort et al., 2020; Liu et al., 2020; Lakshminarayanan et al., 2017; Malinin & Gales, 2021; Schweighofer et al., 2023). Specific attention has been paid to the distinction between different sources of uncertainty. It is commonly agreed to consider two sources (Hüllermeier & Waegeman, 2021) – the first one, **aleatoric** uncertainty, effectively reduces to the inherent ambiguity in label distribution, while the second one, **epistemic** uncertainty, is referred to as the uncertainty due to the "lack of knowledge". Distinguishing between aleatoric and epistemic uncertainties is crucial in practice because it helps identify whether uncertainty can be reduced by gathering more data (epistemic) or if it is inherent to the problem (aleatoric), thus guiding better decision-making and model improvement. Applications of uncertainty disentanglement include active learning (Beluch et al., 2018; Gal et al., 2017), out-of-distribution detection (Kotelevskii et al., 2022; 2023; Mukhoti et al., 2021), and misclassification detection (Vazhentsev et al., 2022).

Despite the practical importance and widespread usage of uncertainty quantification, *there is still no common strict formal definition of both types of uncertainty*. Essentially, various often ad hoc definitions are adopted that lead to even higher number of different measures to quantify either type of uncertainty (see for example (Lakshminarayanan et al., 2017; Gal et al., 2017; Malinin & Gales, 2021; Hüllermeier & Waegeman, 2021; Kotelevskii et al., 2022; Schweighofer et al., 2023)). However, it is not clear how all these measures of uncertainty are related to each other. Do they complement or contradict each other? Are they special cases of some general class of measures? In this paper, we are going to address these questions by introducing a proper statistical approach for predictive uncertainty quantification, reasoning in terms of pointwise risk estimation. Our contributions are as follows:

1. We suggest looking at pointwise risk, which is defined as the expected value of a loss function, as a natural measure of predictive uncertainty; see Section 2.
2. We consider a specific class of loss functions, namely proper scoring rules, that allow us to derive a unified treatment for the wide family of uncertainty measures; see Section 3. In particular, we demonstrate that Excess risk, serving as the measure of epistemic uncertainty, can be represented as Bregman divergence, and we show some special instances of it.
3. We incorporate Bayesian reasoning into our framework (see Section 4), showing that commonly used

[*]Equal contribution [1]CAIT, Skoltech, Moscow, Russia [2]ML department, MBZUAI, Abu Dhabi, UAE. Correspondence to: Nikita Kotelevskii <nikita.kotelevskii@skoltech.ru>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria.

measures of epistemic uncertainty, such as Bayesian Active Learning by Disagreement (BALD) (Houlsby et al., 2011; Gal et al., 2017) and Expected Pairwise Kullback-Leibler divergence (EPKL), are special cases within our general approach. We also highlight the limitations of our framework, elaborating on discussions from (Wimmer et al., 2023; Schweighofer et al., 2023).

4. We experimentally evaluate the proposed uncertainty measures in various tasks, including out-of-distribution detection and misclassification detection; see Section 6. Our results highlight the conditions under which each measure is most effective, providing practical insights for selecting appropriate uncertainty measures.

2. Predictive uncertainty quantification via risks

Assume we have a dataset $D_{tr} = \{(X_i, Y_i)\}_{i=1}^N$, where pairs $X_i \in \mathbb{R}^d, Y_i \in \mathcal{Y}$ are i.i.d. random variables sampled from a joint training distribution $P_{tr}(X, Y)$. We consider a classification task over K classes, i.e. $\mathcal{Y} = \{1, \dots, K\}$. We can express this joint distribution as a product: $P_{tr}(X, Y) = P_{tr}(Y \mid X)P_{tr}(X)$.

In practice, we typically consider a parametric model $P(Y \mid X, \theta)$ with parameters θ to approximate $P_{tr}(Y \mid X)$. We denote the true class probabilities for an input x as $\eta(x) = P_{tr}(Y \mid X = x)$, and the predicted probabilities as $\hat{\eta}_\theta(x) = P(Y \mid X = x, \theta)$. We will often omit the index θ and denote the predicted probability vector by $\hat{\eta}$.

2.1. Pointwise Risk as a Measure of Uncertainty

The goal of uncertainty quantification is to measure the degree of confidence of predictive models, distinguishing between aleatoric and epistemic sources of uncertainty. Despite its importance, there is no unified definition, leading to diverse methods and measures. In the paper, we introduce uncertainty via the statistical concept of risk.

In machine learning, the main concern is the model's "error" at a particular input point x. One way to express this error is through expected risk. Let $\ell\colon \mathbb{R}^K \times \mathcal{Y} \to \mathbb{R}$ be a loss function that measures how well $\hat{\eta}(x)$ matches the true label y. The pointwise risk $\mathrm{R}(\hat{\eta} \mid x)$ for a model $\hat{\eta}$ is defined as:

$$\mathrm{R}(\hat{\eta} \mid x) = \int \ell(\hat{\eta}(x), y)\, dP(y \mid X = x), \qquad (1)$$

Thus, pointwise risk is an expected loss received by a specific predictor $\hat{\eta}$ at a particular input point x. Importantly, pointwise risk, while being a natural measure of expected model error, can not be used directly as a measure of uncertainty as it is not possible to compute it due to unknown true data distribution. We will discuss possible ways to transform pointwise risk into the practical uncertainty measure in Section 4.

Note, that we use distribution P, which might differ from P_{tr}. If $P(X) \neq P_{tr}(X)$ but $P(Y \mid X = x) = P_{tr}(Y \mid X = x)$ for any x, this situation is called "covariate shift". We will assume $P(Y \mid X = x)$ is *valid* for any x, meaning it is a vector of length K regardless of the input. Limitations of this assumption are discussed in Section 7.

2.2. Aleatoric and Epistemic Uncertainties via Risks

Predictive uncertainty can be divided into two sources. **Aleatoric** uncertainty expresses the degree of ambiguity in data and does not depend on the model, being an inherent property of data given a particular choice of design. **Epistemic** uncertainty, which is vaguely defined, but typically associated with the "lack of knowledge" of choosing the right model parameters θ. Sometimes, when the source is not important, practitioners consider **total** uncertainty.

Pointwise risk allows for the following decomposition:

$$\underbrace{\mathrm{R}(\hat{\eta} \mid x)}_{\text{Total risk}} = \underbrace{\mathrm{R}_{\text{Bayes}}(x)}_{\text{Bayes risk}} + \underbrace{\mathrm{R}(\hat{\eta} \mid x) - \mathrm{R}_{\text{Bayes}}(x)}_{\text{Excess risk}}, \qquad (2)$$

where $\mathrm{R}_{\text{Bayes}}$ is the pointwise Bayes risk, defined as:

$$\mathrm{R}_{\text{Bayes}}(x) = \int \ell(\eta(x), y)\, dP(y \mid X = x).$$

The Bayes risk represents the expected error from the true data-generative process $\eta(x) = P(Y \mid X = x)$. It does not depend on the parameters of the model nor the choice of model architecture, and hence can be seen as a measure of *aleatoric* uncertainty. The second term in equation (2) is "Excess risk" and represents the difference between the risks computed for the approximation and for the true model at a given input point x. Thus, it naturally represents a lack of knowledge about the true data distribution, i.e. *epistemic* uncertainty. We note that decomposition (2) was previously considered in the context of uncertainty quantification in (Kotelevskii et al., 2022; Lahlou et al., 2022) but for specific loss functions and without the detailed analysis.

Although the decomposition (2) is useful, it doesn't provide much information about the properties of these risk functions in general cases. Therefore, we consider a specific class of loss functions, strictly proper scoring rules, which allows us to do a theoretical analysis.

3. Risks for Strictly Proper Scoring Rules

Strictly proper scoring rules (Gneiting & Raftery, 2007) represent a class of loss functions that ensure that the minimizing predictive distributions coincide with the data-generative distribution $P(Y \mid X)$. Let's say a forecaster can

A dog, has sufficient probability under P_{tr}. Conditional $\eta(x)$ is **meaningful**.

A cat, has sufficient probability under P_{tr}. Conditional $\eta(x)$ is **meaningful**.

A pigeon, has almost no probability mass under P_{tr}. Conditional $\eta(x)$ is **vague**.

Figure 1: The figure shows different examples of input objects in binary classification problem (cats vs dogs). The limitation of our approach is that $\eta(x) = P_{tr}(Y \mid X = x)$ should be defined even for objects with tiny mass under P_{tr}. See discussion in Section 7.

produce a vector of predicted probabilities $P \in \mathcal{P}_K$, where $\mathcal{P}_K$ is a space of discrete probability distributions over K classes. Then, $\ell(P, y)\colon \mathcal{P}_K \times \mathcal{Y} \to \mathbb{R}$ is the penalty the forecaster would have, given that event y is materialized. Its expected value with respect to some distribution Q we will denote as $\ell(P, Q) = \int \ell(P, y) dQ(y)$.

The scoring rule that for any $P, Q \in \mathcal{P}_K$ has the property that $\ell(P, Q) \geq \ell(Q, Q)$ with equality if and only if $P = Q$ is called *strictly proper* scoring rule. Under mild assumptions (see Theorem 3.2 in (Gneiting & Raftery, 2007)), any strictly proper scoring rule can be represented as:

$$\ell(\eta, i) = \langle G'(\eta)\,,\, \eta \rangle - G'_i(\eta) - G(\eta),$$

where $\langle .\,,\, . \rangle$ is a scalar product, $G\colon \mathcal{P}_K \to \mathbb{R}$ is a strictly convex function, and $G'(\eta) = \{G'_1(\eta), \dots G'_K(\eta)\}$ is a vector of element-wise subgradients.

Risk decompositions for strictly proper scoring rules. Here we present the general results for different types of risk (detailed derivations are in Appendix A).

- **Total Risk (Total Uncertainty)**:

$$\mathrm{R}_{\mathrm{Tot}}(\hat{\eta}_\theta \mid x) = \langle G'(\hat{\eta})\,,\, \hat{\eta} \rangle - G(\hat{\eta}) - \langle G'(\hat{\eta})\,,\, \eta \rangle. \quad (3)$$

  Note, that Total risk depends *linearly* on the true data generative distribution η.

- **Bayes Risk (Aleatoric Uncertainty)**:

$$\mathrm{R}_{\mathrm{Bayes}}(x) = -G(\eta). \quad (4)$$

  Note, that Bayes risk is a concave function of η, since function G is convex.

- **Excess risk (Epistemic Uncertainty)**:

$$\mathrm{R}_{\mathrm{Exc}}(\hat{\eta} \mid x) = D_G\big(\eta \parallel \hat{\eta}\big). \quad (5)$$

  Excess risk is a Bregman divergence (Bregman, 1967) denoted as $D_G\big(\eta \parallel \hat{\eta}\big)$, it is convex in η.

Specific Instances of Proper Scoring Rules. Different choices of the convex function G lead to different proper scoring rules. Table 1 shows the results for some popular cases often used in machine learning algorithms (see detailed derivations in Appendix B).

From Table 1, we see, that some of the risks correspond to well-known aleatoric uncertainty measures. For example, the Bayes risk for the Log score is given by the entropy of the predictive distribution. For the Zero-one score, this component is given by the so-called MaxProb, also widely applied (Geifman & El-Yaniv, 2017; Kotelevskii et al., 2022; Lakshminarayanan et al., 2017). For the Excess risk, we obtain different examples of Bregman divergence which lead to some well-known uncertainty measures when coupled with the Bayesian approach to risk estimation that we discuss in Section 4 below.

Estimating risks. The derived equations are useful but require access to the true data-generative distribution η, which is typically unknown.

One approach to deal with this problem was introduced in (Kotelevskii et al., 2022), where authors considered a specific model $\hat{\eta}_\theta$, namely Nadaraya-Watson kernel regression, as it has useful asymptotic properties to approximate Excess risk. Another approach, the DEUP (Lahlou et al., 2022) proposed a method for estimating Excess risk by directly training a model to predict errors. However, in general cases,

	Log score	Brier score	Zero-one score	Spherical score	Negative log score
$G(\eta)$	$\sum_{k=1}^K \eta_k \log \eta_k$	$-\sum_{k=1}^K \eta_k(1-\eta_k)$	$\max_k \eta_k - 1$	$\|\eta\|_2 - 1$	$-\sum_k \log \eta_k$
Aleatoric (Bayes risk)	$\mathbb{H}\eta$	$1 - \|\eta\|_2^2$	$1 - \max_k \eta_k$	$1 - \|\eta\|_2$	$\sum_k \log \eta_k$
Epistemic (Excess risk)	$D_{\text{KL}}[\eta\|\hat{\eta}]$	$\|\eta - \hat{\eta}\|_2^2$	$\max_k \eta_k - \eta_{\text{argmax}_k \hat{\eta}_k}$	$\|\eta\|_2 \left(1 - \langle \frac{\hat{\eta}}{\|\hat{\eta}\|_2}, \frac{\eta}{\|\eta\|_2} \rangle\right)$	$D_{\text{IS}}[\eta\|\hat{\eta}]$
Total (Risk)	$\text{CE}[\eta\|\hat{\eta}]$	$\|\eta - \hat{\eta}\|_2^2 - \|\eta\|_2^2 + 1$	$1 - \eta_{\text{argmax}_k \hat{\eta}_k}$	$1 - \|\eta\|_2 \langle \frac{\hat{\eta}}{\|\hat{\eta}\|_2}, \frac{\eta}{\|\eta\|_2} \rangle$	$\sum_{k=1}^K (\log \hat{\eta}_k - 1 + \frac{\eta_k}{\hat{\eta}_k})$

Table 1: Resulting expressions for different risks, computed for different strictly proper scoring rules. Dependence on x is omitted for clarity. D_{KL} stands for Kullback-Leibler divergence, CE for Cross-Entropy, and D_{IS} for Itakura–Saito distance. See Appendix B for full derivations.

it is hardly possible to derive these results. In this paper, we consider a Bayesian approach to approximate η that allows us to derive both well-known from the literature and new uncertainty measures based on one unified framework.

4. Bayesian Risk Estimation

The derived equations for risks depend on the true data generative distribution $\eta(x)$ and on some approximation of it $\hat{\eta}(x)$. In particular, $\eta(x)$ appears in all the risks and is unknown. One needs to deal with that to obtain a computable uncertainty measure. In the Bayesian paradigm, one considers a posterior distribution over model parameters $p(\theta \mid D_{tr})$ that immediately leads to a distribution over predictive distributions $\hat{\eta}_\theta \mid D_{tr}$. The goal of this section is to give a complete recipe for risk estimation under the Bayesian approach.

One can think of the risks as functions of $\eta(x)$ and $\hat{\eta}(x)$, namely $g\big(\eta(x), \hat{\eta}(x)\big)$, where g is a shortcut for any risk function. We can approximate the risks with the help of posterior distribution using one of two ideas:

1. **Bayesian averaging of risk.** For example, one can consider $\mathbb{E}_{p(\tilde{\theta}|D_{tr})} g(\hat{\eta}_{\tilde{\theta}}, \hat{\eta})$ to approximate an impact of the true model η. In a fully Bayesian paradigm, the same can be done with $\hat{\eta}$ leading to the fully Bayesian risk estimate $\mathbb{E}_{p(\tilde{\theta}|D_{tr})} \mathbb{E}_{p(\theta|D_{tr})}\, g(\hat{\eta}_{\tilde{\theta}}, \hat{\eta}_\theta)$.

2. **Bayesian model averaging.** Alternatively, one may use posterior predictive distribution given by $\hat{\eta}_{D_{tr}}(x) = \mathbb{E}_{p(\tilde{\theta}|D_{tr})} \hat{\eta}_{\tilde{\theta}}(x)$ and plug it in risk equations instead of η and/or $\hat{\eta}$.

In general, one can consider four approximations of any risk: $\tilde{g}^{(1,1)} = \mathbb{E}_{p(\tilde{\theta}|D_{tr})} \mathbb{E}_{p(\theta|D_{tr})}\, g(\hat{\eta}_{\tilde{\theta}}, \hat{\eta}_\theta)$, $\tilde{g}^{(1,2)} = \mathbb{E}_{p(\tilde{\theta}|D_{tr})}\, g(\hat{\eta}_{\tilde{\theta}}, \hat{\eta}_{D_{tr}})$, $\tilde{g}^{(2,1)} = \mathbb{E}_{p(\theta|D_{tr})}\, g(\hat{\eta}_{D_{tr}}, \hat{\eta}_\theta)$ and $\tilde{g}^{(2,2)} = g(\hat{\eta}_{D_{tr}}, \hat{\eta}_{D_{tr}})$, where by superscripts we denote the index of approximation ideas discussed above. In this section, we present the resulting formulas for the resulting total, aleatoric, and epistemic uncertainty measures with brief remarks on different approximations. For detailed derivations and discussions, refer to Appendix C.

For **Total risk** we obtain the following equations:

$$\tilde{\text{R}}^{(1)}_{\text{Tot}}(x) = \mathbb{E}_{p(\theta|D_{tr})}\Big(\langle G'(\hat{\eta}_\theta)\,,\, \hat{\eta}_\theta\rangle - G(\hat{\eta}_\theta) - \langle G'(\hat{\eta}_\theta)\,,\, \hat{\eta}_{D_{tr}}\rangle\Big),$$

and

$$\tilde{\text{R}}^{(2)}_{\text{Tot}}(x) = \langle G'(\hat{\eta}_{D_{tr}})\,,\, \hat{\eta}_{D_{tr}}\rangle - G(\hat{\eta}_{D_{tr}}) - \langle G'(\hat{\eta}_{D_{tr}})\,,\, \hat{\eta}_{D_{tr}}\rangle = -G(\hat{\eta}_{D_{tr}}).$$

Note, that due to the linear dependence on η, approximations for the ground truth led to the same result of $\hat{\eta}_{D_{tr}}$. The two options above aim to approximate the predictive distribution. The first approximation leads to the Expected Pairwise Proper Scoring Rule (in particular, expected pairwise Cross-Entropy). The second approximation leads to some concave function of $\hat{\eta}_{D_{tr}}$.

For **Bayes risk** one also obtains only two cases as it doesn't depend on $\hat{\eta}$:

$$\tilde{\text{R}}^{(1)}_{\text{Bayes}}(x) = -\mathbb{E}_{p(\tilde{\theta}|D_{tr})} G(\hat{\eta}_{\tilde{\theta}}) \quad \text{and} \quad \tilde{\text{R}}^{(2)}_{\text{Bayes}}(x) = -G(\hat{\eta}_{D_{tr}}).$$

Interestingly, $\tilde{\text{R}}^{(2)}_{\text{Bayes}}(x) = \tilde{\text{R}}^{(2)}_{\text{Tot}}(x)$, which effectively means zero Excess risk. In the Appendix C, we provide a reasoning for which approximation is better for Bayes risk approximation.

For **Excess risk** we obtain the whole family of approximations:

- **Expected Pairwise Bregman Divergence (EPBD):**

$$\tilde{\mathrm{R}}^{(1,1)}_{\mathrm{Exc}}(x) = \mathbb{E}_{p(\tilde{\theta}|D_{tr})}\mathbb{E}_{p(\theta|D_{tr})} D_G\Big(\hat{\eta}_{\tilde{\theta}} \parallel \hat{\eta}_{\theta}\Big),$$

  In a special case of Log score, it is an Expected Pairwise KL (EPKL; (Malinin & Gales, 2021; Schweighofer et al., 2023)).

- **Bregman Information (BI):**

$$\tilde{\mathrm{R}}^{(1,2)}_{\mathrm{Exc}}(x) = \mathbb{E}_{p(\tilde{\theta}|D_{tr})} D_G\Big(\hat{\eta}_{\tilde{\theta}} \parallel \hat{\eta}_{D_{tr}}\Big),$$

  In a special case of Log score, it is Mutual Information (also known as BALD (Gal et al., 2017; Houlsby et al., 2011)).

- **Reverse Bregman Information (RBI):**

$$\tilde{\mathrm{R}}^{(2,1)}_{\mathrm{Exc}}(x) = \mathbb{E}_{p(\theta|D_{tr})} D_G\Big(\hat{\eta}_{D_{tr}} \parallel \hat{\eta}_{\theta}\Big),$$

  Its special case for Log score is known as Reverse Mutual Information (RMI; (Malinin & Gales, 2021)).

- Finally, we obtain that

$$\tilde{\mathrm{R}}^{(2,2)}_{\mathrm{Exc}}(x) = D_G\Big(\hat{\eta}_{D_{tr}} \parallel \hat{\eta}_{D_{tr}}\Big) = 0,$$

  which is coherent with the result obtained for the Total risk, when Excess risk (epistemic uncertainty) is equal to 0.

We observe that the general approach presented in this work allows us to obtain many existing uncertainty measures in the case of the Log score loss function while leading to the whole family of new measures (see Table 1). We refer to Appendix C for additional discussion and to Appendix D for the discussion of connections of these approximations to each other.

5. Related Work

The field of uncertainty quantification for predictive models, especially neural networks, has seen rapid advancements in recent years. Among these, methods allowing explicit uncertainty disentanglement are particularly interesting due to the ability to use estimates of different sources of uncertainty in various downstream tasks. For instance, epistemic uncertainty is effective in out-of-distribution data detection (Hüllermeier & Waegeman, 2021; Kotelevskii et al., 2023; Mukhoti et al., 2021) and active learning (Beluch et al., 2018; Gal et al., 2017). In contrast, aleatoric uncertainty, which is associated with label noise, is useful in misclassification detection (Vazhentsev et al., 2022).

Bayesian methods have become popular because they naturally handle distributions of model parameters, leading to prediction uncertainty. Exact Bayesian inference is very computationally expensive (Izmailov et al., 2021), so many lightweight versions are used in practice (Gal & Ghahramani, 2016; Thin et al., 2021; 2020; Blei et al., 2017; Lakshminarayanan et al., 2017). Early approaches inspired by Bayesian ideas (Gal et al., 2017; Kendall & Gal, 2017; Lakshminarayanan et al., 2017) used information-based measures like BALD (Houlsby et al., 2011) to quantify epistemic uncertainty and measures like entropy or maximum probability for aleatoric uncertainty. These methods, despite different computational costs, are widely used in the field.

However, the practical expense of Bayesian inference, even in its approximate forms, has led to the introduction of more simplified approaches. Some of these methods leverage hidden neural network representations, considering distances in their hidden space as a proxy for epistemic uncertainty estimation (Van Amersfoort et al., 2020; Liu et al., 2020; Kotelevskii et al., 2022; Mukhoti et al., 2021). While they offer the advantage of requiring only a single pass over the network, their notion of epistemic uncertainty, often linked to the distance of an object's representation to training data, captures only a part of the full epistemic uncertainty. Despite this limitation, their efficiency and effectiveness in out-of-distribution detection have made them widely used.

Another class aimed at simplifying Bayesian inference are the so-called second-order models (Sensoy et al., 2018; Malinin & Gales, 2018; 2019; Charpentier et al., 2020; 2021; Kotelevskii et al., 2023; Sale et al., 2023b;a). These models induce a conjugate prior distribution to the predictive distribution but lack a universally accepted approach for uncertainty quantification. Some methods within this class assess the flatness of the predicted second-order distribution as a proxy for epistemic uncertainty (Malinin & Gales, 2018; 2021; Charpentier et al., 2020; 2021), while others introduce a reference distribution, considered to be devoid of specific uncertainty components, and measure distances to them as uncertainty metrics (Sale et al., 2023b;a).

Despite the diversity of these approaches, the arbitrary nature of choosing uncertainty measures has led to ambiguity in understanding uncertainty. This paper aims to address this gap by proposing a unified framework that not only categorizes these diverse methods but also offers a more comprehensive understanding of uncertainty quantification.

Related paper (Adlam et al., 2022) is close to our work and it looks at Bregman divergence loss functions and their decomposition but does not cover uncertainty quantification or Bayesian approximation.

6. Experiments

In this section, we test different uncertainty measures derived from our general framework. Any Bayesian method

that produces multiple samples of model weights (or parameters of the first-order distribution) can be used to compute our proposed measures. For example, Hamiltonian Monte Carlo (HMC) (Neal, 2012) are considered in the Bayesian approximation literature as a "gold standard". However, they typically require a lot of steps to converge, and hence becoming impractical in real scenarios. For this reason, deep ensembles are considered the "practical gold standard" in uncertainty quantification (Lakshminarayanan et al., 2017). Therefore, we use deep ensembles, trained with various strictly proper scoring rules as loss functions for our experiments.

As training (in-distribution) datasets, we consider CIFAR10 and CIFAR100 (Krizhevsky et al., 2009), along with manually created noisy versions where some labels are randomly shuffled (see Section E for details). We evaluate the proposed measures of uncertainty by focusing on two specific problems: *out-of-distribution detection* and *misclassification detection*. For both problems, we use deep ensembles with 20 members. Each ensemble member shares the same architecture but differs due to randomness in initialization and training. We consider two architectures: VGG19 (Simonyan & Zisserman, 2014) and ResNet18 (He et al., 2016) (additional details can be found in Section F).

Our experimental evaluation aims to answer the following questions:

1. Does training with a specific loss function leads to better uncertainty estimates when using the same proper scoring rule for uncertainty measures? For example, does training with Log score (cross-entropy) and evaluating uncertainty with risks based on Log score yield better results than using other scores for uncertainty quantification?
2. Is Excess risk always better than Bayes risk for out-of-distribution detection?
3. Is Bayes risk always better than Excess risk for misclassification detection?
4. Does the choice of approximation strategy matter?

We emphasize that the goal of our experimental evaluation is *not to provide new state-of-the-art measures or compete with other known approaches* for uncertainty quantification. Instead, we aim to *verify whether different uncertainty estimates accurately quantify specific types of uncertainty.*

6.1. Impact of Matching Scoring Rules on Uncertainty Measures

In this section, we address the first question. Our results are shown in Figure 2, which presents the distribution of the area under the ROC curve (AUROC) for different problems when using matching and non-matching scoring rules, averaged over various datasets, risks, and architectures. Note that we excluded two loss functions from the training: the Zero-one score, which is not differentiable, and the Negative log score (Neglog), which is unstable during training.

Our observations indicate that using mismatched scoring rules can lead to poor results for both out-of-distribution detection and misclassification detection. Conversely, when the loss function used during training matches the proper scoring rule applied for uncertainty measures, the results are consistently good. For a more detailed analysis, refer to Appendix G.

Therefore, while the answer to the first question is not absolute, our findings strongly support using matching scoring rules, i.e. the same scoring rule for the training of the model and uncertainty quantification. Using non-matching scoring rules can result in suboptimal performance in some cases.

6.2. Evaluating Excess Risk and Bayes Risk for Out-of-Distribution Detection

In this section, we evaluate different types of risks to identify out-of-distribution (OOD) samples. Since the uncertainty associated with OOD detection is epistemic, we expect that Excess risk and Total risk will perform well for this task, while Bayes risk will likely fail.

We used CIFAR10 and CIFAR100 as in-distribution datasets and SVHN (Netzer et al., 2011) and blurred versions of CIFAR10 and CIFAR100 as out-of-distribution datasets (details in Appendix E). The Gaussian blur augmentation results in "soft-OOD" samples that maintain the same ground-truth labels and meaningful prediction probability vectors.

To evaluate the metrics, we calculated them on both in-distribution and out-of-distribution objects and then computed the AUROC of the resulting ordering. The results are presented in Table 2. Note that the columns for Bayes and Total risks are averaged over all approximations. The Excess risk is averaged over all approximations, except Inner Inner, as it is equal to 0 (see separate evaluation in Appendix J). The results shown are for the case where the proper scoring rules used for loss function and uncertainty estimation match.

We distinguish between two types of out-of-distribution data: "soft-OOD" and "hard-OOD". "Soft-OOD", such as the blurred versions of CIFAR10 and CIFAR100, have predicted probability vectors that are still meaningful. "Hard-OOD" samples, however, have completely non-informative predicted probability vectors (see Section 7 for discussion). For example, when a set of classes is considered during training, but an incoming image does not belong to those classes, the resulting probability distribution over training

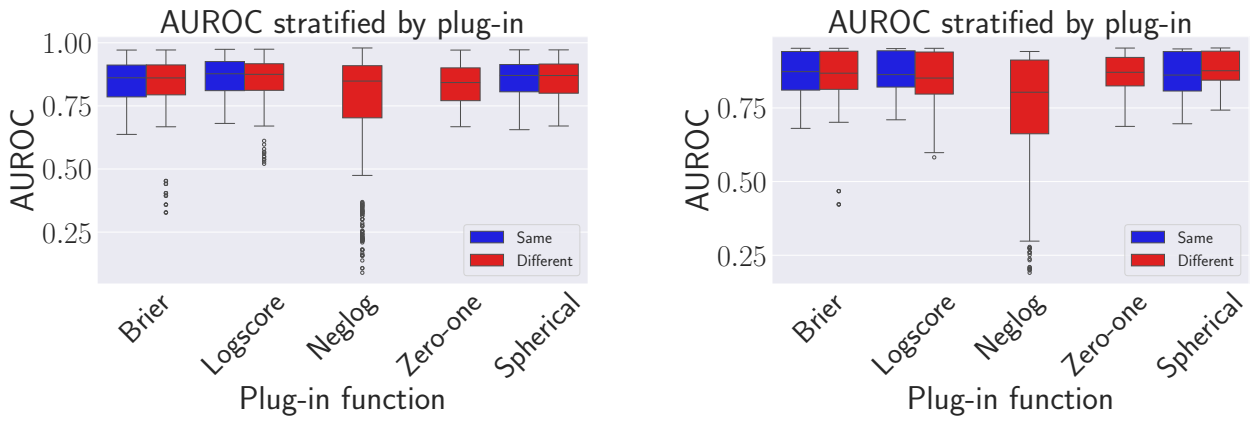

Figure 2: **Left:** AUROC for out-of-distribution detection. **Right:** AUROC for misclassification detection. Results are stratified by the plug-in proper scoring rule used for risk estimations. The legend indicates "Same" for the matching scoring rules and "Different" for the non-matching scoring rules.

classes is meaningless.

For "soft-OOD" datasets, the results meet our expectations: Excess risk (which includes BI, RBI, and EPBD) and Total risk perform well. However, for "hard-OOD" datasets, the results are unexpected—Bayes risk typically outperforms Excess risk.

One possible explanation, discussed in detail in Section 7, is that for "hard-OOD" samples, the target distribution $p(y \mid x)$ is not well-defined. Our Bayesian approximation of this distribution seems suboptimal in these cases. In contrast, for "soft-OOD" samples, the approximation remains reasonable, making Excess risk and Total risk effective.

This highlights a **crucial limitation of Excess risk (including BI, RBI, and EPBD) as a measure of epistemic uncertainty.** These measures naturally appear when approximating Excess risk in a Bayesian way, which assumes a specific form of ground-truth distribution approximation. However, this approximation becomes inaccurate for "hard-OOD" samples, making these measures a poor choice in these cases. This criticism aligns with findings from (Wimmer et al., 2023; Schweighofer et al., 2023; Bengs et al., 2023), which indicate that Bregman Information (in a particular case of Log score) is not an intuitive measure of epistemic uncertainty and does not follow their proposed axioms.

Therefore, the answer to the second question is not absolute as well. For "soft-OOD" samples, where predicted probability vectors remain meaningful, Excess risk is a good choice. For "hard-OOD" samples, Bayes risk might be better. Total risk consistently shows decent results, making it a safe choice when the nature of incoming data is unknown. A comparison of matching/not-matching results for out-of-distribution detection is in Appendix H.

6.3. Is Bayes Risk Always Better than Excess Risk for Misclassification Detection?

Now we consider misclassification detection. As in the previous experiment, we present AUROC obtained for different measures of uncertainty. The results are shown in Table 3, which is structured similarly to Table 2.

Misclassification detection is intuitively connected to aleatoric uncertainty. Therefore, we expect Bayes risk and Total risk to perform well in this task, while all instances of Excess risk should perform worse. It is known that standard versions of CIFAR10 and CIFAR100 lack significant aleatoric uncertainty (Kapoor et al., 2022), making it challenging to demonstrate the usefulness of the appropriate uncertainty measures immediately. To address this, we include noisy versions of these standard datasets, as well as datasets with missing classes, to introduce more noise (see details in Appendix E).

As expected, Bayes risk and Total risk outperform Excess risk for misclassification detection, except for the clean CIFAR10 dataset where the results are very close. Moreover, the difference in performance becomes more significant as more aleatoric noise is introduced into the training dataset.

Hence, the answer to this question is mostly positive. Bayes risk and Total risk are better for misclassification detection, and their advantage becomes more significant with the increase in aleatoric noise in the datasets. A comparison of matching/not-matching results for misclassification detection is in Appendix I.

Dataset				
InD	OOD	Bayes	Excess	Total
CIFAR10	Blurred CIFAR10(*)	84.38	**87.80**	85.90
	Blurred CIFAR100	94.48	**95.65**	95.42
	CIFAR100	**91.05**	90.08	90.87
	SVHN	**94.54**	93.31	94.44
CIFAR100	Blurred CIFAR10	87.96	85.36	**90.32**
	Blurred CIFAR100(*)	71.40	**77.55**	74.15
	CIFAR10	**79.35**	72.70	79.27
	SVHN	**84.90**	73.99	84.59

Table 2: AUROC for OOD detection. Best results in **bold**, second-best underline. By asterisk (*) we denote "soft-OOD".

Dataset	Bayes	Excess	Total
CIFAR10	94.53	94.65	**94.78**
CIFAR100	86.47	82.90	**86.83**
Missed class CIFAR10	**93.73**	83.23	91.14
Noisy CIFAR10	**81.00**	74.35	80.97
Noisy CIFAR100	**82.64**	72.30	82.45

Table 3: AUROC for misclassification detection. Best results in **bold**, second-best underline.

6.4. Does the Choice of Approximation Strategy Matter?

Due to space limitations, we provide a detailed discussion in Appendix J. In summary, our experiments show that the choice of approximation strategy is not always crucial. Different approximation strategies can behave differently across various datasets and experiments, making it difficult to conclusively determine if one is better than another.

7. Limitations

We see two limitations to our approach.

Valid conditional $\eta(x) = p(y \mid x)$ for all x. This assumption implies, that regardless of the input x, the form of the probability distribution $\eta(x)$ will not change. This means, that even for inputs, that do not belong to $P_{tr}(X)$, the conditional should produce some categorical vector over the same number of classes. Let us consider an example of a binary classification problem, where we want our model to distinguish between cats and dogs (see Figure 1). In this case, the distribution of covariates $P_{tr}(X)$ is the distribution over images of all possible cats and dogs. An image of a pigeon under this distribution should have a negligible probability. Now imagine, that somehow it happened that x is actually an image of a pigeon. Under our assumption, $\eta(x)$ should be valid, so it should produce a vector of probabilities over two classes: Cats and Dogs, despite an input being an apparent out-of-distribution object. There is no good way to define $\eta(x)$ for such input objects, hence we say it is vague. However, for unusual (but still in-distribution) inputs, like rare dog breeds, $\eta(x)$ is meaningful.

Incorporation of Bayesian reasoning for estimation of η. In practice, we do not have access to $\eta(x)$. Hence, we suggested approximating it using the Bayesian approach and proposed two ideas to do it (inner and outer expectations). For Bayes risk the best Bayesian estimate is given by outer expectation (see Appendix C). However, this is not the case for Excess risk. It appears (see discussion in Appendix C) that Excess risk depends on the estimate of the Total risk. But we never know in advance for a particular input x, in which regime (overestimated or underestimated Total risk) we are. Thus, we do not know what is the best choice for an approximation to epistemic uncertainty.

8. Conclusion

In this paper, we developed a general framework for predictive uncertainty estimation using pointwise risk estimation and strictly proper scoring rules as loss functions. We proposed pointwise risk as a natural measure of predictive uncertainty and derived general results for total, epistemic, and aleatoric uncertainties, demonstrating that epistemic uncertainties can be represented as a Bregman divergence within this framework.

We incorporated Bayesian reasoning into our framework, showing that commonly used measures of epistemic uncertainty, such as Bayesian Active Learning by Disagreement (BALD) and Expected Pairwise Kullback-Leibler divergence (EPKL), are special cases within our general approach. We also discussed the limitations of our framework, elaborating on recent critiques in the literature (Wimmer et al., 2023; Schweighofer et al., 2023).

Finally, in our experiments on image datasets, we evaluated these measures for out-of-distribution detection and misclassification detection tasks and discussed which measures are most suitable for each scenario.

References

Adlam, B., Gupta, N., Mariet, Z., and Smith, J. Understanding the bias-variance tradeoff of bregman divergences. *arXiv preprint arXiv:2202.04167*, 2022.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

Bengs, V., Hüllermeier, E., and Waegeman, W. On second-order scoring rules for epistemic uncertainty quantification. In *International Conference on Machine Learning*, pp. 2078–2091. PMLR, 2023.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3): 200–217, 1967.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020.

Charpentier, B., Borchert, O., Zügner, D., Geisler, S., and Günnemann, S. Natural posterior network: Deep bayesian predictive uncertainty for exponential family distributions. In *International Conference on Learning Representations*, 2021.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.

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

Heaton, J. B., Polson, N. G., and Witte, J. H. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 2017.

Houlsby, N., Huszár, F., Ghahramani, Z., and Lengyel, M. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

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

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. G. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pp. 4629–4640. PMLR, 2021.

Kapoor, S., Maddox, W. J., Izmailov, P., and Wilson, A. G. On uncertainty, tempering, and data augmentation in bayesian classification. *Advances in Neural Information Processing Systems*, 35:18211–18225, 2022.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kotelevskii, N., Artemenkov, A., Fedyanin, K., Noskov, F., Fishkov, A., Shelmanov, A., Vazhentsev, A., Petiushko, A., and Panov, M. Nonparametric uncertainty quantification for single deterministic neural network. *Advances in Neural Information Processing Systems*, 35:36308–36323, 2022.

Kotelevskii, N., Horváth, S., Nandakumar, K., Takáč, M., and Panov, M. Dirichlet-based uncertainty quantification for personalized federated learning with improved posterior networks. *arXiv preprint arXiv:2312.11230*, 2023.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lahlou, S., Jain, M., Nekoei, H., Butoi, V. I., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. Deup: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2022.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Malinin, A. and Gales, M. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Malinin, A. and Gales, M. Uncertainty estimation in autoregressive structured prediction. *In International Conference on Learning Representations.*, 2021.

Mozaffari, S., Al-Jarrah, O. Y., Dianati, M., Jennings, P., and Mouzakitis, A. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):33–47, 2020.

Mukhoti, J., van Amersfoort, J., Torr, P. H., and Gal, Y. Deep deterministic uncertainty for semantic segmentation. *arXiv preprint arXiv:2111.00079*, 2021.

Neal, R. M. Mcmc using hamiltonian dynamics. *arXiv preprint arXiv:1206.1901*, 2012.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A. Y., et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*. Granada, Spain, 2011.

Ozbayoglu, A. M., Gudelek, M. U., and Sezer, O. B. Deep learning for financial applications: A survey. *Applied Soft Computing*, 93:106384, 2020.

Pfau, D. A generalized bias-variance decomposition for bregman divergences. *Unpublished manuscript*, 2013.

Sale, Y., Bengs, V., Caprio, M., and Hüllermeier, E. Second-order uncertainty quantification: A distance-based approach. *arXiv preprint arXiv:2312.00995*, 2023a.

Sale, Y., Hofman, P., Wimmer, L., Hüllermeier, E., and Nagler, T. Second-order uncertainty quantification: Variance-based measures. *arXiv preprint arXiv:2401.00276*, 2023b.

Schweighofer, K., Aichberger, L., Ielanskyi, M., and Hochreiter, S. Introducing an improved information-theoretic measure of predictive uncertainty. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Shen, D., Wu, G., and Suk, H.-I. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

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

Thin, A., Kotelevskii, N., Denain, J.-S., Grinsztajn, L., Durmus, A., Panov, M., and Moulines, E. Metflow: a new efficient method for bridging the gap between markov chain monte carlo and variational inference. *arXiv preprint arXiv:2002.12253*, 2020.

Thin, A., Kotelevskii, N., Doucet, A., Durmus, A., Moulines, E., and Panov, M. Monte carlo variational auto-encoders. In *International Conference on Machine Learning*, pp. 10247–10257. PMLR, 2021.

Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pp. 9690–9700. PMLR, 2020.

Vazhentsev, A., Kuzmin, G., Shelmanov, A., Tsvigun, A., Tsymbalov, E., Fedyanin, K., Panov, M., Panchenko, A., Gusev, G., Burtsev, M., et al. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8237–8252, 2022.

Wimmer, L., Sale, Y., Hofman, P., Bischl, B., and Hüllermeier, E. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence*, pp. 2282–2292. PMLR, 2023.

A. Derivations of different risks with proper scoring rules

We will start with the derivation of Total risk. In what follows, we will omit dependency on x for $\eta(x)$ and $\hat{\eta}_\theta(x)$.

$$\mathrm{R}_{\text{Tot}}(\hat{\eta}_\theta \mid x) = \int \ell(\hat{\eta}_\theta, y) dP(y \mid x) = \sum_{k=1}^{K} \Big(\langle G'(\hat{\eta}_\theta)\,,\, \hat{\eta}_\theta \rangle - G'_k(\hat{\eta}_\theta) - G(\hat{\eta}_\theta) \Big) \eta_k = \\ \langle G'(\hat{\eta}_\theta)\,,\, \hat{\eta}_\theta \rangle - G(\hat{\eta}_\theta) - \langle G'(\hat{\eta}_\theta)\,,\, \eta \rangle.$$

Let us now consider Bayes risk:

$$\mathrm{R}_{\text{Bayes}}(x) = \int \ell(\eta, y) dP(y \mid x) = \sum_{k=1}^{K} \Big(\langle G'(\eta)\,,\, \eta \rangle - G'_k(\eta) - G(\eta) \Big) \eta_k = \\ \langle G'(\eta)\,,\, \eta \rangle - \langle G'(\eta)\,,\, \eta \rangle - G(\eta) = -G(\eta).$$

Finally, let us consider Excess risk:

$$\mathrm{R}_{\text{Exc}}(\hat{\eta}_\theta \mid x) = \underbrace{\int \ell(\hat{\eta}_\theta, y) dP(y \mid x)}_{\text{Total risk}} - \underbrace{\int \ell(\eta, y) dP(y \mid x)}_{\text{Bayes risk}} = \\ \langle G'(\hat{\eta}_\theta)\,,\, \hat{\eta}_\theta \rangle - G(\hat{\eta}_\theta) - \langle G'(\hat{\eta}_\theta)\,,\, \eta \rangle + G(\eta) = \\ G(\eta) - G(\hat{\eta}_\theta) - \langle G'(\hat{\eta}_\theta)\,,\, \eta - \hat{\eta}_\theta \rangle := D_G\Big(\eta \| \hat{\eta}_\theta\Big).$$

B. Derivation of risks for specific choices of scoring rules

In this section, we will derive specific equations for Total, Bayes, and Excess pointwise risks to get the estimates of total, aleatoric, and epistemic uncertainties correspondingly. We will omit subscript θ in this section, indicating an estimate by using a hat.

Recall equations for proper scoring rule and different risks:

$$\ell(\eta, i) = \langle G'(\eta)\,,\, \eta \rangle - G'_i(\eta) - G(\eta),$$
$$\mathrm{R}_{\text{Tot}} = \langle G'(\hat{\eta})\,,\, \hat{\eta} \rangle - G(\hat{\eta}) - \langle G'(\hat{\eta})\,,\, \eta \rangle,$$
$$\mathrm{R}_{\text{Bayes}} = -G(\eta),$$
$$\mathrm{R}_{\text{Exc}} = G(\eta) - G(\hat{\eta}) + \langle G'(\hat{\eta})\,,\, \hat{\eta} - \eta \rangle.$$

B.1. Log score (Cross-Entropy)

$$G(\eta) = \sum_{k=1}^{K} \eta_k \log \eta_k,$$
$$G'(\eta)_k = 1 + \log \eta_k,$$

$$\ell(\eta, i) = \langle 1 + \log \eta\,,\, \eta \rangle - 1 - \log \eta_i - \sum_{k=1}^{K} \eta_k \log \eta_k = \\ \sum_{k=1}^{K} \eta_k \log \eta_k + 1 - 1 - \log \eta_i - \sum_{k=1}^{K} \eta_k \log \eta_k = -\log \eta_i,$$

$$\mathrm{R}_{\text{Tot}} = \sum_{k=1}^{K}\Big((1+\log\hat{\eta}_k)\hat{\eta}_k - \hat{\eta}_k\log\hat{\eta}_k - (1+\log\hat{\eta}_k)\eta_k\Big) = \sum_{k=1}^{K}\Big(\hat{\eta}_k\log\hat{\eta}_k - \hat{\eta}_k\log\hat{\eta}_k - \eta_k\log\hat{\eta}_k\Big) = \mathrm{CE}\Big[\eta\|\hat{\eta}\Big],$$

$$\mathrm{R}_{\text{Bayes}} = -\sum_{k=1}^{K}\eta_k\log\eta_k = \mathbb{H}\eta,$$

$$\mathrm{R}_{\text{Exc}} = \mathrm{R}_{\text{Tot}} - \mathrm{R}_{\text{Bayes}} = \mathrm{CE}\Big[\eta\|\hat{\eta}\Big] - \mathbb{H}\eta = \mathrm{KL}\Big[\eta\|\hat{\eta}\Big].$$

B.2. Quadratic score (Brier Score)

$$G(\eta) = -\sum_{k=1}^{K}\eta_k(1-\eta_k),$$

$$G'(\eta)_k = 2\eta_k - 1,$$

$$\ell(\eta, i) = \langle 2\eta - 1\,,\,\eta\rangle - 2\eta_i + 1 + \sum_{k=1}^{K}\eta_k(1-\eta_k) = 2\sum_{k=1}^{K}\eta_k^2 - 1 - 2\eta_i + 1 + 1 - \sum_{k=1}^{K}\eta_k^2 = \sum_{k=1}^{K}\eta_k^2 - 2\eta_i + 1,$$

since constant does not affect optimization, we will use the following:

$$\ell(\eta, i) = \sum_{k=1}^{K}\eta_k^2 - 2\eta_i.$$

$$\mathrm{R}_{\text{Tot}} = \sum_{k=1}^{K}(2\hat{\eta}_k - 1)\hat{\eta}_k + \sum_{k=1}^{K}\hat{\eta}_k(1-\hat{\eta}_k) - \sum_{k=1}^{K}(2\hat{\eta}_k - 1)\eta_k = \sum_{k=1}^{K}(\hat{\eta}_k^2 - 2\hat{\eta}_k\eta_k + \eta_k^2 - \eta_k^2) + 1 = \|\hat{\eta} - \eta\|_2^2 - \|\eta\|_2^2 + 1,$$

$$\mathrm{R}_{\text{Bayes}} = \sum_{k=1}^{K}\eta_k(1-\eta_k) = 1 - \|\eta_k\|_2^2,$$

$$\mathrm{R}_{\text{Exc}} = \mathrm{R}_{\text{Tot}} - \mathrm{R}_{\text{Bayes}} = \|\hat{\eta} - \eta\|_2^2 - \|\eta\|_2^2 + 1 - 1 + \|\eta_k\|_2^2 = \|\hat{\eta} - \eta\|_2^2.$$

B.3. Zero-one score

$$G(\eta) = \max_k \eta_k - 1,$$

$$G'(\eta)_k = \mathbb{I}[k = \arg\max_j \eta_j],$$

$$\ell(\eta, i) = \langle \mathbb{I}[k = \arg\max_j \eta_j]\,,\,\eta\rangle - \mathbb{I}[i = \arg\max_j \eta_j] - \max_k \eta_k + 1 = \max_k \eta_k - \mathbb{I}[i = \arg\max_j \eta_j] - \max_k \eta_k + 1 = 1 - \mathbb{I}[i = \arg\max_j \eta_j],$$

$$\mathrm{R}_{\text{Tot}} = \sum_{k=1}^{K}\Big(\hat{\eta}_k \mathbb{I}[k = \arg\max_j \hat{\eta}_k] - \eta_k \mathbb{I}[k = \arg\max_j \hat{\eta}_k]\Big) - \max_k \hat{\eta}_k + 1 =$$

$$\max_k \hat{\eta}_k - \eta_{\arg\max_j \hat{\eta}_k} - \max_k \hat{\eta}_k + 1 = 1 - \eta_{\arg\max_j \hat{\eta}_k},$$

$$\mathrm{R}_{\text{Bayes}} = 1 - \max_k \eta_k,$$

$$\mathrm{R}_{\text{Exc}} = \mathrm{R}_{\text{Tot}} - \mathrm{R}_{\text{Bayes}} = 1 - \eta_{\arg\max_j \hat{\eta}_k} - 1 + \max_k \eta_k = \eta_{\arg\max_j \eta_k} - \eta_{\arg\max_j \hat{\eta}_k}.$$

B.4. Spherical score

$$G(\eta) = \|\eta\|_2 - 1,$$

$$G'(\eta)_k = \frac{\eta_k}{\|\eta\|_2},$$

$$\ell(\eta, i) = \langle \frac{\eta}{\|\eta\|_2}\,,\, \eta\rangle - \frac{\eta_i}{\|\eta\|_2} - \|\eta\|_2 + 1 = \|\eta\|_2 - \frac{\eta_i}{\|\eta\|_2} - \|\eta\|_2 + 1 = 1 - \frac{\eta_i}{\|\eta\|_2},$$

$$\mathrm{R}_{\text{Tot}} = \sum_{k=1}^{K}\Big(\frac{\hat{\eta}_k \hat{\eta}_k}{\|\hat{\eta}\|_2} - \frac{\eta_k \hat{\eta}_k}{\|\hat{\eta}\|_2}\Big) - \|\hat{\eta}\|_2 + 1 = 1 - \sum_{k=1}^{K}\frac{\eta_k \hat{\eta}_k}{\|\hat{\eta}\|_2} = 1 - \|\eta\|_2 \langle \frac{\eta}{\|\eta\|_2}\,,\, \frac{\hat{\eta}}{\|\hat{\eta}\|_2}\rangle,$$

$$\mathrm{R}_{\text{Bayes}} = 1 - \|\eta\|_2,$$

$$\mathrm{R}_{\text{Exc}} = \mathrm{R}_{\text{Tot}} - \mathrm{R}_{\text{Bayes}} = 1 - \|\eta\|_2 \langle \frac{\eta}{\|\eta\|_2}\,,\, \frac{\hat{\eta}}{\|\hat{\eta}\|_2}\rangle + \|\eta\|_2 - 1 = \|\eta\|_2\Big(1 - \langle \frac{\eta}{\|\eta\|_2}\,,\, \frac{\hat{\eta}}{\|\hat{\eta}\|_2}\rangle\Big).$$

B.5. Negative log score

$$G(\eta) = -\sum_{k=1}^{K} \log \eta_k,$$

$$G'(\eta)_k = -\frac{1}{\eta_k},$$

$$\ell(\eta, i) = \langle -\frac{1}{\eta}\,,\, \eta\rangle + \frac{1}{\eta_i} + \sum_{k=1}^{K} \log \eta_k = -K + \frac{1}{\eta_i} + \sum_{k=1}^{K} \log \eta_k,$$

since constant does not affect optimization, we will have:

$$\ell(\eta, i) = \frac{1}{\eta_k} + \sum_{k=1}^{K} \log \eta_k,$$

$$\mathrm{R}_{\text{Tot}} = \sum_{k=1}^{K}\Big(-\frac{\hat{\eta}_k}{\hat{\eta}_k} + \frac{\eta_k}{\hat{\eta}_k} + \log \hat{\eta}_k\Big) = \sum_{k=1}^{K}\Big(\frac{\eta_k}{\hat{\eta}_k} + \log \hat{\eta}_k - 1\Big),$$

$$\mathrm{R}_{\text{Bayes}} = \sum_{k=1}^{K} \log \eta_k,$$

$$\mathrm{R}_{\text{Exc}} = \mathrm{R}_{\text{Tot}} - \mathrm{R}_{\text{Bayes}} = \sum_{k=1}^{K}\Big(\frac{\eta_k}{\hat{\eta}_k} + \log \hat{\eta}_k - 1 - \log \eta_k\Big) = \sum_{k=1}^{K}\Big(\frac{\eta_k}{\hat{\eta}_k} - \log \frac{\eta_k}{\hat{\eta}_k} - 1\Big) = D_{\text{IS}}[\eta \| \hat{\eta}].$$

C. Approximations

In this section, we provide derivations for different types of Bayesian approximations. First, let us note, that Bayes risk depends only on the ground truth density (hence has only one argument), while Total and Excess risks are functions of both ground truth density and our approximation (they have two arguments).

Given a set of models (ensemble or samples from a posterior distribution over model parameters), we can use these approximation ideas for both of the arguments.

Total risk. Let us start with Total risk. Since R_{Tot} *depends linearly* on η, both approximations of the ground truth density lead to the same result:

$$\tilde{\mathrm{R}}_{\text{Tot}}(\theta \mid x) = \langle G'(\hat{\eta}_\theta)\,,\, \hat{\eta}_\theta\rangle - G(\hat{\eta}_\theta) - \langle G'(\hat{\eta}_\theta)\,,\, \hat{\eta}_{D_{tr}}\rangle, \tag{6}$$

where $\hat{\eta}_{D_{tr}}(x) = \mathbb{E}_{p(\tilde{\theta}|D_{tr})}\hat{\eta}_{\tilde{\theta}}(x)$ denotes posterior predictive distribution.

We also have a second argument, for which we also can incorporate this Bayesian reasoning. To do so, we will use an ensemble. Again, two approximation ideas, which lead to the following:

$$\tilde{\mathrm{R}}^{(1)}_{\text{Tot}}(x) = \mathbb{E}_{p(\theta|D_{tr})}\Big(\langle G'(\hat{\eta}_\theta)\,,\, \hat{\eta}_\theta\rangle - G(\hat{\eta}_\theta) - \langle G'(\hat{\eta}_\theta)\,,\, \hat{\eta}_{D_{tr}}\rangle\Big),$$

and

$$\tilde{\mathrm{R}}^{(2)}_{\text{Tot}}(x) = \langle G'(\hat{\eta}_{D_{tr}})\,,\, \hat{\eta}_{D_{tr}}\rangle - G(\hat{\eta}_{D_{tr}}) - \langle G'(\hat{\eta}_{D_{tr}})\,,\, \hat{\eta}_{D_{tr}}\rangle = -G(\hat{\eta}_{D_{tr}}),$$

where by superscripts we denote the index of approximation idea.

As we will see below, the second case effectively corresponds to the situation with zero Excess risk (no epistemic uncertainty).

Bayes risk. For Bayes risks, these two approximation ideas lead to the following results:

$$\tilde{\mathrm{R}}^{(1)}_{\text{Bayes}}(x) = -\mathbb{E}_{p(\tilde{\theta}|D_{tr})}G(\hat{\eta}_{\tilde{\theta}}) \quad \text{and} \quad \tilde{\mathrm{R}}^{(2)}_{\text{Bayes}}(x) = -G(\hat{\eta}_{D_{tr}}).$$

However, it is not clear, which of these approximations is better. To investigate it, we assume that there exists a vector of true parameters θ^*, for which $\eta(x) = p(y \mid x) = p(y \mid x, \theta^*)$ for any input point x. For example for neural networks, which are flexible enough to represent any function, this is a mild assumption.

Note, that according to equation (4), Bayes risk is concave. Using $g(.) = -\mathbb{E}_x G(.)$, the following follows from Jensen's inequality:

$$\mathbb{E}_{p(\theta|D_{tr})}g(\hat{\eta}_\theta) \leq g(\mathbb{E}_{p(\theta|D_{tr})}\hat{\eta}_\theta).$$

At the same time, we know that for Bayes risk the following must hold for any θ:

$$g(\eta) = g\big(p(y \mid x, \theta^*)\big) \leq g(\hat{\eta}_\theta),$$

and in particular:

$$g(\eta) \leq \mathbb{E}_{p(\theta|D_{tr})}g(\hat{\eta}_\theta).$$

Hence, we have the following:

$$g(\eta) \leq \mathbb{E}_{p(\theta|D_{tr})}g(\hat{\eta}_\theta) \leq g(\mathbb{E}_{p(\theta|D_{tr})}\hat{\eta}_\theta). \tag{7}$$

From equation (7) we see that for estimating Bayes risk, it is beneficial to use **the first** approximation idea, as it leads to a tighter upper-bound on a true risk, under the assumption that true parameter θ^* belongs to the set of possible parameters of the model.

Excess risk. For Excess risk these two approximation ideas for the ground truth distribution lead to the following results:

$$\tilde{\mathrm{R}}^{(1)}_{\text{Exc}}(\theta \mid x) = \mathbb{E}_{p(\tilde{\theta}|D_{tr})}D_G\Big(\hat{\eta}_{\tilde{\theta}} \parallel \hat{\eta}_\theta\Big) \quad \text{and} \quad \tilde{\mathrm{R}}^{(2)}_{\text{Exc}}(\theta \mid x) = D_G\Big(\hat{\eta}_{D_{tr}} \parallel \hat{\eta}_\theta\Big).$$

Furthermore, using results of (Pfau, 2013), we can rewrite the first approximation as follows:

$$\tilde{\mathrm{R}}^{(1)}_{\mathrm{Exc}}(\theta \mid x) = D_G\Big(\hat{\eta}_{D_{tr}} \parallel \hat{\eta}_\theta\Big) + \mathbb{E}_{p(\tilde{\theta}|D_{tr})} D_G\Big(\hat{\eta}_{\tilde{\theta}} \parallel \hat{\eta}_{D_{tr}}\Big) = \tilde{\mathrm{R}}^{(2)}_{\mathrm{Exc}}(\theta \mid x) + \mathbb{E}_{p(\tilde{\theta}|D_{tr})} D_G\Big(\hat{\eta}_{\tilde{\theta}} \parallel \hat{\eta}_{D_{tr}}\Big).$$

We see that the first approximation contains the second one, and as Bregman divergence is non-negative, $\tilde{\mathrm{R}}^{(1)}_{\mathrm{Exc}}(\theta \mid x) \geq \tilde{\mathrm{R}}^{(2)}_{\mathrm{Exc}}(\theta \mid x)$ (which also follows from the convexity of Bregman in its first argument). Moreover, the difference between estimates is equal to $\mathbb{E}_{p(\tilde{\theta}|D_{tr})} D_G\Big(\hat{\eta}_{\tilde{\theta}} \| \hat{\eta}_{D_{tr}}\Big)$, which is known (Banerjee et al., 2005) as a **Bregman Information**. A specific case of Bregman information for Log score (see Table 1) is BALD (Gal et al., 2017; Houlsby et al., 2011), a well-known epistemic uncertainty measure. We see, that *BALD is only a part of the epistemic uncertainty estimate*. Worth noting, that in recent work (Wimmer et al., 2023; Schweighofer et al., 2023) concerns regarding this expression as a proper measure of epistemic uncertainty were raised.

Furthermore, similarly to the Total risk estimation, we can use ensembles to incorporate the Bayesian approximation idea for the second argument. Thus, we end up with four different estimates of Excess risk:

- **Expected Pairwise Bregman Divergence (EPBD):**

  $$\tilde{\mathrm{R}}^{(1,1)}_{\mathrm{Exc}}(x) = \mathbb{E}_{p(\tilde{\theta}|D_{tr})} \mathbb{E}_{p(\theta|D_{tr})} D_G\Big(\hat{\eta}_{\tilde{\theta}} \parallel \hat{\eta}_\theta\Big).$$

  Note, that since KL divergence is a special case of Bregman divergence, Expected Pairwise KL (EPKL (Malinin & Gales, 2021; Schweighofer et al., 2023)) is one of the special cases of this Excess risk estimate.

- **Bregman Information (BI):**

  $$\tilde{\mathrm{R}}^{(1,2)}_{\mathrm{Exc}}(x) = \mathbb{E}_{p(\tilde{\theta}|D_{tr})} D_G\Big(\hat{\eta}_{\tilde{\theta}} \parallel \hat{\eta}_{D_{tr}}\Big),$$

  which special case is BALD (Gal et al., 2017; Houlsby et al., 2011).

- **Reverse Bregman Information (RBI):**

  $$\tilde{\mathrm{R}}^{(2,1)}_{\mathrm{Exc}}(x) = \mathbb{E}_{p(\theta|D_{tr})} D_G\Big(\hat{\eta}_{D_{tr}} \parallel \hat{\eta}_\theta\Big).$$

  Its special case for Log score is known as **Reverse Mutual Information** (Malinin & Gales, 2021).

- Finally, we obtain

  $$\tilde{\mathrm{R}}^{(2,2)}_{\mathrm{Exc}}(x) = D_G\Big(\hat{\eta}_{D_{tr}} \parallel \hat{\eta}_{D_{tr}}\Big) = 0,$$

  which is coherent with the result obtained for the Total risk, when Excess risk (epistemic uncertainty) is equal to 0.

However, it is not clear, what estimate of the Excess risk we should use. Indeed, neither of these estimates are upper nor lower bounds for the true Excess risk. This is because contrary to Bayes risk, we don't have any idea if Excess risk with ground truth η reaches any extreme. For another explanation, see Figure 3 and discussion in Section 7.

In Figure 3, for simplicity, we consider only the Bayesian approximation of the first argument (ground-truth probability). In black, we have actual (real risks), while in color we have different estimates of risks. Also, as two-sided arrows, we show the Excess risk.

If we *underestimate* the Total risk (see top plot in Figure 3), the best choice for Excess risk will be $\tilde{\mathrm{R}}^{(1)}_{\mathrm{Exc}}$, as despite being a lower bound on Excess risk, it is the best we can do ($\tilde{\mathrm{R}}^{(2)}_{\mathrm{Exc}}$ in this case will be even worse). However, if we *overestimate* Total risk, then there is no single best choice. In the bottom plot, when $\tilde{\mathrm{R}}_{\mathrm{Tot}}$ significantly overestimates $\mathrm{R}_{\mathrm{Tot}}$, the second idea for estimating Excess risk gives a better estimate, despite the first idea for Bayes risk still better.

Hence, the best estimate of Excess risk depends on how well we estimate Total risk. But we never know in advance for a particular input x, in which regime (overestimated or underestimated Total risk) we are. Thus, there is no best choice among these risks to approximate epistemic uncertainty.

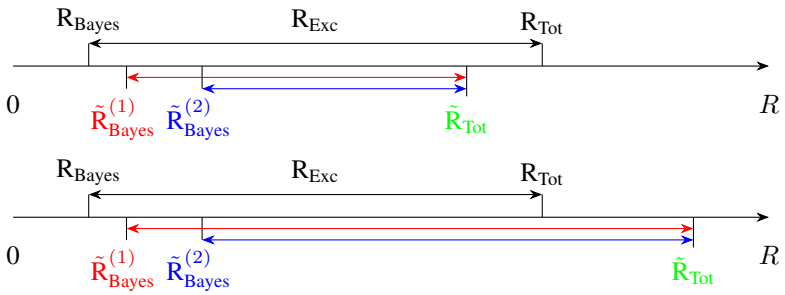

Figure 3: Different situations for risk estimates. Risks typed in black and above the axis are the true ones. Risks, typed in color, and below are estimates. Two-pointed arrows show Excess risks.

Top. $\tilde{\mathrm{R}}_{\mathrm{Tot}}$ underestimates $\mathrm{R}_{\mathrm{Tot}}$, $\tilde{\mathrm{R}}^{(1)}_{\mathrm{Bayes}}$ better estimates $\mathrm{R}_{\mathrm{Bayes}}$, and $\tilde{\mathrm{R}}^{(1)}_{\mathrm{Exc}}$ better estimates $\mathrm{R}_{\mathrm{Exc}}$.

Bottom. $\tilde{\mathrm{R}}_{\mathrm{Tot}}$ overestimates $\mathrm{R}_{\mathrm{Tot}}$, $\tilde{\mathrm{R}}^{(1)}_{\mathrm{Bayes}}$ better estimates $\mathrm{R}_{\mathrm{Bayes}}$, and $\tilde{\mathrm{R}}^{(2)}_{\mathrm{Exc}}$ better estimates $\mathrm{R}_{\mathrm{Exc}}$. We see, that for different estimates of $\mathrm{R}_{\mathrm{Tot}}$, we have different best approximations for $\mathrm{R}_{\mathrm{Exc}}$. See discussion in Section 7.

D. Relations between the estimates

In this section, we discuss how the measures of uncertainty are connected. In the main text, we discussed several ways how one can estimate risk given an ensemble of models, posterior, or samples from it. In what follows, we show how one can further decompose these estimates of Excess risk.

Let us start with $\tilde{\mathrm{R}}^{(1,1)}_{\mathrm{Exc}}(x)$. Using results of (Pfau, 2013), we have:

$$\tilde{\mathrm{R}}^{(1,1)}_{\mathrm{Exc}}(x) = \mathbb{E}_{p(\theta|D_{tr})}\mathbb{E}_{p(\tilde{\theta}|D_{tr})} D_G\left(\hat{\eta}_{\tilde{\theta}} \| \hat{\eta}_{\theta}\right) =$$
$$\mathbb{E}_{p(\theta|D_{tr})} D_G\left(\hat{\eta}_{D_{tr}} \| \hat{\eta}_{\theta}\right) + \mathbb{E}_{p(\tilde{\theta}|D_{tr})} D_G\left(\hat{\eta}_{\tilde{\theta}} \| \hat{\eta}_{D_{tr}}\right) = \tilde{\mathrm{R}}^{(2,1)}_{\mathrm{Exc}}(x) + \tilde{\mathrm{R}}^{(1,2)}_{\mathrm{Exc}}(x).$$

Since all of these estimates are non-negative, the following holds true:

$$\tilde{\mathrm{R}}^{(1,1)}_{\mathrm{Exc}}(x) \geq \tilde{\mathrm{R}}^{(2,1)}_{\mathrm{Exc}}(x) \geq \tilde{\mathrm{R}}^{(2,2)}_{\mathrm{Exc}}(x) = 0,$$

and

$$\tilde{\mathrm{R}}^{(1,1)}_{\mathrm{Exc}}(x) \geq \tilde{\mathrm{R}}^{(1,2)}_{\mathrm{Exc}}(x) \geq \tilde{\mathrm{R}}^{(2,2)}_{\mathrm{Exc}}(x) = 0$$

for any x.

Moreover, one can show that the following holds:

$$\tilde{\mathrm{R}}^{(1)}_{\mathrm{Tot}}(x) = \tilde{\mathrm{R}}^{(1)}_{\mathrm{Bayes}}(x) + \tilde{\mathrm{R}}^{(1,1)}_{\mathrm{Exc}}(x) = \tilde{\mathrm{R}}^{(2)}_{\mathrm{Bayes}}(x) + \tilde{\mathrm{R}}^{(2,1)}_{\mathrm{Exc}}(x),$$

and

$$\tilde{\mathrm{R}}^{(2)}_{\mathrm{Tot}}(x) = \tilde{\mathrm{R}}^{(2)}_{\mathrm{Bayes}}(x) + \tilde{\mathrm{R}}^{(2,2)}_{\mathrm{Exc}}(x) = \tilde{\mathrm{R}}^{(1)}_{\mathrm{Bayes}}(x) + \tilde{\mathrm{R}}^{(1,2)}_{\mathrm{Exc}}(x).$$

Additionally, Bregman information can be received as follows:

$$BI(x) = \tilde{\mathrm{R}}^{(1,1)}_{\mathrm{Exc}}(x) - \tilde{\mathrm{R}}^{(2,1)}_{\mathrm{Exc}}(x) = \tilde{\mathrm{R}}^{(1,2)}_{\mathrm{Exc}}(x) - \tilde{\mathrm{R}}^{(2,2)}_{\mathrm{Exc}}(x) = \tilde{\mathrm{R}}^{(2)}_{\mathrm{Bayes}}(x) - \tilde{\mathrm{R}}^{(1)}_{\mathrm{Bayes}}(x).$$

Reverse Bregman Information:

$$RBI(x) = \tilde{\mathrm{R}}^{(2,1)}_{\mathrm{Exc}}(x) - \tilde{\mathrm{R}}^{(2,2)}_{\mathrm{Exc}}(x) = \tilde{\mathrm{R}}^{(1,1)}_{\mathrm{Exc}}(x) - \tilde{\mathrm{R}}^{(1,2)}_{\mathrm{Exc}}(x) = \tilde{\mathrm{R}}^{(1)}_{\mathrm{Tot}}(x) - \tilde{\mathrm{R}}^{(2)}_{\mathrm{Tot}}(x).$$

E. Hand-crafted datasets

In this section, we describe the noisy versions of CIFAR10 and CIFAR100 datasets created for our experiments. We created three noisy datasets, each discussed below.

E.1. Noisy labels datasets

In the dataset, the images are the same as in the original dataset (covariates are not changed). However, some of the labels are randomly swapped. Hence, only labels were changed, while covariates were kept as in the original dataset. The motivation for the creation of this dataset is due to the fact that conventional image classification datasets essentially contain no aleatoric uncertainty (Kapoor et al., 2022). To mitigate the limitation, which is critical for our evaluation, we introduce the label noise manually. By nature, this noise is aleatoric.

CIFAR10. We decide to do the following pairs of labels that are randomly swapped: 1 to 7, 7 to 1, 3 to 8, 8 to 3, 2 to 5, and 5 to 2.

CIFAR100. We decided to randomly swap the following pairs of labels: 1 to 7, 7 to 1, 3 to 8, 8 to 3, 2 to 5, 5 to 2, 10 to 20, 20 to 10, 40 to 50, 50 to 40, 90 to 99, 99 to 90, 25 to 75, 75 to 25, 17 to 71, 71 to 17, 13 to 31, 31 to 13, and 24 to 42, 42 to 24.

E.2. Blurred datasets

When predicting by a model trained to particular data, on some another dataset, where covariates are drastically different ("hard-OOD"), predictions are not meaningful anymore.

To introduce "soft-OOD" datasets, where labels remain the same but covariates are altered (yet the predicted vectors are still reasonable), we applied a Gaussian blur to the images in CIFAR10 and CIFAR100. Specifically, before the same transformations that we did for standard validation splits, we did Gaussian blur with kernel size of $(3, 3)$ and sigma of $(0.1, 2.0)$. These blurred datasets help model "soft-OOD" scenarios.

E.3. Missing class dataset

Essentially, Excess risk-based uncertainty measures evaluate disagreement between models in an ensemble. It is interesting to study this disagreement in the case when some of the classes are missing for some members within the ensemble. For this, we consider a special version of the CIFAR10 dataset, when we totally ignore some classes for some members of the ensemble. It results in an increase disagreement between ensemble members.

In our experiments, we had overall 20 ensemble members. For members 1 and 2 class 0 was missed. For members 3 and 4 class 1 was missed. And for members 5 and 6 class 2 was missed. All other members of the ensemble had all the classes.

F. Training details

Training procedures for each dataset were similar. We used either ResNet18 or VGG19 architectures.

For CIFAR10-based datasets, we used code from this repository: `https://github.com/kuangliu/pytorch-cifar`. The training procedure consisted of 200 epochs with a cosine annealing learning rate. For an optimizer, we use SGD with momentum and weight decay. For more details see the code.

In Figure 4 we present performance summary statistics of the ensembles of different architectures. Specifically, we show accuracy, macro averaged precision, recall, and F1-score. In the left figure, the distribution for ResNet18, trained with different loss functions. In the right figure, VGG19.

For CIFAR100-based datasets, we used code from this repository: `https://github.com/weiaicunzai/pytorch-cifar100`. The training procedure consisted of 200 epochs with learning rate decay at particular milestones: [60, 120, 160]. For an optimizer, we use SGD with momentum and weight decay. For more details see the code.

Similarly to CIFAR10, in Figure 5 we present performance summary statistics of the ensembles of different architectures. In the left figure, the distribution for ResNet18, trained with different loss functions. In the right figure, VGG19.

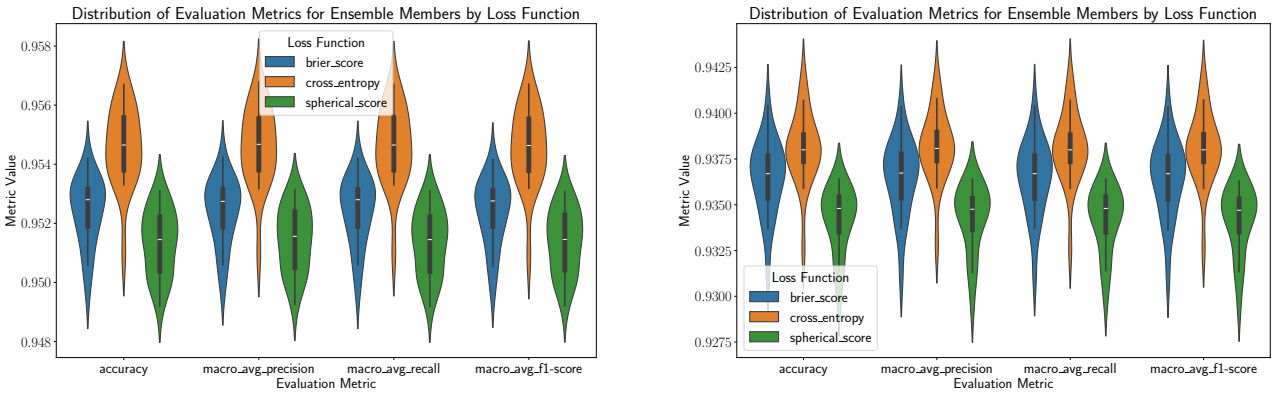

Figure 4: Violin plots for different training loss functions and different metrics. The training dataset is **CIFAR10**. Left: ResNet18. Right: VGG-19.

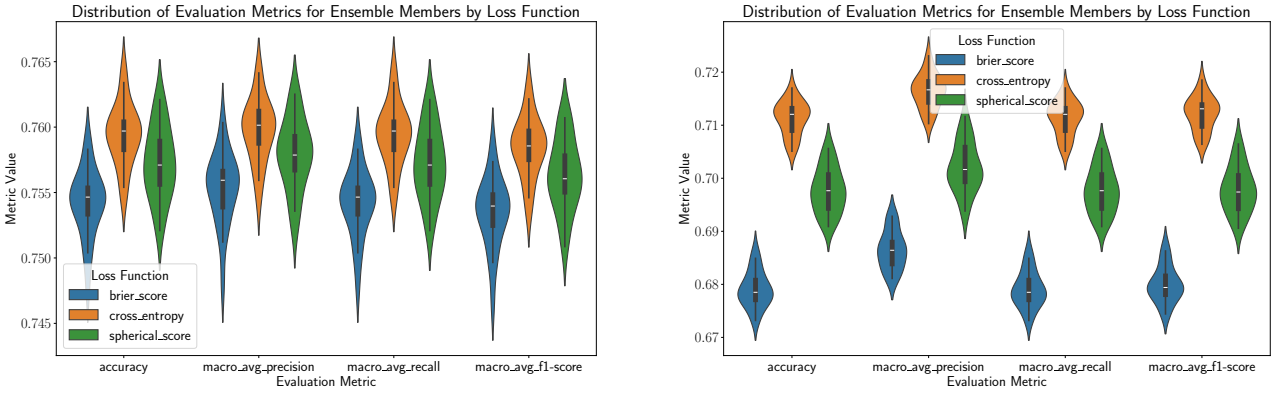

Figure 5: Violin plots for different training loss functions and different metrics. The training dataset is **CIFAR100**. Left: ResNet18. Right: VGG-19.

G. Additional experiments on matching / not-matching proper scoring rules

In this section, we provide additional results on the matching and non-matching proper scoring rules used for loss function and risk (uncertainty measures) computations.

Figure 6 displays a histogram of AUROC scores for out-of-distribution detection, comparing the matching and non-matching cases. For the matching case, the results are meaningful, showing a significant margin from an AUROC score of 0.5, which corresponds to a constant predictor. In contrast, the non-matching case includes values less than 0.5, indicating a drawback in combining different loss functions and uncertainty measures.

From Figure 2 in the main text, we see that the most unstable scoring rule is Neglog. One might think this instability is due to its computational issues, and excluding it might equalize the results for matching and non-matching cases. To test this, we plot an additional figure (Figure 7) excluding the Neglog scoring rule.

However, even after excluding Neglog, we still observe noise in the results, particularly from combinations of the Spherical loss function with other scoring rules for risk computations. This demonstrates that the effect of matching and non-matching scoring rules persists beyond the instability of any single rule.

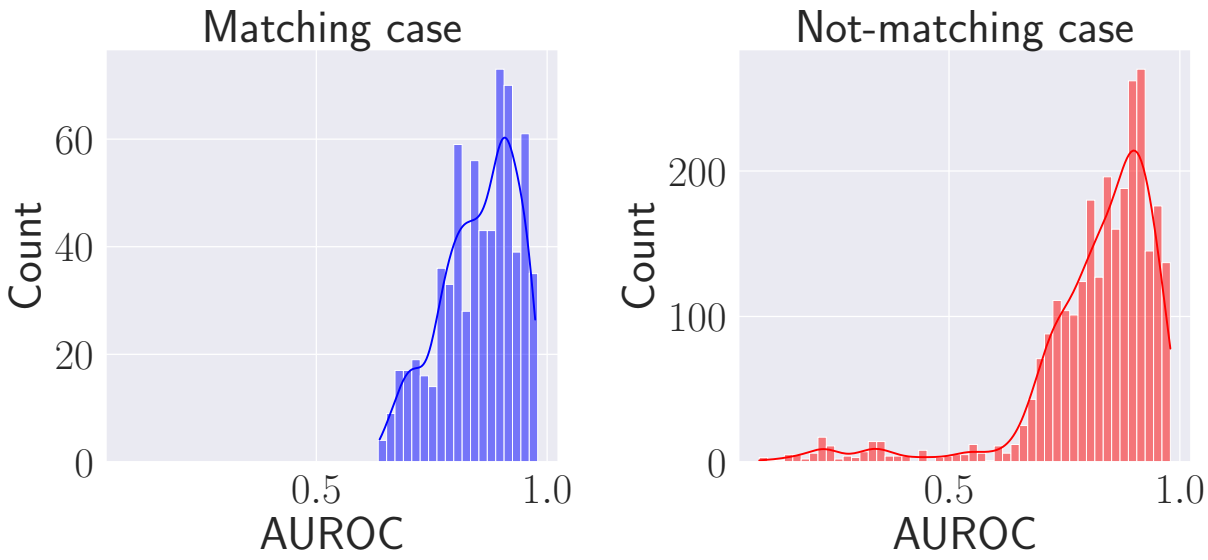

Figure 6: AUROC for out-of-distribution detection. Left: When scoring rules, used for loss function and risks computation, match. Right: When they are not matching. All proper scoring rules are considered.

H. Additional experiments on out-of-distribution detection

In this section, we provide an additional comparison of matching and not-matching results for out-of-distribution detection. For this, we add Table 4 and compare columns "Matching" and "Not-Matching" in this table element-wise. Column "Matching" corresponds to matching proper scoring rule used for loss function and uncertainty measures. Column "Not-Matching" corresponds to different ones. We see, that results in the left part of Table 4, except two values (in red), are systematically better, than in its right part. However, those two values are very close to each other.

I. Additional experiments misclassification detection

Similarly to the previous section with out-of-distribution detection, we compare columns "Matching" and "Not-Matching" in Table 5 element-wise. Column "Matching" corresponds to matching proper scoring rule used for loss function and uncertainty measures. Column "Not-Matching" corresponds to different ones. We see, that results in the left part of Table 5, are systematically better, than in its right part.

J. Other approximations

In this section, we discuss choices of Bayesian approximation strategy (Inner or Outer) separately for different problems. Outer in this notation means the first approach of the Bayesian approximation, when we perform Bayesian averaging of risk, and effectively apply an outer expectation. Inner corresponds to the second approach of Bayesian approximation, namely Bayesian model averaging when we first find the expectation and then plug it in the risk. Effectively we do inner expectation in this case.

J.1. Out-of-distribution detection

Here, we extend the results presented in the main part, and demonstrate, how the approximation strategy of risks influences the results of out-of-distribution detection. We present results in Table 6, where matching scoring rules were used for loss function and risks computations, and Table 7, where they are different.

We see, that results in Table 6 (matching scoring rules) are systematically better, than results in Table 7, when there was no

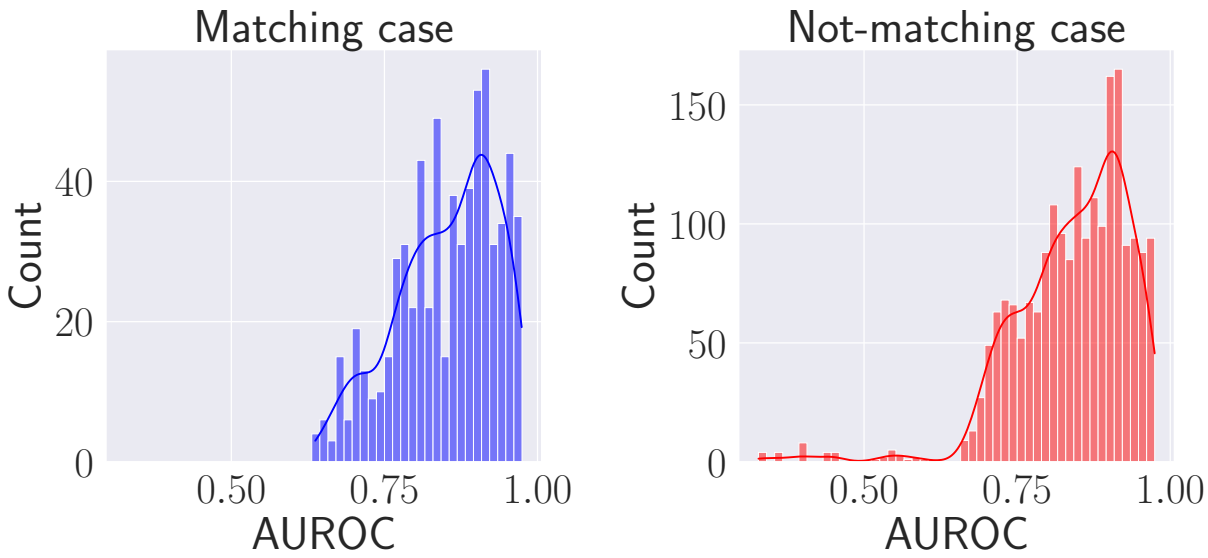

Figure 7: AUROC for out-of-distribution detection. Left: When scoring rules, used for loss function and risks computation, match. Right: When they are not matching. **Neglog proper scoring rule excluded from consideration.**

matching.

Bayes risk. We see, that there is no the only leader of specific approximation. Nevertheless, in most cases and on average, Inner approximations seem slightly better. Average values for matching case: Bayes(O) - 85.47 ± 8.12, Bayes(I) - 86.54 ± 7.69. Average values for not matching case: Bayes(O) - 85.09 ± 8.32, Bayes(I) - 86.41 ± 7.72. In both cases, the Inner approximation is slightly (on average) better, than the Outer.

Total risk. Similarly to Bayes risk, we see that measures are very close to each other. Average values for matching case: Total(O) - 87.20 ± 7.17, Total(I) - 86.54 ± 7.69. Average values for not matching case: Total(O) - 85.26 ± 8.37, Total(I) - 86.41 ± 7.72.

Excess risk. In these Tables we have three out of four different approximations of Excess risk: Bregman Information (BI), Reverse Bregman Information (RBI), and Expected Pairwise Bregman Divergence (EPBD). We do not consider Inner Inner approximation, as it is zero. As in the previous cases, there is no sole leader in these measures.

J.2. Misclassification detection

Similar to the previous section, here we demonstrate extended results of misclassification detection. We present results in Table 8, where matching scoring rules were used for loss function and risks computations, and Table 9, where they are different.

We see, that results in Table 8 (matching scoring rules) are systematically better than results in Table 9, when there was no matching.

Bayes risk. As before, there is no sole leader. Average values for matching case: Bayes(O) - 87.62 ± 6.48, Bayes(I) - 87.72 ± 5.99. Average values for not matching case: Bayes(O) - 86.71 ± 7.23, Bayes(I) - 87.06 ± 6.40.

Total risk. Similarly to Bayes risk, we see that measures are very close to each other. Average values for matching case: Total(O) - 86.75 ± 5.70, Total(I) - 87.72 ± 5.99. Average values for not matching case: Total(O) - 82.54 ± 7.97, Total(I) - 87.06 ± 6.40.

Dataset		Matching			Not-Matching		
		Bayes	Excess	Total	Bayes	Excess	Total
InD	OOD						
CIFAR10	Blurred CIFAR10(*)	**84.38**	**87.80**	85.90	83.91	86.59	**86.00**
	Blurred CIFAR100	**94.48**	**95.65**	**95.42**	94.25	94.95	95.34
	CIFAR100	**91.05**	**90.08**	**90.87**	90.96	88.21	90.77
	SVHN	**94.54**	**93.31**	**94.44**	94.46	91.90	94.30
CIFAR100	Blurred CIFAR10	**87.96**	**85.36**	**90.32**	87.65	76.68	87.96
	Blurred CIFAR100(*)	**71.40**	**77.55**	**74.15**	70.99	72.83	73.55
	CIFAR10	**79.35**	**72.70**	**79.27**	78.87	64.19	76.64
	SVHN	84.90	**73.99**	**84.59**	**84.91**	65.46	82.12

Table 4: AUROC for OOD detection. Left (Matching): For computation of OOD measures, we used loss and corresponding risks, generated by the **matching function** G. Right (Not-matching): Loss and corresponding risks generated by **not-matching functions** G. The best results (element-wise, between Matching and Not-matching) are in **bold**. In red are those where Different is better. By asterisk (*) we denote "soft-OOD".

Dataset	Matching			Not-Matching		
	Bayes	Excess	Total	Bayes	Excess	Total
CIFAR10	**94.53**	**94.65**	**94.78**	94.30	92.72	94.39
CIFAR100	**86.47**	**82.90**	**86.83**	85.57	71.62	83.13
Missed class CIFAR10	**93.73**	**83.23**	**91.14**	93.55	79.71	89.75
Noisy CIFAR10	**81.00**	**74.35**	**80.97**	79.49	70.17	78.52
Noisy CIFAR100	**82.64**	**72.30**	**82.45**	81.49	60.23	78.19

Table 5: AUROC for misclassification detection. Left (Matching): For computation of OOD measures, we used loss and corresponding risks, generated by the **matching function** G. Right (Not-Matching): Loss and corresponding risks generated by **not-matching functions** G. The best results (element-wise, between Matching and Not-Matching) are in **bold**.

Excess risk. In these Tables we have three out of four different approximations of Excess risk: Bregman Information (BI), Reverse Bregman Information (RBI), and Expected Pairwise Bregman Divergence (EPBD). We do not consider Inner Inner approximation, as it is zero. As in the previous cases, there is no sole leader in these measures.

Dataset		Metrics						
InD	OOD	Bayes(O)	Bayes(I)	Total(O)	Total(I)	BI	RBI	EPBD
CIFAR10	Blurred CIFAR10(*)	83.26	85.49	86.30	85.49	87.63	**87.93**	87.84
	Blurred CIFAR100	93.75	95.22	95.62	95.22	95.60	**95.68**	95.66
	CIFAR100	**91.11**	90.99	90.75	90.99	90.20	89.96	90.07
	SVHN	**94.55**	94.54	94.34	94.54	93.39	93.23	93.31
CIFAR100	Blurred CIFAR10	86.74	89.18	**91.45**	89.18	85.83	84.48	85.76
	Blurred CIFAR100(*)	70.18	72.62	75.68	72.62	76.76	**78.03**	77.87
	CIFAR10	79.25	**79.44**	79.10	**79.44**	73.49	71.82	72.78
	SVHN	**84.95**	84.85	84.33	84.85	75.07	72.80	74.11

Table 6: AUROC for out-of-distribution detection, when **matching** scoring rules were used for loss/risks. Here we extend different approximations. Notation: O - Outer, I - Inner. (R)BI - (Reverse) Bregman Information, EPBD - Expected Pairwise Bregman Divergence. Best results in **bold**, second-best underline. By asterisk (*) we denote "soft-OOD".

Dataset		Metrics						
InD	OOD	Bayes(O)	Bayes(I)	Total(O)	Total(I)	BI	RBI	EPBD
CIFAR10	Blurred CIFAR10(*)	82.56	85.26	**86.74**	85.26	86.57	86.46	86.73
	Blurred CIFAR100	93.45	95.05	**95.64**	95.05	95.00	94.71	95.12
	CIFAR100	90.92	**90.99**	90.55	**90.99**	88.34	88.04	88.25
	SVHN	94.41	**94.51**	94.09	**94.51**	91.94	91.75	92.00
CIFAR100	Blurred CIFAR10	86.36	**88.94**	86.98	**88.94**	76.98	75.04	78.02
	Blurred CIFAR100(*)	69.42	72.56	**74.54**	72.56	71.87	72.99	73.63
	CIFAR10	78.71	**79.03**	74.26	**79.03**	64.67	63.30	64.61
	SVHN	84.87	**84.95**	79.29	**84.95**	66.20	64.34	65.84

Table 7: AUROC for out-of-distribution detection, when **not-matching** scoring rules were used for loss/risks. Notation: O - Outer, I - Inner. (R)BI - (Reverse) Bregman Information, EPBD - Expected Pairwise Bregman Divergence. Best results in **bold**, second-best underline. By asterisk (*) we denote "soft-OOD".

Dataset	Metrics						
InD	Bayes(O)	Bayes(I)	Total(O)	Total(I)	BI	RBI	EPBD
CIFAR10	94.30	94.76	**94.79**	94.76	94.68	94.60	94.65
CIFAR100	86.12	86.81	**86.86**	86.81	83.74	81.92	83.04
Missed class CIFAR10	**94.47**	92.99	89.30	92.99	86.34	80.86	82.49
Noisy CIFAR10	80.57	**81.42**	80.53	**81.42**	74.77	73.98	74.31
Noisy CIFAR100	**82.66**	82.62	82.28	82.62	73.28	71.25	72.37

Table 8: AUROC for misclassification detection, when **matching** scoring rules were used for loss/risks. Notation: O - Outer, I - Inner. (R)BI - (Reverse) Bregman Information, EPBD - Expected Pairwise Bregman Divergence. Best results in **bold**, second-best underline.

Dataset	Metrics						
InD	Bayes(O)	Bayes(I)	Total(O)	Total(I)	BI	RBI	EPBD
CIFAR10	94.11	**94.50**	<u>94.29</u>	**94.50**	92.53	92.59	93.04
CIFAR100	<u>85.19</u>	**85.96**	80.30	**85.96**	72.45	70.00	72.40
Missed class CIFAR10	**94.28**	<u>92.83</u>	86.67	<u>92.83</u>	81.24	77.62	80.26
Noisy CIFAR10	<u>78.52</u>	**80.45**	76.59	**80.45**	71.02	69.18	70.33
Noisy CIFAR100	<u>81.44</u>	**81.54**	74.85	**81.54**	61.37	58.88	60.45

Table 9: AUROC for misclassification detection, when **not-matching** scoring rules were used for loss/risks. Notation: O - Outer, I - Inner. (R)BI - (Reverse) Bregman Information, EPBD - Expected Pairwise Bregman Divergence. Best results in **bold**, second-best <u>underline</u>.