Uncertainty Quantification for Regression Using Proper Scoring Rules

Anonymous authors
Paper under double-blind review

Abstract

Quantifying predictive uncertainty is essential to enable reliable decision-making with machine learning models. Recent work proposed a unified framework for uncertainty quantification (UQ) based on proper scoring rules, but has focused exclusively on classification. Extending these results to regression remains an open problem. In this paper, we introduce a UQ framework for regression that applies to any proper scoring rule, including the CRPS, logarithmic, squared-error, and quadratic scores. We derive closed-form expressions under practical parametric assumptions and show how to estimate them using an ensemble of predictors. The framework allows separating and estimating aleatoric from epistemic types of predictive uncertainty, and it recovers popular, well-established variance- and entropy-based UQ measures as special instances. Finally, a broad evaluation on synthetic and real-world datasets provides guidance for selecting reliable UQ measures in practice.

1 Introduction

Machine learning models are widely deployed, yet their predictions can be unreliable. Recognizing when this occurs is essential in safety-critical applications (Helou et al., 2020; Su et al., 2023). Uncertainty quantification (UQ) addresses precisely this need (Hüllermeier & Waegeman, 2021; Gal & Ghahramani, 2016).

One of UQ's goals is to separate two types of predictive uncertainty. Aleatoric uncertainty (AU) arises from irreducible nondeterminism in the data, while epistemic uncertainty (EU) arises from limited knowledge about the underlying process (Hüllermeier & Waegeman, 2021). The distinction between them is crucial for decision-making in downstream tasks (Kendall & Gal, 2017; Kotelevskii et al., 2024).

A natural idea to formalize UQ measures is to link uncertainty to the expected error at a given point. Early works explored this connection for specific losses and model families (Kotelevskii et al., 2022; Lahlou et al., 2023). Later works generalized the approach using proper scoring rules (Gneiting & Raftery, 2007), deriving uncertainty measures via Bayesian approximations to pointwise risk components. This led to a unified framework that encompasses many known UQ measures for classification (Kotelevskii et al., 2025; Schweighofer et al., 2025; Hofman et al., 2026). However, many real-world applications involve continuous outcomes (Gal & Ghahramani, 2016; Amini et al., 2020; Valdenegro-Toro & Mori, 2022), and this framework has not yet been extended to regression.

In particular, we derive a family of uncertainty measures for regression that decompose into aleatoric and epistemic components and provide closed-form approximations for popular scoring rules under standard parametric assumptions. Among these scoring rules, we consider CRPS, widely used in meteorology (Alet et al., 2025; Lang et al., 2025), but also more traditional scores such as the logarithmic or quadratic score. Furthermore, our framework recovers established entropy- and variance-based measures (Depeweg et al., 2018; Bülte et al., 2025), providing a principled understanding of their relationships.

Our contribution is to turn the classical proper score risk decomposition into a constructive framework for regression uncertainty quantification. Unlike classification, regression requires an explicit predictive family, and this modeling choice determines which uncertainty measures can be computed and interpreted. We

instantiate the framework for Gaussian ensemble predictors while keeping the construction applicable to other predictive families.

Specifically, our contributions are:

- A proper-scoring-rule framework for regression UQ. We extend risk-based uncertainty quantification from classification to regression and derive total, aleatoric, and epistemic uncertainty measures from proper scoring rules.
- Practical estimators for common scoring rules. Under Gaussian ensemble predictors, we derive closed-form or tractable estimators for CRPS, logarithmic, quadratic, and squared-error scores. These estimators recover established entropy- and variance-based measures as special cases, while also yielding less standard CRPS- and quadratic-score uncertainty measures.
- An empirical study of which risk component is useful for which task. We evaluate the resulting measures on synthetic and real-world benchmarks for selective prediction, out-of-distribution detection, and active learning. Rather than seeking a universally dominant uncertainty score, our experiments identify which risk components are most appropriate for different downstream uses.

2 Background on Proper Scoring Rules

We start by defining proper scoring rules and introducing the necessary notation, following (Gneiting & Raftery, 2007; Waghmare & Ziegel, 2026).

We consider a supervised regression problem, with input $x \in \mathcal{X}$ and output $y \in \mathcal{Y}$. Let $\mathcal{P}$ be a convex class of probability measures on $\mathcal{Y}$. A predictive distribution is any probability measure $\widehat{P} \in \mathcal{P}$. In the classification case, predictive distributions appear naturally, as models usually output a categorical distribution $\widehat{P} = \widehat{P}(y \mid x)$ over class labels. In regression, however, the standard approach is to produce only a point prediction $\widehat{y} = f(x) \in \mathcal{Y}$, with no distributional model of $y \mid x$ and hence no notion of predictive uncertainty. Nevertheless, to introduce an explicit representation of uncertainty, one can choose a predictive distribution over outcomes $\widehat{P}(y \mid x)$, with the Gaussian being the simplest and most widely considered option.

Once a predictive distribution is chosen, we need a way to evaluate its faithfulness of representing the actual distribution of observed outcomes. Scoring rules provide a principled way to do so. A scoring rule is an extended real-valued function $S\colon \mathcal{P} \times \mathcal{Y} \to \bar{\mathbb{R}} = [-\infty, +\infty]$ such that $S(\widehat{P}, \cdot)$ is quasi-integrable for all $\widehat{P} \in \mathcal{P}$. If the model predicts a distribution $\widehat{P}$ and an outcome y materializes, the forecaster's penalty is $S(\widehat{P}, y)$. Intuitively, a scoring rule assigns a value depending on the compatibility between the distribution and the outcome.

Usually, we are interested in assessing the predictive quality of the forecast $\widehat{P}$ on average over possible values of y. For this, let us consider the expected score under P:

$$S(\widehat{P}, P) = \int S(\widehat{P}, y)\, dP(y) \tag{1}$$

A practically important family of scoring rules are *proper scoring rules*, which incentivize learning the true data distribution. A scoring rule S is called proper relative to $\mathcal{P}$ if

$$S(P, P) \leq S(\widehat{P}, P) \quad \text{for all} \quad \widehat{P}, P \in \mathcal{P}. \tag{2}$$

The scoring rule is called *strictly* proper if the equality in Equation (2) holds iff $\widehat{P} = P$. Thus, any (strictly) proper scoring rule can be used to evaluate predicted distributions, ensuring that its minimum is attained at the true data distribution.

Let us introduce the following definitions, useful for our framework.

Definition 2.1. The *entropy function* is the expected score of a measure against itself:

$$H(P) = S(P, P) = \int S(P, y)\, dP(y).$$

Table 1: Examples of proper scoring rules for regression. Here $F_{\widehat{P}}$ is the CDF and $\widehat{p}$ is the PDF of the predictive distribution $\widehat{P}$.

Score name	**Definition**
Continuous ranked probability	$\mathrm{CRPS}(\widehat{P}, y) = \int_{\mathbb{R}} \left(F_{\widehat{P}}(t) - \mathbb{I}\{y \leq t\}\right)^2 dt$
Logarithmic	$\mathrm{LS}(\widehat{P}, y) = -\log \widehat{p}(y)$
Quadratic	$\mathrm{QS}(\widehat{P}, y) = -2\,\widehat{p}(y) + \int_{\mathbb{R}} \widehat{p}(t)^2 dt$
Squared error	$\mathrm{SE}(\widehat{P}, y) = \left(y - \mathbb{E}_{Y \sim \widehat{P}}[Y]\right)^2$

It represents the smallest achievable error, attained by the Bayes-optimal (pointwise) predictor.

Definition 2.2. The *divergence function* is the difference between the expected scores of the predicted and true distributions:

$$d(\widehat{P}, P) = S(\widehat{P}, P) - S(P, P) = S(\widehat{P}, P) - H(P).$$

It quantifies how much worse it is to predict with the distribution $\widehat{P}$, rather than with the true distribution P.

In the next section, we will interpret these notions point-by-point, with the true conditional distribution of the response and the conditional predictive distribution, as a decomposition of the predictive risk into Bayes and excess components.

Proper scoring rules for regression. Several proper scoring rules exist for continuous distributions (Gneiting & Raftery, 2007). We consider some of the most common proper scoring rules, such as the *continuous ranked probability score (CRPS), logarithmic, quadratic, and squared error scores* (defined in Table 1). All of these scores are strictly proper except for the squared error score, which is only proper. In what follows, we derive practical uncertainty estimates for regression based on these scores.

3 Regression Uncertainty Measures via Proper Scoring Rules

In this section, we introduce our uncertainty quantification framework for regression and derive practical expressions for uncertainty measures using proper scoring rules. We follow the approach of (Kotelevskii et al., 2025; Hofman et al., 2026), which associates predictive uncertainty with the expected error at a given point.

Our proposed framework consists of four steps to derive suitable UQ measures: choosing (I) a parametric assumption for the predicted distribution, (II) a risk quantity aligned with the desired UQ task, (III) a proper scoring rule to instantiate the risk quantity, and (IV) an approximation strategy. The first step is particular to regression. We exemplify the framework under the common Gaussian parametric assumption, while noting that the framework could be applied under other parametric assumptions, such as Laplace, or to nonparametric predictive distributions under different approximation schemes. The remaining steps follow the classification framework of Kotelevskii et al. (2025). In the following, we instantiate the framework by selecting the risk quantity, proper scoring rule, and Bayesian risk approximation under the Gaussian parametric assumption.

3.1 Pointwise Risk as an Uncertainty Measure

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ be a training set of n i.i.d. samples from the joint distribution $P(X, Y)$, where $X_i \in \mathcal{X} = \mathbb{R}^d$ and $Y_i \in \mathcal{Y} = \mathbb{R}$. We denote the true conditional distribution of Y at $X = x$ by $P(Y \mid X = x)$. The goal of statistical learning is to approximate this conditional distribution with a model $\widehat{P}(Y \mid X = x)$ using the training data. For brevity, we assume that everything is implicitly conditional on $X = x$, so we write $P = P(Y \mid X = x)$ and $\widehat{P} = \widehat{P}(Y \mid X = x)$. We denote the corresponding densities by p and $\widehat{p}$, and the cumulative distribution functions by F_P and $F_{\widehat{P}}$.

The expected score $S(\widehat{P}, P)$ in Equation (1) can be viewed as an expected prediction error of the model $\widehat{P}$ at a given point. Such an object is usually called *pointwise risk* in the statistical learning literature. In this

work, we adopt the view of uncertainty as an expected error. If the expected error of a model at a particular point is significant, we say that the associated predictive uncertainty in this point is also significant [1].

There are three key risk quantities that we consider in our framework:

- **Total risk.** We interpret the expected score between the estimated $\widehat{P}$ and the true P distributions as the total risk, capturing the overall uncertainty of a prediction:

$$R_{\text{Tot}}(\widehat{P}, P) = S(\widehat{P}, P).$$

- **Bayes risk** is the risk incurred when predicting with the *true* distribution itself:

$$R_{\text{Bayes}}(P) = S(P, P) = H(P).$$

  Since it does not depend on the model, the nature of this error is purely *aleatoric*. It is equal to the entropy (in terms of Definition 2.1) of the data and, for any proper score, it equals the smallest possible expected error of prediction.
- **Excess risk** measures how much worse the model is compared to the perfect prediction:

$$R_{\text{Exc}}(\widehat{P}, P) = R_{\text{Tot}}(\widehat{P}, P) - R_{\text{Bayes}}(P) = d(\widehat{P}, P).$$

  It is equal to the score's divergence (see Definition 2.2) and captures the part of the error not described by the randomness in the data itself; it naturally relates to *epistemic* uncertainty.

Example. Consider a fixed input x, and suppose that the true conditional distribution is $P(\cdot \mid x) = \mathcal{N}(0, 1)$, while the model predicts $\widehat{P}(\cdot \mid x) = \mathcal{N}(\mu, \sigma^2)$. Under the logarithmic score $S(\widehat{P}, y) = -\log \widehat{p}(y)$:

$$R_{\text{Tot}}(\widehat{P}, P) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1+\mu^2}{2\sigma^2},\ R_{\text{Bayes}} = H(P) = \frac{1}{2}\log(2\pi e),\ R_{\text{Exc}}(\widehat{P}, P) = \frac{1}{2}\left[\log\sigma^2 + \frac{1+\mu^2}{\sigma^2} - 1\right].$$

As we can see, when $\widehat{P} = P$, the excess term is zero. If the model predicts $\widehat{P} = \mathcal{N}(1, 1)$, the Bayes term remains unchanged but the excess term becomes $\frac{1}{2}$, reflecting a mismatch.

The risks above depend on the true distribution P, which is unknown in a practical setting. In the next section, we consider the same pointwise decomposition and develop various approximations of these risks.

3.2 Bayesian Risk Approximation

Following Kotelevskii et al. (2025); Schweighofer et al. (2025); Hofman et al. (2026), we consider Bayesian approximations to the risk expressions. We assume that we have a predictive parametric model, with the vector of parameters θ following a posterior distribution of $p(\theta \mid \mathcal{D})$, giving a *distribution* over predictive distributions $\widehat{P}_\theta$.

[1] *Remark.* The difference with the more classical statistical decision theory is that the prediction is now in the form of **distribution**, while the observation is still a single number.

Table 2: Bayes-risk approximations under different proper scoring rules. We use the following notation: $\widehat{P}_{\text{ens}} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{N}\left(\mu_i, \sigma_i^2\right)$, $\mathcal{N}\left(a \mid b, s^2\right) = \frac{1}{s}\phi(\frac{a-b}{s})$, $A(\mu, \sigma) = 2\sigma\phi\left(\frac{\mu}{\sigma}\right) + \mu\left[2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right]$, and Φ, ϕ are CDF and PDF of a standard normal. For $\widehat{R}^3_{\text{Bayes}}$, different plug-in estimates of the variance σ_*^2 can be used (see Appendix A.1.2).

App.	CRPS	Logarithmic	Quadratic	SE
$\widehat{R}^1_{\text{Bayes}}$	$\frac{1}{M\sqrt{\pi}}\sum_{i=1}^{M}\sigma_i$	$\frac{1}{2M}\sum_{i=1}^{M}\log\left(2\pi e{\sigma_i}^2\right)$	$-\frac{1}{2M\sqrt{\pi}}\sum_{i=1}^{M}\frac{1}{\sigma_i}$	$\frac{1}{M}\sum_{i=1}^{M}\sigma_i^2$
$\widehat{R}^2_{\text{Bayes}}$	$\frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}A(\mu_{ij}, \sigma_{ij})$	$-\int_{\mathbb{R}}\widehat{p}_{\text{ens}}(y)\log\widehat{p}_{\text{ens}}(y)dy$	$-\int_{\mathbb{R}}\widehat{p}_{\text{ens}}(y)^2 dy$	$\frac{1}{M}\sum_{i=1}^{M}(\sigma_i^2 + \mu_i^2) - \mu_*^2$
$\widehat{R}^3_{\text{Bayes}}$	$\sqrt{\sigma_*^2/\pi}$	$\frac{1}{2}\log\left(2\pi e\sigma_*^2\right)$	$-\frac{1}{2\sqrt{\pi}\sigma_*}$	σ_*^2

Table 3: Chosen combinations of approximations for risk estimation.

Predicted $\widehat{P}$	True P		
	Bayesian averaging	$\widehat{P}_{\text{ens}}$	$\widehat{P}_*$
Bayesian averaging	$\widehat{R}^{1,1}$	–	–
$\widehat{P}_{\text{ens}}$	$\widehat{R}^{2,1}$	–	–
$\widehat{P}_*$	$\widehat{R}^{3,1}$	$\widehat{R}^{3,2}$	–

Table 4: Excess risks $\widehat{R}^{1,1}_{\text{Exc}}$ (Bayesian averaging of both P and $\widehat{P}$) for different proper scoring rules.

Scoring Rule	$\widehat{R}^{1,1}_{\text{Exc}}$
CRPS	$\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left[A(\mu_{ij},\sigma_{ij}) - \frac{\sigma_i+\sigma_j}{\sqrt{\pi}}\right]$
Logarithmic	$\frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left[\frac{\sigma_j^2+(\mu_i-\mu_j)^2}{\sigma_i^2} - 1\right]$
Quadratic	$-\frac{2}{M}\sum_{i=1}^{M} H(\widehat{P}_i) - \frac{2}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\mathcal{N}\left(\mu_i \mid \mu_j, \sigma_i^2+\sigma_j^2\right)$
SE	$2\widehat{\text{Var}}\left[\mu_i\right] = \frac{2}{M}\sum_{i=1}^{M}\left(\mu_i - \mu_*\right)^2$

Although the framework is model-agnostic, we make practical parametric assumptions on $\widehat{P}_\theta$ to obtain tractable closed-form expressions. Specifically, we assume a Gaussian predictive distribution $\widehat{P}_\theta = \mathcal{N}\left(\mu, \sigma^2\right)$ with parameters μ, σ depending on the model parameters and location x. *Remark*: in the experiments we train a neural network to directly output the Gaussian parameters for each input x, but to simplify the notation, we absorb dependencies on the trained weights and x into a single θ. Other parametric assumptions, e.g., the Laplace distribution, could be used to derive different measures. Following the standard way of estimating posterior expectations $\mathbb{E}_{p(\theta|\mathcal{D})}[\cdot]$, we consider Monte-Carlo sampling and estimate, e.g., the posterior predictive as an ensemble of Gaussians: $\widehat{P}_{\text{ens}} = \frac{1}{M}\sum_{i=1}^{M}\widehat{P}_i = \frac{1}{M}\sum_{i=1}^{M}\mathcal{N}\left(\mu_i, \sigma_i^2\right)$, where $\theta_i \sim p(\theta \mid \mathcal{D})$.

From this perspective, we estimate any of the risks above in the following generic ways:

1. **Bayesian averaging.** Posterior expectation of risk: $\mathbb{E}_{p(\theta|\mathcal{D})}\left[R(\widehat{P}, \widehat{P}_\theta)\right]$.
2. **Posterior predictive distribution.** Average the distributions $\widehat{P}_{\text{ens}} = \mathbb{E}_{p(\theta|\mathcal{D})}\left[\widehat{P}_\theta\right]$, and use it as a plug-in estimate: $R(\widehat{P}, \widehat{P}_{\text{ens}})$.
3. **Single Gaussian surrogate.** We consider two Gaussian approximations for the posterior $\widehat{P}_*, \bar{P}_*$ (see Appendix A.1.2), and use them as plug-ins estimates: $R(\widehat{P}, \widehat{P}_* \text{ or } \bar{P}_*)$.

Note that either of these approaches can also be used to derive a specific estimate for $\widehat{P}$ as well, see (Kotelevskii et al., 2025; Schweighofer et al., 2025; Hofman et al., 2026).

Practical approximation. Combining the three approximation strategies for P with the three choices for $\widehat{P}$ yields nine (in general) possible pairs $(\widehat{P}, P)$. We choose only a subset of them based on the understanding that the approximations for P should be at least as accurate as those for $\widehat{P}$. The choices are summarized in Table 3: notice that the true distribution is estimated by Bayesian averaging or by the full posterior predictive distribution, and the prediction distribution is estimated using either of the three approaches.

Therefore, any risk estimate considered from now on carries two superscripts $\widehat{R}^{i,j}$, stating the choice of approximation strategies. Numbers correspond to the generic way to approximate risks we outlined above, e.g. 1 denotes Bayesian averaging of risks. For Bayes risk, we need only approximate the true distribution P, for which we have three options. We exemplify the resulting expressions of Monte-Carlo estimates for popular regression proper scoring rules for the approximation of Bayes risk in Table 2, and $\widehat{R}^{1,1}_{\text{Exc}}$ in Table 4. We refer to Appendix A for complete derivations and results under other approximation choices.

Misspecification. The definitions of total, Bayes, and excess risk do not assume a Gaussian model. Gaussianity enters only when we build closed-form estimators from ensemble members. If the predictive family cannot represent skewness, heavy tails, or multimodality, the resulting AU/EU estimates are biased proxies for the corresponding proper scoring rule risks. This is a modeling limitation rather than a limitation of risk decomposition. In such cases, the same construction can be applied with a richer predictive family, such as Laplace, mixtures, or nonparametric estimators, but closed forms may be replaced by numerical integration or Monte Carlo estimates. Thus, our Gaussian formulas should be read as a practical instantiation, not as an explicit assumption of the framework.

Generalization of earlier work. Bülte et al. (2025) recently studied entropy- and variance-based uncertainty measures for regression:

$$\mathrm{H}(\mathbb{E}_{p(\theta|\mathcal{D})}[\widehat{P}_\theta]) = \mathbb{E}_{p(\theta|\mathcal{D})}[\mathrm{H}(\widehat{P}_\theta)] + \mathbb{E}_{p(\theta|\mathcal{D})}[d_{\mathrm{KL}}(\widehat{P}_\theta, \mathbb{E}_{p(\theta|\mathcal{D})}[\widehat{P}_\theta])],$$
$$\mathrm{Var}(\mathbb{E}_{p(\theta|\mathcal{D})}[\widehat{P}_\theta]) = \mathbb{E}_{p(\theta|\mathcal{D})}[\mathrm{Var}(\widehat{P}_\theta)] + \mathrm{Var}_{p(\theta|\mathcal{D})}[\mathbb{E}(\widehat{P}_\theta)],$$

where $\mathrm{H}(\cdot)$ is the Shannon entropy and $d_{\mathrm{KL}}(\cdot,\cdot)$ is the Kullback-Leibler divergence.

In our framework, these measures arise naturally as the *logarithmic* and *squared error* special cases (using $\widehat{R}^{2,1}_{\mathrm{Tot}} = \widehat{R}^{1}_{\mathrm{Bayes}} + \widehat{R}^{2,1}_{\mathrm{Exc}}$ risk approximation); see the derivations in Appendices A.2.2 and A.2.4, and in Example 2.1 of (Bülte et al., 2025).

Our framework thus comprises well-established entropy- and variance-based measures, as well as measures based on CRPS (widely considered in meteorology and time-series forecasting) and the quadratic score. New measures can readily be derived under our framework for different parametric assumptions and other choices of proper scoring rules.

4 Related Work

Axiomatic approaches in classification. Despite the maturity of uncertainty quantification, the formalization of uncertainty measures appeared only recently. Most of the efforts in this formalization were made in the context of classification, where the authors tried to formalize what makes a "good" uncertainty score. Specifically, Wimmer et al. (2023); Sale et al. (2024) propose different sets of axioms for classification, but there seems to be no existing measure that satisfies all of them.

Risk-based uncertainty measures. The idea of viewing predictive uncertainty as *pointwise risk*, the expected value of a loss, was first proposed for classification in (Kotelevskii et al., 2022) and for both settings in (Lahlou et al., 2023). However, they considered a specific class of models and a specific set of loss functions. Based on this view, (Schweighofer et al., 2025; Kotelevskii et al., 2025; Hofman et al., 2026) showed that the risk of any proper scoring rule admits a clean decomposition into *Bayes risk* (aleatoric part) and *excess risk* (epistemic part), where the former corresponds to the generalized entropy, while the latter is the notion of a Bregman divergence (Bregman, 1967). However, generalizing the ideas of using proper scoring rules to generate uncertainty measures for regression was not considered.

Towards regression. Recent work (Bülte et al., 2025) introduces axioms for regression and assesses entropy and variance for them. However, they do not introduce other measures that can be derived from considering proper scoring rules. Our paper fills this gap by deriving regression-specific scores from proper scoring rules, specifically the CRPS score (Alet et al., 2025; Lang et al., 2025).

Practical deep-learning baselines. Bayesian dropout (Gal & Ghahramani, 2016), Deep Ensembles (Lakshminarayanan et al., 2017), and heteroscedastic losses (Kendall & Gal, 2017) remain standard tools of uncertainty quantification in regression in Deep Learning, because they scale well, although they lack a firm axiomatic footing. Deep Evidential Regression (Amini et al., 2020) offers a single-pass alternative that outputs Normal-Inverse-Gamma parameters, giving closed-form aleatoric and epistemic variances, yet recent studies question how reliably DER captures epistemic risk (Juergens et al., 2024).

5 Experiments

We investigate the empirical behavior of the measures introduced in Section 3. We begin by examining how these measures respond to different perturbations of the posterior and showcase their behavior on a synthetic regression task. Then we evaluate their utility on three downstream problems: selective prediction, out-of-distribution detection, and active learning. To further assess the relationship between measures, we evaluate rank correlations between measures and examine when different scores lead to different decisions. Unless stated otherwise, posterior expectations are computed using a Monte-Carlo approximation via heteroscedastic deep ensembles (Lakshminarayanan et al., 2017); training details are provided in Appendix E.1.

Table 5: Overview of the established UQ measures recovered by our framework.

Common UQ baseline	Expressed in our framework as
Mean predicted variance	SE score, $\widehat{R}^{1}_{\text{Bayes}} = M^{-1}\sum_{i=1}^{M}\sigma_i^2$
Ensemble predictions variance	SE score, $\widehat{R}^{1,1}_{\text{Exc}} \sim \widehat{\text{Var}}[\mu_i]$
Predictive variance	SE score, $\widehat{R}^{2,1}_{\text{Tot}} = \widehat{R}^{2}_{\text{Bayes}} = M^{-1}\sum_i \sigma_i^2 + \widehat{\text{Var}}[\mu_i]$
Predictive entropy	Log score, $\widehat{R}^{2,1}_{\text{Tot}} = \widehat{R}^{2}_{\text{Bayes}} = H(\widehat{P}_{\text{ens}})$ = mixture entropy
Disagreement and Mutual information-type measures	Log score, $\widehat{R}^{2,1}_{\text{Exc}}$ Lakshminarayanan et al. (2017) and Log score, $\bar{R}^{3,1}_{\text{Exc}}$ Bülte et al. (2025), see A.2.2, eq. (20)

Since several standard regression UQ measures are recovered by our framework, we report them using the corresponding risk notation in Table 5 and compare them with the newly derived CRPS- and quadratic-score variants, all computed from the same predictive ensemble.

5.1 Characterization of Uncertainty Measures

In Section 3.2 we introduced different approximation strategies for the uncertainty measures that rely on Bayesian modelling and posterior distribution of (μ, σ). Before moving on to larger experiments where we train ensembles of models as a proxy to samples from the posterior, we first conduct synthetic experiment.

Here we consider the case when the posterior is known: a uniform distribution in the predicted means $\mu \sim \mathcal{U}(-1, 1)$ and a uniform distribution in the predicted variances $\sigma^2 \sim \mathcal{U}(1, 2)$. This closed-form posterior allows us to evaluate posterior expectations exactly and thus compute uncertainty measures directly without relying on sampling methods such as deep ensembles.

We investigate the behavior of the different uncertainty measures by applying shifts to the posterior, inspired by the investigation in (Bülte et al., 2025, Figure 2). We consider four different shifts to the posterior: (a) a location shift on the predicted means, (b) a location shift on the predicted variances, (c) a scale shift of the predicted means, and (d) a scale shift of the predicted variances; see Figure 1 (left column).

Location shift of predicted means. The distribution of predicted means shifts to $\mu \sim \mathcal{U}(1, 3)$, thus the location of the posterior shifts, but not its scale (Figure 1, first row). None of the measures that we examine change under this shift. This is desirable, as uncertainty measures should not be sensitive to the magnitude of mean predictions.

Location shift of predicted variances. The distribution of predicted variances shifts to $\sigma^2 \sim \mathcal{U}(2, 3)$, thus the location of the posterior shifts, but not its scale. The results are shown in Figure 1 (second row). It is expected that all uncertainty measures increase under such a shift. However, we observe that excess risk measures either decrease or remain unchanged (for $\widehat{R}^{3a,2}_{\text{Exc}}$ under CRPS and Quadratic, and for all excess risks under SE score). Here, staying the same is a desirable outcome over decreasing.

Scale shift of predicted means. The distribution of predicted means shifts to $\mu \sim \mathcal{U}(-2, 2)$, thus the scale of the posterior shifts, but not its location. The results are shown in Figure 1 (third row). It is expected that all uncertainty measures increase under such a shift. We observe that $\widehat{R}^{1}_{\text{Bayes}}$ and $\widehat{R}^{3b}_{\text{Bayes}}$ under all scores are invariant under such a shift, the same for $\widehat{R}^{3b,2}_{\text{Tot}}$ under SE. This could reduce their viability in downstream tasks. All other measures increase under the scale shift in the predicted means as expected.

Scale shift of predicted variances. The distribution of predicted variances shifts to $\sigma^2 \sim \mathcal{U}(0.5, 2.5)$, thus the scale of the posterior shifts, but not its location. The results are shown in Figure 1 (fourth row). It is expected that total and epistemic uncertainty will increase under such a shift, while aleatoric uncertainty will remain unchanged.

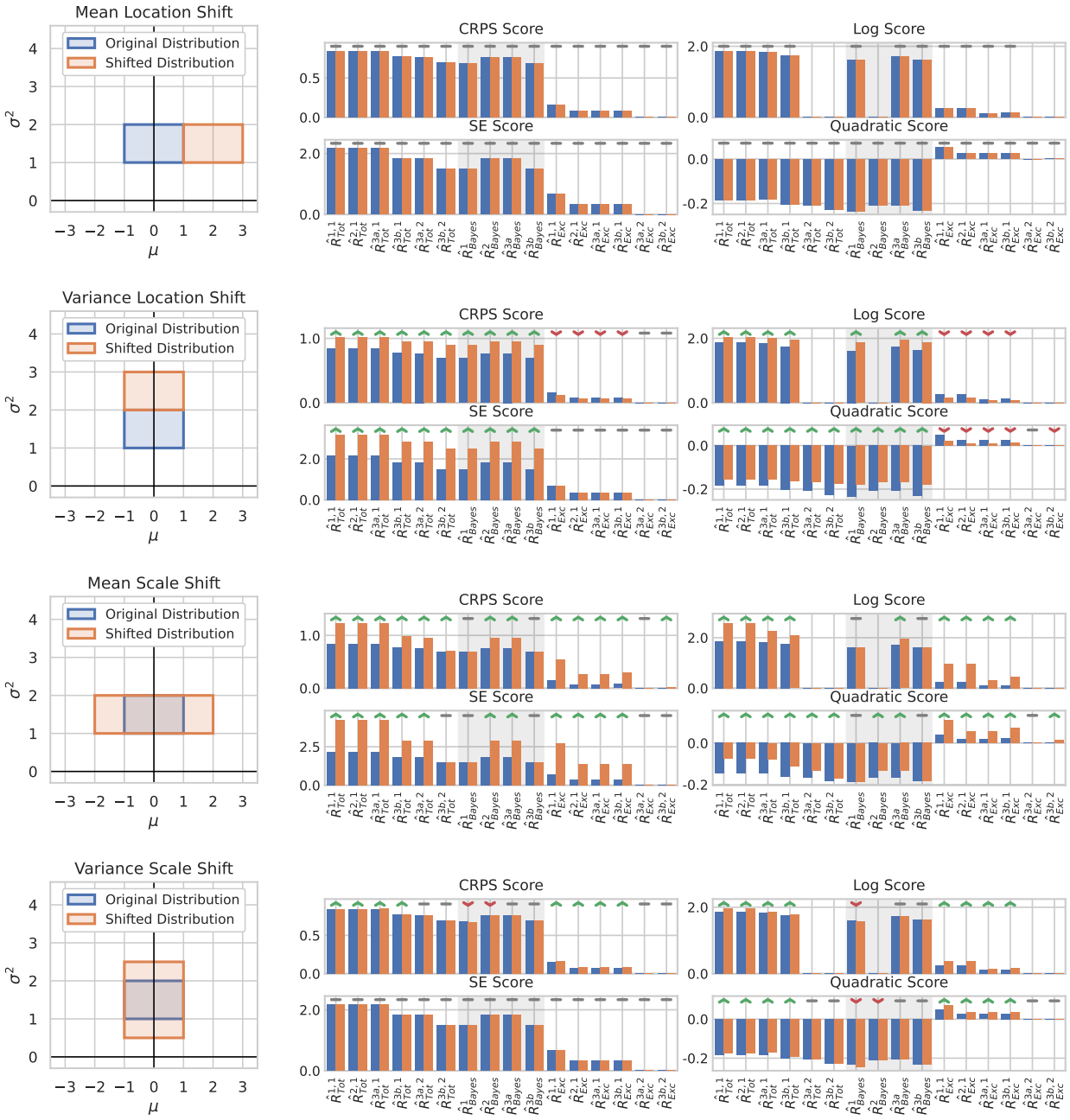


Figure 1: Behavior of uncertainty measures under location and scale shifts of the posterior of parameters (μ, σ^2). We assume that $\mu \sim \mathcal{U}(-1, 1)$ and $\sigma^2 \sim \mathcal{U}(1, 2)$ under the initial posterior distribution. The shifts are shown in the leftmost column, and the plots on the right show whether measures increase, decrease, or are independent of those shifts.

General observations. Across all shifts, we observe that $\widehat{R}^{3a,2}_{\text{Exc}}$ and $\widehat{R}^{3b,2}_{\text{Exc}}$ have very low magnitudes, and for the SE score, they are equal to zero. The magnitude of the excess risks is often much lower than the Bayes risks, and they scale most with the mean scale shift, which is expected.

5.2 Synthetic Datasets

Having characterized how the measures react to controlled shifts of the posterior, we now turn to a synthetic data example that verifies these behaviors against common intuition. Specifically, inputs far from the training

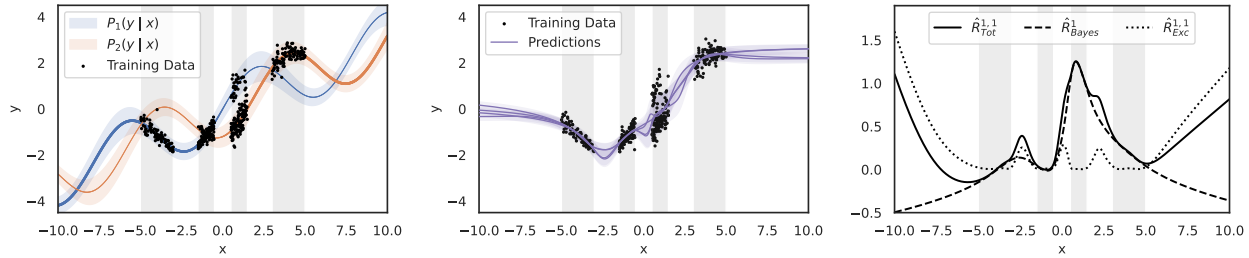


Figure 2: Uncertainty with the logarithmic scoring rule for the synthetic task. **Left:** The ground truth distribution is a mixture of two distributions. Line thickness indicates the weight of the mixture component. Intervals where training data is sampled are highlighted in gray. **Middle:** Predictive distributions of different models trained on the training data. **Right:** Resulting Total, Bayes, and Excess risks.

Table 6: Results for selective prediction (PRR ↓) for different scoring rules, averaged over datasets. Value format is *mean* ± *std*. The best result is underlined, all within its standard deviation in **bold**.

Measure	CRPS	Logarithmic	Squared error	Quadratic
$\widehat{R}^{1,1}_{\text{Tot}}$	**0.318** (.007)	**0.323** (.007)	**0.320** (.007)	**0.319** (.007)
$\widehat{R}^{2,1}_{\text{Tot}}$	**0.318** (.007)	**0.323** (.007)	**0.320** (.007)	**0.319** (.007)
$\widehat{R}^{3a,1}_{\text{Tot}}$	**0.319** (.007)	**0.321** (.007)	**0.320** (.007)	**0.323** (.007)
$\widehat{R}^{3b,1}_{\text{Tot}}$	0.327 (.007)	0.327 (.007)	0.327 (.007)	0.327 (.007)
$\widehat{R}^{3a,2}_{\text{Tot}}$	0.327 (.007)	–	0.327 (.007)	0.329 (.007)
$\widehat{R}^{3b,2}_{\text{Tot}}$	0.352 (.007)	–	0.357 (.007)	0.347 (.007)
$\widehat{R}^{1}_{\text{Bayes}}$	0.356 (.007)	0.356 (.007)	0.357 (.007)	0.356 (.007)
$\widehat{R}^{2}_{\text{Bayes}}$	0.327 (.007)	–	0.327 (.007)	0.326 (.007)
$\widehat{R}^{3a}_{\text{Bayes}}$	0.327 (.007)	0.327 (.007)	0.327 (.007)	0.327 (.007)
$\widehat{R}^{3b}_{\text{Bayes}}$	0.357 (.007)	0.357 (.007)	0.357 (.007)	0.357 (.007)
$\widehat{R}^{1,1}_{\text{Exc}}$	0.339 (.006)	0.397 (.006)	**0.325** (.006)	0.514 (.007)
$\widehat{R}^{2,1}_{\text{Exc}}$	0.339 (.006)	0.397 (.006)	**0.325** (.006)	0.514 (.007)
$\widehat{R}^{3a,1}_{\text{Exc}}$	0.340 (.006)	0.400 (.006)	**0.325** (.006)	0.523 (.006)
$\widehat{R}^{3b,1}_{\text{Exc}}$	0.341 (.006)	0.398 (.007)	**0.325** (.006)	0.511 (.007)
$\widehat{R}^{3a,2}_{\text{Exc}}$	0.504 (.006)	–	–	0.689 (.004)
$\widehat{R}^{3b,2}_{\text{Exc}}$	0.378 (.006)	–	–	0.503 (.007)

Table 7: Results for OOD detection (AUROC ↑) for different scoring rules, averaged over datasets. Value format is *mean* ± *std*. The best result is underlined, all within its standard deviation in **bold**.

Measure	CRPS	Logarithmic	Squared error	Quadratic
$\widehat{R}^{1,1}_{\text{Tot}}$	0.794 (.012)	0.802 (.011)	0.792 (.012)	0.800 (.012)
$\widehat{R}^{2,1}_{\text{Tot}}$	0.794 (.012)	0.802 (.011)	0.792 (.012)	0.800 (.012)
$\widehat{R}^{3a,1}_{\text{Tot}}$	0.794 (.012)	0.797 (.012)	0.792 (.012)	0.800 (.012)
$\widehat{R}^{3b,1}_{\text{Tot}}$	0.779 (.013)	0.785 (.013)	0.777 (.012)	0.784 (.013)
$\widehat{R}^{3a,2}_{\text{Tot}}$	0.777 (.012)	–	0.777 (.012)	0.777 (.013)
$\widehat{R}^{3b,2}_{\text{Tot}}$	0.727 (.009)	–	0.714 (.010)	0.740 (.011)
$\widehat{R}^{1}_{\text{Bayes}}$	0.714 (.010)	0.713 (.010)	0.714 (.010)	0.713 (.010)
$\widehat{R}^{2}_{\text{Bayes}}$	0.777 (.012)	–	0.777 (.012)	0.777 (.012)
$\widehat{R}^{3a}_{\text{Bayes}}$	0.777 (.012)	0.777 (.012)	0.777 (.012)	0.777 (.012)
$\widehat{R}^{3b}_{\text{Bayes}}$	0.714 (.010)	0.714 (.010)	0.714 (.010)	0.714 (.010)
$\widehat{R}^{1,1}_{\text{Exc}}$	**0.825** (.010)	**0.827** (.011)	**0.820** (.010)	**0.816** (.016)
$\widehat{R}^{2,1}_{\text{Exc}}$	**0.825** (.010)	**0.827** (.011)	**0.820** (.010)	**0.816** (.016)
$\widehat{R}^{3a,1}_{\text{Exc}}$	**0.825** (.010)	**0.826** (.011)	**0.820** (.010)	**0.816** (.016)
$\widehat{R}^{3b,1}_{\text{Exc}}$	**0.825** (.010)	**0.826** (.011)	**0.820** (.010)	**0.817** (.015)
$\widehat{R}^{3a,2}_{\text{Exc}}$	0.795 (.006)	–	–	0.783 (.005)
$\widehat{R}^{3b,2}_{\text{Exc}}$	**0.826** (.010)	–	–	**0.822** (.012)

distribution should have greater epistemic uncertainty, while regions where the data-generating process is inherently ambiguous should yield higher aleatoric uncertainty.

In our 1D regression experiment (see results in Figure 2 and its description in Appendix E.2), the *excess risk* estimate, a proxy for epistemic uncertainty, increases for inputs outside the training data support, where the model cannot reliably infer the underlying dependency. The *Bayes risk* estimate, capturing aleatoric uncertainty, is largest where the mapping from x to y is least determined. In our synthetic example, this occurs in the region where the probabilities of sampling from the two generating curves are similar. Consequently, the *total risk* is high whenever either component is high.

For clarity, we present a Log-score instantiation (which leads to entropy-based measures) in the main text and visualize the corresponding approximations of Bayes, excess, and total risk in Figure 2. Log-score-based uncertainty measures are widely used in practice (Bülte et al., 2025), and our empirical results support this common choice. We also observe the same qualitative behavior for other instantiations, with additional visualizations provided in Appendix E.2.

5.3 Selective Prediction

In selective prediction, the model's performance is evaluated on a specific subset that is defined as having low uncertainty. Therefore, uncertainty measures the ability to indicate whether the prediction is correct or not. We considered multiple datasets; see Appendix E.3 for details. Empirical findings for selective prediction in classification (e.g., Kotelevskii et al., 2025; Schweighofer et al., 2025) suggest that total uncertainty performs well in this task, which we seek to validate for the regression setting.

The results are in Table 6, which are averages over all the considered datasets (see also Table 11 for detailed results). Performances are measured as prediction-reject ratios or PRRs (see Appendix E.3), where a lower PRR indicates more effective rejection of inaccurate predictions using the uncertainty score. The best performing measures are $\widehat{R}^{1,1}_{\text{Tot}}$, $\widehat{R}^{2,1}_{\text{Tot}}$ and $\widehat{R}^{3a,1}_{\text{Tot}}$ for all considered scoring rules. Furthermore, all excess risks under the SE score (equivalent to each other) perform similarly well. Noteworthy, excess risks for the quadratic score are performing very poorly compared to all other measures; the same for $\widehat{R}^{3a,2}_{\text{Exc}}$ for the CRPS scoring rule. Excess risks under the widely considered Log-score perform slightly worse than most other measures. Thus, we recommend using $\widehat{R}^{1,1}_{\text{Tot}}$ for selective prediction: it leads to the overall best performance under CRPS, and is consistently among the best across scoring rules.

5.4 Out-of-Distribution Detection

Out-of-distribution (OOD) detection is an important problem, as models trained on in-distribution (ID) data have no performance guarantees beyond this distribution. Uncertainty quantification helps address this by assigning high uncertainty values to OOD inputs, indicating they should be further inspected or rejected. For our OOD experiments, we considered a dataset consisting of a mosaic of four MNIST (LeCun et al., 1998) images, where the target is given as the number formed by those four digits, as an ID dataset. We considered multiple OOD datasets; for details; see Appendix E.4.

This task is evaluated using the AUROC of the uncertainty score to distinguish between ID and OOD data. Results are provided in Table 7, averaged over all considered ID and OOD data pairs. We refer to Table 12 in the Appendix for detailed results per ID/OOD pair. We find that excess risk measures are generally the most suitable for this task, particularly $\widehat{R}^{1,1}_{\text{Exc}}$. This aligns with the assumptions on excess risk in prior work (Lahlou et al., 2023; Hofman et al., 2026; Kotelevskii et al., 2025).

Table 8: Results for active learning (average rank ↓ over datasets and five seeds using different scores as acquisition functions). Value format is *mean* ± *std*. The best result is underlined, all within its standard deviation in **bold**.

Measure	CRPS	Logarithmic	Squared error	Quadratic
$\widehat{R}^{1,1}_{\text{Tot}}$	16.6 (3.6)	15.7 (2.6)	16.9 (4.8)	29.2 (2.1)
$\widehat{R}^{2,1}_{\text{Tot}}$	16.6 (3.6)	15.7 (2.6)	16.9 (4.8)	29.2 (2.1)
$\widehat{R}^{3a,1}_{\text{Tot}}$	16.6 (4.3)	15.0 (2.4)	16.9 (4.8)	32.6 (2.3)
$\widehat{R}^{3b,1}_{\text{Tot}}$	17.8 (3.3)	15.5 (3.6)	16.9 (5.3)	29.5 (5.1)
$\widehat{R}^{3a,2}_{\text{Tot}}$	17.6 (4.2)	–	16.7 (5.0)	31.7 (1.9)
$\widehat{R}^{3b,2}_{\text{Tot}}$	18.3 (2.8)	–	16.5 (2.9)	31.3 (4.2)
$\widehat{R}^{1}_{\text{Bayes}}$	19.1 (4.4)	19.1 (2.2)	16.5 (2.9)	31.1 (3.1)
$\widehat{R}^{2}_{\text{Bayes}}$	16.6 (2.8)	–	16.7 (5.0)	32.5 (1.1)
$\widehat{R}^{3a}_{\text{Bayes}}$	17.7 (3.7)	18.0 (3.6)	16.7 (5.0)	30.3 (3.8)
$\widehat{R}^{3b}_{\text{Bayes}}$	19.2 (3.0)	21.5 (2.6)	16.5 (2.9)	31.8 (4.2)
$\widehat{R}^{1,1}_{\text{Exc}}$	17.2 (3.6)	**13.9** (1.3)	17.5 (3.7)	16.1 (3.4)
$\widehat{R}^{2,1}_{\text{Exc}}$	17.2 (3.6)	**13.9** (1.3)	17.5 (3.7)	16.1 (3.4)
$\widehat{R}^{3a,1}_{\text{Exc}}$	15.9 (3.7)	17.3 (6.0)	16.0 (5.2)	**12.6** (2.2)
$\widehat{R}^{3b,1}_{\text{Exc}}$	15.1 (4.0)	**14.4** (3.9)	16.0 (5.2)	**13.4** (3.9)
$\widehat{R}^{3a,2}_{\text{Exc}}$	15.3 (3.0)	–	–	17.0 (2.3)
$\widehat{R}^{3b,2}_{\text{Exc}}$	15.9 (3.5)	–	–	15.9 (2.5)
Random	20.5 (5.9)	20.5 (5.9)	20.5 (5.9)	20.5 (5.9)

5.5 Active Learning

Another widely considered real-world task that uses uncertainty information is active learning. Uncertainty guides which samples are most interesting to obtain labels for and add to the training dataset. Generally, excess risks (epistemic uncertainty) are considered to be better guidance than total or Bayes risks (Mukhoti et al., 2023).

The results are provided in Table 8, showing average ranks over the considered datasets; see Appendix E.5 for details. As expected, we observe that excess risk measures perform best, but there is no clear best approximation strategy. Most outperform random sampling, yet there are some exceptions: the total and Bayes risks under the quadratic score and $\widehat{R}^{3b}_{\text{Bayes}}$ under the log score.

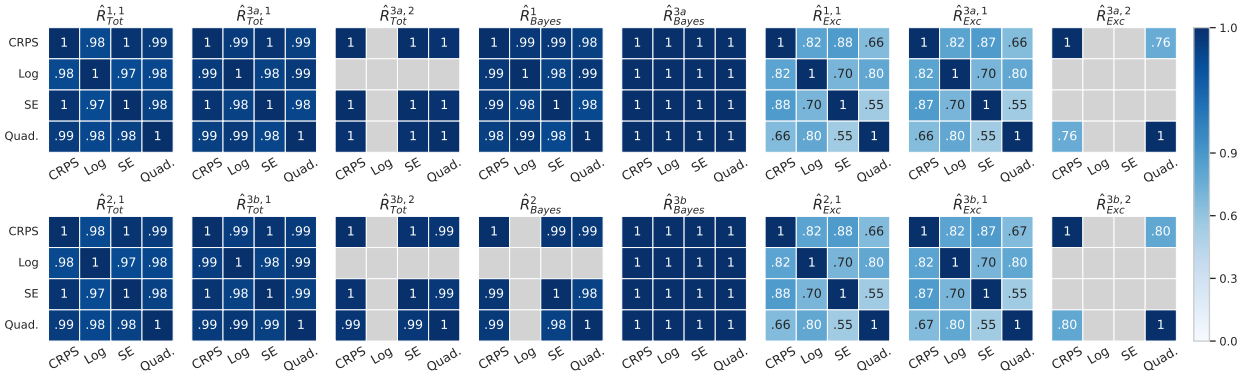


Figure 3: Kendall's τ_b rank correlation between scoring rules for the considered risk approximations. Correlations averaged over datasets. Correlations are high for total and Bayes risk and substantially lower for excess risks.

5.6 Rank Correlation Between Measures

Finally, we inspect the rank correlation between the measures within our framework. We use the same datasets and models as for selective prediction and calculate the rank correlations induced by different uncertainty measures. In Figure 4, we observe that $\widehat{R}^{1}_{\text{Bayes}}$ is strongly correlated to $\widehat{R}^{3b}_{\text{Bayes}}$ (same for $\widehat{R}^{2}_{\text{Bayes}}$, $\widehat{R}^{3a}_{\text{Bayes}}$). Furthermore, $\widehat{R}^{1,1}_{\text{Tot}}$ and $\widehat{R}^{2,1}_{\text{Tot}}$ are equivalent due to the linearity of expectation, the same holds for $\widehat{R}^{1,1}_{\text{Exc}}$ and $\widehat{R}^{2,1}_{\text{Exc}}$. Regarding the differences between induced rankings for different scoring rules for the same approximations (see Figure 3), a striking finding is that for both $\widehat{R}^{3a}_{\text{Bayes}}$ and $\widehat{R}^{3b}_{\text{Bayes}}$, all scoring rules are perfectly correlated, the same for $\widehat{R}^{3a,2}_{\text{Tot}}$ for those scoring rules it is computable for. Rank correlations are overall very high among approximations of total and Bayes risks, and only vary substantially across excess risks. Although the total and excess risk estimates are not perfectly uncorrelated, performance on previous tasks shows the usefulness of using different risks for specific tasks. The related work of Mucsányi et al. (2024) investigates this for classification.

6 Discussion

General findings and failure modes. Our experiments show that the specific choice of a proper scoring rule and approximation strategy has a relatively minor effect on downstream performance. The exception is $\bar{R}^{3a,2}_{\text{Exc}}$, which we found performs significantly worse throughout all considered experiments. Furthermore, this particular risk approximation was found to have undesirable characteristics across scoring rules, such as invariance under mean scale shift, linking theoretical behavior to downstream performance. We hypothesize that this is caused by the fact that $\bar{R}^{3a,2}_{\text{Exc}}$ uses two approximation strategies that average out the disagreement between individual models. Additionally, we found that no measure of Bayes risk, regardless of scoring rule or approximation strategy, is best in any of the considered tasks.

Limitations. Robustness of the uncertainty measures we consider is limited to the setting of selective prediction and out-of-distribution detection. The former corresponds to the failure detection setting, while the latter is modelled as covariate shift. The uncertainty measures derived from our framework perform well and consistent in those setting, but a broader investigation would be required to study the other aspects of a trustworthy AI system, like subgroup fairness and adversarial robustness.

Recommendations. Our empirical findings in Section 5 show that different risks are desirable as measures in different tasks, leading to the following recommendations. For selective prediction, we find that total risk leads to the best results across particular scoring rules and approximation strategies. For both out-of-distribution detection and active learning, we find that excess risks perform best. While most scoring rules

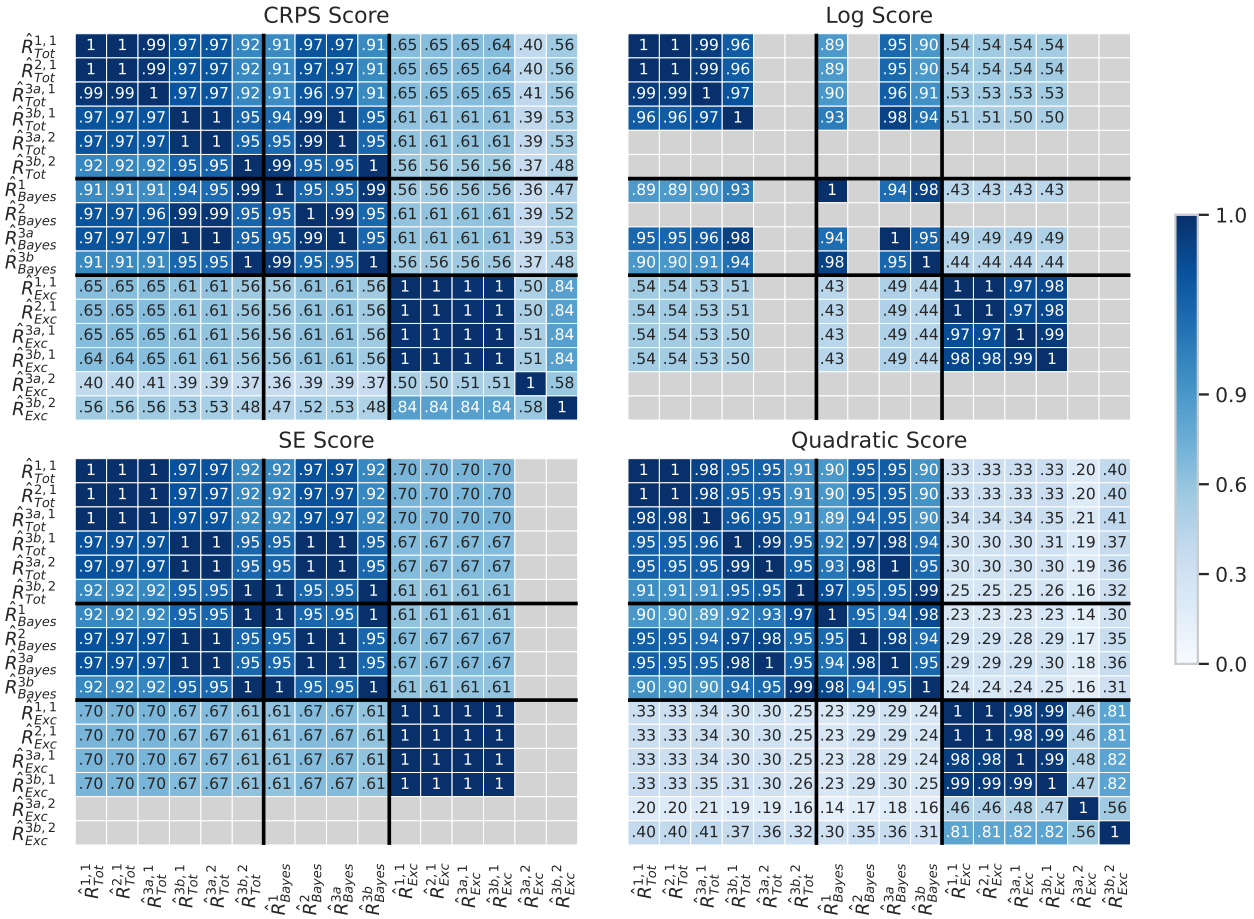


Figure 4: Kendall's τ_b rank correlation between different risk approximations for different scoring rules. Correlations are averaged over all considered datasets. Correlations between different total and Bayes risks are high, between total and excess risks as well as Bayes and excess risks substantially lower. Under some scoring rules, different approximations result in the same ranking and thus perfect rank correlation.

and Bayesian approximation strategies give reasonable results, we recommend $\bar{R}^{1,1}_{\text{Tot}}$ and $\bar{R}^{1,1}_{\text{Exc}}$ for any of the considered scoring rules as a default choice.

7 Conclusion

We developed a unified framework for regression uncertainty quantification by deriving practical predictive-uncertainty measures from proper scoring rule pointwise risk decompositions. The framework allows one to obtain measures of total, aleatoric, and epistemic uncertainty that generalize existing variance- and entropy-based approaches and enable the derivation of new measures. We showcase our general framework under the Gaussian parametric assumption, using Deep Ensembles to compute Monte-Carlo approximations of posterior expectations to estimate the risks. We chose this particular instantiation due to its widespread use in practice.

Our synthetic experiments confirm that these measures perform reasonably under controlled perturbations and show decent performance on practical downstream tasks. The analysis also shows in which scenarios different scoring rules agree and when they diverge, offering guidance for practitioners. Our goal is not to provide a universally superior uncertainty estimator. Instead, we show that the risk component - total, Bayes, or excess - is often more important than the specific proper score, and we provide task-tailored recommendations for choosing this component. Overall, our work bridges the gap between theoretical considerations and practical deep learning methods, providing a principled and versatile foundation for regression UQ.

References

Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*, 2025. doi: 10.48550/arXiv.2506.10772. URL https://arxiv.org/abs/2506.10772.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html.

Rafael Ballester-Ripoll, Enrique G Paredes, and Renato Pajarola. Sobol tensor trains for global sensitivity analysis. *Reliability Engineering & System Safety*, 183:311–322, 2019. doi: 10.1016/j.ress.2018.11.007. URL https://doi.org/10.1016/j.ress.2018.11.007.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pp. 591–596, 2011. URL http://ismir2011.ismir.net/papers/OS6-1.pdf.

Lev M Bregman. 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. doi: 10.1016/0041-5553(67)90040-7. URL https://doi.org/10.1016/0041-5553(67)90040-7.

Christopher Bülte, Yusuf Sale, Timo Löhr, Paul Hofman, Gitta Kutyniok, and Eyke Hüllermeier. An axiomatic assessment of entropy-and variance-based uncertainty quantification in regression. *arXiv preprint arXiv:2504.18433*, 2025. doi: 10.48550/arXiv.2504.18433. URL https://arxiv.org/abs/2504.18433.

Krisztian Buza. Feedback prediction for blogs. In *Data Analysis, Machine Learning and Knowledge Discovery*, pp. 145–152. Springer, 2013. doi: 10.1007/978-3-319-01595-8_16. URL https://doi.org/10.1007/978-3-319-01595-8_16.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017. doi: 10.1109/IJCNN.2017.7966217. URL https://arxiv.org/abs/1702.05373.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016. doi: 10.1109/CVPR.2016.350. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1184–1193. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/depeweg18a.html.

Jean-Louis Durrieu, Jean-Philippe Thiran, and Finnian Kelly. Lower and upper bounds for approximation of the kullback-leibler divergence between gaussian mixture models. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4833–4836, 2012. doi: 10.1109/ICASSP.2012.6289001. URL https://doi.org/10.1109/ICASSP.2012.6289001.

Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. A proactive intelligent decision support system for predicting the popularity of online news. In *Progress in Artificial Intelligence*, pp. 535–546, Cham, 2015. Springer International Publishing. doi: 10.1007/978-3-319-23485-4_53. URL https://doi.org/10.1007/978-3-319-23485-4_53.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138. URL https://arxiv.org/abs/0912.0902.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://stat.uw.edu/research/tech-reports/strictly-proper-scoring-rules-prediction-and-estimation-revised.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Marieka A Helou, Deborah DiazGranados, Michael S Ryan, and John W Cyrus. Uncertainty in decision making in medicine: a scoping review and thematic analysis of conceptual models. *Academic Medicine*, 95(1):157–165, 2020. doi: 10.1097/ACM.0000000000002902. URL https://doi.org/10.1097/ACM.0000000000002902.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. doi: 10.48550/arXiv.1606.08415. URL https://arxiv.org/abs/1606.08415.

John R. Hershey and Peder A. Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pp. IV–317–IV–320, 2007. doi: 10.1109/ICASSP.2007.366913. URL https://doi.org/10.1109/ICASSP.2007.366913.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Uncertainty quantification for machine learning: One size does not fit all. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(26):21726–21734, 2026. doi: 10.1609/aaai.v40i26.39323. URL https://ojs.aaai.org/index.php/AAAI/article/view/39323.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. doi: 10.1007/s10994-021-05946-3. URL https://doi.org/10.1007/s10994-021-05946-3.

Alexander Immer, Emanuele Palumbo, Alexander Marx, and Julia Vogt. Effective bayesian heteroscedastic regression with deep neural networks. *Advances in Neural Information Processing Systems*, 36:53996–54019, 2023. doi: 10.52202/075280-2349. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a901d5540789a086ee0881a82211b63d-Abstract-Conference.html.

Mira Juergens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22624–22642. PMLR, 2024. URL https://proceedings.mlr.press/v235/juergens24a.html.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017. doi: 10.48550/arXiv.1703.04977. URL https://papers.nips.cc/paper/7141-what-uncertainties-do-we-need.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. doi: 10.48550/arXiv.1412.6980. URL https://arxiv.org/abs/1412.6980.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://papers.neurips.cc/paper/8925-batchbald-efficient-and-diverse-batch-acquisition-for-deep-bayesian-active-learning.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. Nonparametric uncertainty quantification for single deterministic neural network. *Advances in Neural Information Processing Systems*, 35:36308–36323, 2022. doi: 10.52202/068431-2631. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/eb7389b039655fc5c53b11d4a6fa11bc-Abstract-Conference.html.

Nikita Kotelevskii, Samuel Horváth, Karthik Nandakumar, Martin Takác, and Maxim Panov. Dirichlet-based uncertainty quantification for personalized federated learning with improved posterior networks. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 7127–7135, 2024. doi: 10.24963/ijcai.2024/788. URL https://www.ijcai.org/proceedings/2024/0788.

Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, Eric Moulines, and Maxim Panov. From risk to uncertainty: Generating predictive uncertainty measures via bayesian estimation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cWfpt2t37q.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor I Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=eGLdVRvvfQ. Expert Certification.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017. URL https://papers.neurips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.

Simon Lang, Martin Leutbecher, and Pedro Maciel. A multi-scale loss formulation for learning a probabilistic model with proper score optimisation. *arXiv preprint arXiv:2506.10868*, 2025. doi: 10.48550/arXiv.2506.10868. URL https://arxiv.org/abs/2506.10868.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL http://yann.lecun.com/exdb/publis/pdf/lecun-98.pdf.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. *Advances in Neural Information Processing Systems*, 37:50972–51038, 2024. doi: 10.52202/079017-1614. URL https://papers.nips.cc/paper_files/paper/2024/hash/5afa9cb1e917b898ad418216dc726fbd-Abstract-Datasets_and_Benchmarks_Track.html.

Jishnu Mukhoti, Andreas Kirsch, Joost Van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24384–24394, 2023. doi: 10.1109/CVPR52729.2023.02336. URL https://openaccess.thecvf.com/content/CVPR2023/html/Mukhoti_Deep_Deterministic_Uncertainty_A_New_Simple_Baseline_CVPR_2023_paper.html.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, December 2011. URL https://research.google/pubs/reading-digits-in-natural-images-with-unsupervised-feature-learning/.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994. doi: 10.1109/ICNN.1994.374138. URL https://doi.org/10.1109/ICNN.1994.374138.

Prashant Rana. Physicochemical Properties of Protein Tertiary Structure. UCI Machine Learning Repository, 2013. URL https://archive.ics.uci.edu/dataset/265/physicochemical+properties+of+protein+tertiary+structure.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-order uncertainty quantification: A distance-based approach. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43060–43076. PMLR, 2024. URL https://proceedings.mlr.press/v235/sale24a.html.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971. doi: 10.1080/01621459.1971.10482346. URL https://doi.org/10.1080/01621459.1971.10482346.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, and Sepp Hochreiter. On information-theoretic measures of predictive uncertainty. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 3605–3640. PMLR, 2025. URL https://proceedings.mlr.press/v286/schweighofer25a.html.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aPOpXlnV1T.

Nicki Skafte, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 32, 2019. URL https://papers.nips.cc/paper/8862-reliable-training-and-estimation-of-variance-networks.

Sanbao Su, Yiming Li, Sihong He, Songyang Han, Chen Feng, Caiwen Ding, and Fei Miao. Uncertainty quantification of collaborative detection for self-driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5588–5594. IEEE, 2023. doi: 10.1109/ICRA48891.2023.10160367. URL https://doi.org/10.1109/ICRA48891.2023.10160367.

Pınar Tüfekci. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126–140, 2014. doi: 10.1016/j.ijepes.2014.02.027. URL https://doi.org/10.1016/j.ijepes.2014.02.027.

Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1508–1516. IEEE, 2022. doi: 10.1109/CVPRW56347.2022.00157. URL https://openaccess.thecvf.com/content/CVPR2022W/LXCV/html/Valdenegro-Toro_A_Deeper_Look_Into_Aleatoric_and_Epistemic_Uncertainty_Disentanglement_CVPRW_2022_paper.html.

Kris Villez. Analytical expressions to compute the continuous ranked probability score (crps). Technical Report TR-004-01-0, Swiss Federal Institute of Aquatic Science and Technology, 2018. URL https://krisvillez.gitlab.io/profile/pdf/KV_T004_v1.0.pdf.

Kartik Waghmare and Johanna Ziegel. Proper scoring rules for estimation and forecast evaluation. *Annual Review of Statistics and Its Application*, 13:271–296, 2026. doi: 10.1146/annurev-statistics-042424-050626. URL https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-042424-050626.

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 *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2282–2292. PMLR, 2023. URL https://proceedings.mlr.press/v216/wimmer23a.html.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. doi: 10.48550/arXiv.1708.07747. URL https://arxiv.org/abs/1708.07747.

A Risk Estimation for Proper Scoring Rules

This appendix develops the general machinery for estimating pointwise risk under proper scoring rules in regression. It formalizes Bayes/total/excess risks and derives tractable approximation for CRPS, logarithmic, quadratic, and squared error scores under a Gaussian assumption.

A.1 Bayesian Risk Estimation

This section specifies Bayesian estimators for the three risk components, showing how to combine posterior averaging, posterior-predictive (mixture) plug-ins, and Gaussian surrogates to obtain computable expressions.

A.1.1 Gaussian Ensemble Model

We will consider a specific tractable example where our Bayesian model is a finite Gaussian ensemble. This model arises e.g. when we train multiple neural network copies $\{\widehat{P}_i\}_{i=1}^M$ that predict Gaussian parameters starting from different weights initializations and is sometimes called *deep ensembles* (Lakshminarayanan et al., 2017):

$$\widehat{P}_i = \mathcal{N}(\mu_i, \sigma_i^2), \quad \mu_i \in \mathbb{R}, \sigma_i \in \mathbb{R}_+, i = 1, \ldots, M.$$

We can interpret this as a finite i.i.d. sample drawn from a posterior distribution over parameters of a one-dimensional Gaussian. Our prediction is the posterior predictive, which equals the full Gaussian mixture distribution:

$$\widehat{P}_{\text{ens}} = \frac{1}{M}\sum_{i=1}^{M} \mathcal{N}\left(\mu_i, \sigma_i^2\right), \quad \widehat{p}_{\text{ens}}(y) = \frac{1}{M}\sum_{i=1}^{M} \widehat{p}_i(y).$$

A.1.2 Ensemble Approximations with a Single Gaussian

Previous work on ensembles (Lakshminarayanan et al., 2017) has used the following Gaussian approximation to the full mixture distribution:

$$\widehat{P}_* = \mathcal{N}(\mu_*, \sigma_*^2), \quad \mu_* = \frac{1}{M}\sum_{i=1}^{M} \mu_i, \quad \sigma_*^2 = \frac{1}{M}\sum_{i=1}^{M} (\sigma_i^2 + \mu_i^2) - \mu_*^2, \tag{3}$$

where μ_* and σ_*^2 are the mean and variance of the full mixture distribution. Another, simpler Gaussian approximation is to set the variance equal to the average variance of the components (Bülte et al., 2025):

$$\bar{P}_* = \mathcal{N}(\mu_*, \bar{\sigma}_*^2), \quad \mu_* = \frac{1}{M}\sum_{i=1}^{M} \mu_i, \quad \bar{\sigma}_*^2 = \frac{1}{M}\sum_{i=1}^{M} \sigma_i^2. \tag{4}$$

A.2 Tractable Risk Estimation Using Gaussian Ensembles

We assume that we only have access to a single Gaussian ensemble model and propose to use in to approximate both $\widehat{P}$ and P. Based on the results of the previous section, we arrive at the following four approximations: (1) Bayesian risk averaging, (2) full mixture distribution $\widehat{P}_{\text{ens}}$, and (3) $\widehat{P}_*$. A specific risk estimate can thus be denoted as $\widehat{R}^{i,j}$, $i, j = 1, \ldots, 3$. These cases roughly correspond to the ones for classification that appeared in (Kotelevskii et al., 2025). We recreate the summary table from the main section here in Table 9 for clarity. In the table empty cells correspond to combinations that do not seem particularly useful (approximation for ground truth should be at least as complex as for the prediction). In what follows, we will provide expressions for these estimates for different proper scoring rules.

A.2.1 CRPS Score

The notation and derivations of various quantities related to CRPS for the ensemble model can be found in the Appendix B.1. In particular, we will use the following notation:

$$A(\mu, \sigma) := 2\sigma\phi\left(\frac{\mu}{\sigma}\right) + \mu\left[2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right], \quad \mu_{ij} := \mu_i - \mu_j, \quad \sigma_{ij} := \sqrt{\sigma_i^2 + \sigma_j^2}.$$

Table 9: Chosen combinations of estimates for excess risk estimation.

Predicted $\widehat{P}$	True P		
	Bayesian averaging	$\widehat{P}_{\text{ens}}$	$\widehat{P}_*$
Bayesian averaging	$\widehat{R}^{1,1}$	–	–
$\widehat{P}_{\text{ens}}$	$\widehat{R}^{2,1}$	–	–
$\widehat{P}_*$	$\widehat{R}^{3,1}$	$\widehat{R}^{3,2}$	–

Bayes risk. Since Bayes risk is equal to the entropy of the true distribution $H(P)$, we only need to choose an approximation scheme for P:

- **Bayesian averaging**: $\widehat{R}^1_{\text{Bayes}} = \frac{1}{M}\sum_{i=1}^M H(\widehat{P}_i) = \frac{1}{M\sqrt{\pi}}\sum_{i=1}^M \sigma_i$.
- **Posterior predictive**: $\widehat{R}^2_{\text{Bayes}} = H(\widehat{P}_{\text{ens}}) = \frac{1}{2}\frac{1}{M^2}\sum_{i=1}^M\sum_{j=1}^M A(\mu_{ij},\sigma_{ij})$.
- **Gaussian approximation**: $\widehat{R}^3_{\text{Bayes}} = H(\widehat{P}_*) = \sqrt{\frac{\sigma_*^2}{\pi}}$.

Excess risk. Since the divergence function for CRPS is symmetric, many of $\widehat{R}^{i,j}_{\text{Exc}}$ quantities coincide. Here we present all possible combinations of approximations for the Excess risk in the Gaussian ensemble model.

- **Bayesian averaging for both**. Also known as *expected pairwise Bregman divergence*.

$$\widehat{R}^{1,1}_{\text{Exc}} = \frac{1}{M^2}\sum_{i=1}^M\sum_{j=1}^M d(\widehat{P}_i,\widehat{P}_j) = \frac{1}{M^2}\sum_{i=1}^M\sum_{j=1}^M\left[A(\mu_{ij},\sigma_{ij}) - \frac{\sigma_i+\sigma_j}{\sqrt{\pi}}\right] \tag{5}$$

$$= \frac{1}{M^2}\sum_{i=1}^M\sum_{j=1}^M A(\mu_{ij},\sigma_{ij}) - \frac{2}{M}\sum_{i=1}^M\frac{\sigma_i}{\sqrt{\pi}}. \tag{6}$$

- **Gaussian mixture and Bayesian averaging**. This approximation is also referred to as *Bregman information* (and *reverse Bregman information* for non-symmetric divergences).

$$\widehat{R}^{2,1}_{\text{Exc}} = \widehat{R}^{1,2}_{\text{Exc}} = \frac{1}{M}\sum_{i=1}^M d(\widehat{P}_i,\widehat{P}_{\text{ens}}) = \frac{1}{M}\sum_{i=1}^M\Big[\text{CRPS}(\widehat{P}_i,\widehat{P}_{\text{ens}}) - H(\widehat{P}_{\text{ens}})\Big] \tag{7}$$

$$= \frac{1}{2}\frac{1}{M^2}\sum_{i=1}^M\sum_{j=1}^M A(\mu_{ij},\sigma_{ij}) - \frac{1}{M}\sum_{i=1}^M\frac{\sigma_i}{\sqrt{\pi}}. \tag{8}$$

- **Moment-matched Gaussian approximation and Bayesian averaging**:

$$\widehat{R}^{3,1}_{\text{Exc}} = \widehat{R}^{1,3}_{\text{Exc}} = \frac{1}{M}\sum_{j=1}^M d(\widehat{P}_*,\widehat{P}_j) = \frac{1}{M}\sum_{j=1}^M\left[A(\mu_{*j},\sigma_{*j}) - \frac{\sigma_*+\sigma_j}{\sqrt{\pi}}\right] \tag{9}$$

$$= \frac{1}{M}\sum_{j=1}^M A(\mu_{*j},\sigma_{*j}) - \frac{\sigma_*}{\sqrt{\pi}} - \frac{\frac{1}{M}\sum_{j=1}^M\sigma_j}{\sqrt{\pi}}. \tag{10}$$

- **Moment-matched Gaussian approximation and mixture**:

$$\widehat{R}^{3,2}_{\text{Exc}} = \widehat{R}^{2,3}_{\text{Exc}} = d(\widehat{P}_*,\widehat{P}_{\text{ens}}) = \text{CRPS}(\widehat{P}_*,\widehat{P}_{\text{ens}}) - H(\widehat{P}_{\text{ens}}) \tag{11}$$

$$= \frac{1}{M}\sum_{j=1}^M A(\mu_{*j},\sigma_{*j}) - \frac{\sigma_*}{\sqrt{\pi}} - H(\widehat{P}_{\text{ens}}). \tag{12}$$

We notice a peculiar relation: $2\widehat{R}^{2,1}_{\text{Exc}} = \widehat{R}^{1,1}_{\text{Exc}}$. Similar results were obtained for a score with symmetric divergence in (Kotelevskii et al., 2025) in the case of classification.

A.2.2 Logarithmic Score

The logarithmic score is a proper scoring rule for both discrete and continuous distributions, which is defined as follows:

$$\text{LS}(\widehat{P}, y) = -\log \widehat{p}(y), \tag{13}$$

where $\widehat{p}(y)$ is the probability density function of the distribution $\widehat{P}$ evaluated at y. This is an example of a *local scoring rule*: it depends only on the density value at the point y and does not take into account the entire distribution.

The entropy function for the logarithmic score is the negative Shannon entropy:

$$H(P) = \text{LS}(P, P) = \int_{\mathbb{R}} \text{LS}(P, y)\, p(y)\, dy = -\int_{\mathbb{R}} p(y) \log p(y)\, dy. \tag{14}$$

The divergence function of the logarithmic score is the well-known Kullback-Leibler divergence:

$$d(\widehat{P}, P) = \int_{\mathbb{R}} \big[-\log \widehat{p}(y)\big]\, p(y)\, dy - \int_{\mathbb{R}} \big[-\log p(y)\big]\, p(y)\, dy = \int_{\mathbb{R}} p(y) \log \frac{p(y)}{\widehat{p}(y)}\, dy = D_{\text{KL}}(P \,\|\, \widehat{P}). \tag{15}$$

KL-divergence between two Gaussian mixtures can not be expressed in a closed form, but it can be approximated with MC methods or bounded; see (Hershey & Olsen, 2007; Durrieu et al., 2012). In particular, the work (Durrieu et al., 2012) provides lower and upper bounds for the KL divergence between two Gaussian mixtures using variational approximations.

Detailed derivations of the quantities presented in the following are deferred to Section B.2.

Bayes risk. For the Gaussian ensemble model, we can derive the following approximations of the Bayes risk for the log score:

- $\widehat{R}^{1}_{\text{Bayes}} = \frac{1}{M}\sum_{i=1}^{M} H(\widehat{P}_i) = \frac{1}{2M}\sum_{i=1}^{M} \log\left(2\pi e \sigma_i^2\right)$.
- $\widehat{R}^{2}_{\text{Bayes}} = H(\widehat{P}_{\text{ens}}) = -\int_{\mathbb{R}} \log\left(\frac{1}{M}\sum_{i=1}^{M} \widehat{p}_i(y)\right) \cdot \left(\frac{1}{M}\sum_{i=1}^{M} \widehat{p}_i(y)\right) dy$ – this expression is known to have no closed form. Using Jensen's inequality, we can derive a lower bound:

$$\widehat{R}^{2}_{\text{Bayes}} \geq \frac{1}{M}\sum_{i=1}^{M} H(\widehat{P}_i) = \widehat{R}^{1}_{\text{Bayes}}.$$

- $\widehat{R}^{3}_{\text{Bayes}} = H(\widehat{P}_*) = \frac{1}{2}\log\left(2\pi e \sigma_*^2\right)$.

Excess risk. In the case of the Gaussian ensemble, we can use the following approximations:

- **Bayesian averaging for both**:

$$\widehat{R}^{1,1}_{\text{Exc}} = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} d(\widehat{P}_i, \widehat{P}_j) = \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left[\frac{\sigma_j^2 + (\mu_i - \mu_j)^2}{\sigma_i^2} - 1\right]. \tag{16}$$

- **Gaussian mixture and Bayesian averaging**:

$$\widehat{R}^{2,1}_{\text{Exc}} = \frac{1}{M}\sum_{i=1}^{M} d(\widehat{P}_{\text{ens}}, \widehat{P}_i) = \frac{1}{M}\sum_{i=1}^{M} D_{\text{KL}}(\widehat{P}_i \,\|\, \widehat{P}_{\text{ens}}), \tag{17}$$

$$\widehat{R}^{1,2}_{\text{Exc}} = \frac{1}{M}\sum_{i=1}^{M} d(\widehat{P}_i, \widehat{P}_{\text{ens}}) = \frac{1}{M}\sum_{i=1}^{M} D_{\text{KL}}(\widehat{P}_{\text{ens}} \,\|\, \widehat{P}_i). \tag{18}$$

There is no nice analytical form for those since D_{KL} contains the log of the mixture density. The expression for $\widehat{R}^{2,1}_{\mathrm{Exc}}$ was used for discrete distributions in (Lakshminarayanan et al., 2017) as "disagreement".

- **Moment-matched Gaussian approximation and Bayesian averaging**:

$$\widehat{R}^{3,1}_{\mathrm{Exc}} = \frac{1}{M}\sum_{i=1}^{M} d(\widehat{P}_*, \widehat{P}_i) = \frac{1}{2}\left[\log\left(\sigma_*^2\right) - \frac{1}{M}\sum_{i=1}^{M}\log\left(\sigma_i^2\right)\right]. \tag{19}$$

Another, simpler approximation of the mixture is $\bar{P}_* = \mathcal{N}(\mu_*, \bar{\sigma}_*^2)$, where $\bar{\sigma}_*^2 = \frac{1}{M}\sum_{i=1}^{M}\sigma_i^2$. With this, we recover the approximation from (Bülte et al., 2025):

$$\bar{R}^{3,1}_{\mathrm{Exc}} = \frac{1}{2}\left[\log\left(\bar{\sigma}_*^2\right) - \frac{1}{M}\sum_{i=1}^{M}\log\left(\sigma_i^2\right) + \frac{1}{\bar{\sigma}_*^2 M}\sum_{i=1}^{M}\left(\mu_i - \mu_*\right)^2\right]. \tag{20}$$

- **Moment-matched Gaussian approximation and mixture**:

$$\widehat{R}^{3,2}_{\mathrm{Exc}} = D_{\mathrm{KL}}(\widehat{P}_{\mathrm{ens}} \,\|\, \widehat{P}_*). \tag{21}$$

A.2.3 Quadratic Score

In our results below, we will use the following notation:

$$\mathcal{N}\left(a \mid b, s^2\right) = \frac{1}{\sqrt{2\pi s^2}}\exp\left(-\frac{(a-b)^2}{2s^2}\right) = \frac{1}{\sqrt{\sigma_i^2+\sigma_j^2}}\,\phi\left(\frac{\mu_i-\mu_j}{\sqrt{\sigma_i^2+\sigma_j^2}}\right),$$

see Appendix B.3 for more details and derivations of the below quantities.

Bayes risk. For the Gaussian ensemble model, we can derive the following approximations of the Bayes risk for the quadratic score:

- $\widehat{R}^1_{\mathrm{Bayes}} = \frac{1}{M}\sum_{i=1}^{M} H(\widehat{P}_i) = -\frac{1}{2M\sqrt{\pi}}\sum_{i=1}^{M}\frac{1}{\sigma_i}$.
- $\widehat{R}^2_{\mathrm{Bayes}} = H(\widehat{P}_{\mathrm{ens}}) = -\int_{\mathbb{R}}\left(\frac{1}{M}\sum_{i=1}^{M}\widehat{p}_i(y)\right)^2 dy = \frac{1}{2M}\sum_{i=1}^{M} H(\widehat{P}_i) + \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\mathrm{QS}(\widehat{P}_i, \widehat{P}_j)$.
  Alternatively, we can express the entropy of the quadratic score for a Gaussian mixture as follows: $\widehat{R}^2_{\mathrm{Bayes}} = -\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\mathcal{N}\left(\mu_i \mid \mu_j, \sigma_i^2+\sigma_j^2\right)$.
- $\widehat{R}^3_{\mathrm{Bayes}} = H(\widehat{P}_*) = -\frac{1}{2\sqrt{\pi\sigma_*^2}}$.

Excess risk. The divergence is symmetric for the quadratic score, so the situation is analogous to the case of CRPS. We have two different approximations of Excess risk:

- **Expected Pairwise Bregman Divergence (EPBD)**: if we employ Bayesian approach for estimation both P and $\widehat{P}$, we get the following approximation: $\widehat{R}^{1,1}_{\mathrm{Exc}} = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} d(\widehat{P}_i, \widehat{P}_j)$, where closed form of $d(\widehat{P}_i, \widehat{P}_j)$ is given by (see equation (70)):

$$d(\widehat{P}_i, \widehat{P}_j) = \frac{1}{2\sqrt{\pi}\sigma_i} + \frac{1}{2\sqrt{\pi}\sigma_j} - 2\mathcal{N}\left(\mu_i \mid \mu_j, \sigma_i^2+\sigma_j^2\right).$$

We combine these results to get the final expression:

$$\widehat{R}^{1,1}_{\mathrm{Exc}} = \underbrace{\frac{1}{M\sqrt{\pi}}\sum_{i=1}^{M}\frac{1}{\sigma_i}}_{-2\widehat{R}^1_{\mathrm{Bayes}}} - \frac{2}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\mathcal{N}\left(\mu_i \mid \mu_j, \sigma_i^2+\sigma_j^2\right).$$

Notice that the first part equal $-2\widehat{R}^1_{\mathrm{Bayes}}$.

- **Bregman Information (BI)**: here we combine Bayesian averaging and the mixture distribution. Using equation (74) for the divergence between the mixture and its component we get:

$$\begin{aligned}\widehat{R}^{2,1}_{\text{Exc}} = \widehat{R}^{1,2}_{\text{Exc}} &= \frac{1}{M}\sum_{i=1}^{M} d(\widehat{P}_i, \widehat{P}_{\text{ens}}) \\ &= \frac{1}{M}\sum_{i=1}^{M}\left[-\frac{2}{M}\left(\sum_{j=1}^{M}\mathcal{N}\big(\mu_i \mid \mu_j, \sigma_i^2+\sigma_j^2\big)\right) - H(\widehat{P}_i) - H(\widehat{P}_{\text{ens}})\right] \\ &= \underbrace{\frac{1}{2M\sqrt{\pi}}\sum_{i=1}^{M}\frac{1}{\sigma_i}}_{-\widehat{R}^1_{\text{Bayes}}} - \frac{2}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\mathcal{N}\left(\mu_i \mid \mu_j, \sigma_i^2+\sigma_j^2\right) - \widehat{R}^2_{\text{Bayes}} = \frac{1}{2}\widehat{R}^{1,1}_{\text{Exc}}.\end{aligned}$$

- **Moment-matched Gaussian approximation and Bayesian averaging**:

$$\begin{aligned}\widehat{R}^{3,1}_{\text{Exc}} = \widehat{R}^{1,3}_{\text{Exc}} &= \frac{1}{M}\sum_{i=1}^{M} d(\widehat{P}_*, \widehat{P}_i) \\ &= \frac{1}{M}\sum_{i=1}^{M}\left[-H(\widehat{P}_i) - H(\widehat{P}_*) - 2\mathcal{N}\left(\mu_* \mid \mu_i, \sigma_*^2+\sigma_i^2\right)\right] \\ &= \underbrace{\frac{1}{2M\sqrt{\pi}}\sum_{i=1}^{M}\frac{1}{\sigma_i}}_{-\widehat{R}^1_{\text{Bayes}}} - \frac{2}{M}\sum_{i=1}^{M}\mathcal{N}\left(\mu_* \mid \mu_i, \sigma_*^2+\sigma_i^2\right) + \underbrace{\frac{1}{2\sqrt{\pi\sigma_*^2}}}_{-\widehat{R}^3_{\text{Bayes}}}.\end{aligned}$$

- **Moment-matched Gaussian approximation and mixture**:

$$\widehat{R}^{3,2}_{\text{Exc}} = \widehat{R}^{2,3}_{\text{Exc}} = d(\widehat{P}_*, \widehat{P}_{\text{ens}}) = -\frac{2}{M}\left(\sum_{j=1}^{M}\mathcal{N}\big(\mu_* \mid \mu_j, \sigma_*^2+\sigma_j^2\big)\right) - \underbrace{H(\widehat{P}_*)}_{\widehat{R}^3_{\text{Bayes}}} - \underbrace{H(\widehat{P}_{\text{ens}})}_{\widehat{R}^2_{\text{Bayes}}}.$$

A.2.4 SE Score

We can use the same three approaches to estimate the risk components for the SE score as we did for other proper scoring rules. Due to the symmetry of the divergence, the central label and central prediction coincide again. They are equal to the full mixture distribution, which we again denote as $\widehat{P}_{\text{ens}}$. The mean and variance of this distribution are denoted as μ_* and σ_*^2. We have previously introduced them to construct our Gaussian approximation in equation (3). Here they are again for clarity:

$$\mathbb{E}\big[\widehat{P}_{\text{ens}}\big] = \mu_* = \frac{1}{M}\sum_{i=1}^{M}\mu_i,\ \operatorname{Var}\big[\widehat{P}_{\text{ens}}\big] = \sigma_*^2 = \frac{1}{M}\sum_{i=1}^{M}(\sigma_i^2+\mu_i^2) - \mu_*^2.$$

For a more detailed treatment of the SE score, please refer to Section C.

Bayes risk.

- **Bayesian averaging**: $\widehat{R}^1_{\text{Bayes}} = \frac{1}{M}\sum_{i=1}^{M} H_{\text{SE}}(\widehat{P}_i) = \frac{1}{M}\sum_{i=1}^{M}\operatorname{Var}_{Y\sim\widehat{P}_i}[Y] = \frac{1}{M}\sum_{i=1}^{M}\sigma_i^2$.
- **Posterior predictive**: $\widehat{R}^2_{\text{Bayes}} = H_{\text{SE}}(\widehat{P}_{\text{ens}}) = \operatorname{Var}_{Y\sim\widehat{P}_{\text{ens}}}[Y] = \sigma_*^2 = \frac{1}{M}\sum_{i=1}^{M}(\sigma_i^2+\mu_i^2) - \mu_*^2$.
- **Gaussian approximation**: $\widehat{R}^3_{\text{Bayes}} = H_{\text{SE}}(\widehat{P}_*) = \operatorname{Var}_{Y\sim\widehat{P}_*}[Y] = \sigma_*^2 = \widehat{R}^2_{\text{Bayes}}$.

Excess risk. For the symmetric divergence of SE we get:

- **Bayesian averaging**: $\widehat{R}^{1,1}_{\text{Exc}} = \frac{1}{M^2}\sum_{i=1}^M \sum_{j=1}^M d_{\text{SE}}(\widehat{P}_i, \widehat{P}_j) = \frac{1}{M^2}\sum_{i=1}^M \sum_{j=1}^M (\mu_j - \mu_i)^2 = 2\widehat{\text{Var}}\left[\mu_i\right]$.
- **Posterior predictive and Bayesian averaging**: $\widehat{R}^{2,1}_{\text{Exc}} = \widehat{R}^{1,2}_{\text{Exc}} = \frac{1}{M}\sum_{i=1}^M d_{\text{SE}}(\widehat{P}_i, \widehat{P}_{\text{ens}}) = \frac{1}{M}\sum_{i=1}^M (\mathbb{E}_{X\sim\widehat{P}_i}[X] - \mathbb{E}_{Y\sim\widehat{P}_{\text{ens}}}[Y])^2 = = \frac{1}{M}\sum_{i=1}^M (\mu_i - \mu_*)^2 = \widehat{\text{Var}}\left[\mu_i\right]$.
- **Moment-matched Gaussian approximation and Bayesian averaging**: since SE score only looks at the mean of the distribution, using this Gaussian approximation will give the same result as the previous section.
$$\widehat{R}^{3,1}_{\text{Exc}} = \widehat{R}^{1,3}_{\text{Exc}} = \frac{1}{M}\sum_{i=1}^M (\mu_i - \mu_*)^2 = \widehat{\text{Var}}\left[\mu_i\right].$$
- **Moment-matched Gaussian approximation and posterior predictive**: the means of these distributions coincide so we get a zero.
$$\widehat{R}^{3,2}_{\text{Exc}} = \widehat{R}^{2,3}_{\text{Exc}} = 0.$$

B Score Computation and Tools

B.1 CRPS

Continuous ranked probability score has multiple equivalent representations:

$$\text{CRPS}(P, y) = \int_{\mathbb{R}} \big(F_P(t) - \mathbb{I}\{y \le t\}\big)^2 dt. \tag{22}$$

$$\text{CRPS}(P, y) = \mathbb{E}_{X\sim P}\big|X - y\big| - \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X - X'\big|. \tag{23}$$

$$\text{CRPS}(P, y) = \int |x - y|\, dP(x) - \frac{1}{2}\int\int |x - x'|\, dP(x)\, dP(x'). \tag{24}$$

Expected CRPS between two distributions P and Q can be expressed as:

$$\text{CRPS}(P, Q) = \mathbb{E}_{X\sim P, Y\sim Q}\big|X - Y\big| - \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X - X'\big|. \tag{25}$$

B.1.1 Entropy Function of CRPS

$$\begin{aligned} H(P) &= \int \text{CRPS}(P, y)\, dP(y) = \int \left[\mathbb{E}_{X\sim P}\big|X - y\big| - \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X - X'\big|\right] dP(y) \\ &= \int \mathbb{E}_{X\sim P}\big|X - y\big|\, dP(y) - \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X - X'\big| = \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X - X'\big|. \end{aligned}$$

B.1.2 Divergence Function of CRPS

$$\begin{aligned} d(P, Q) &= \int \text{CRPS}(P, y)\, dQ(y) - H(Q) \\ &= \int \left[\mathbb{E}_{X\sim P}\big|X - y\big| - \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X - X'\big|\right] dQ(y) - H(Q) \\ &= \mathbb{E}_{X\sim P, Y\sim Q}\big|X - Y\big| - \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X - X'\big| - \frac{1}{2}\mathbb{E}_{Y,Y'\sim Q}\big|Y - Y'\big| \\ &= \mathbb{E}_{X\sim P, Y\sim Q}\big|X - Y\big| - H(P) - H(Q). \end{aligned}$$

Another expression for the divergence function of CRPS is given by:

$$d(P, Q) = \int_{\mathbb{R}} \big(F_P(y) - F_Q(y)\big)^2 dy. \tag{26}$$

B.1.3 CRPS for Gaussians

Notation. Some commonly used notation:

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \text{ is the standard normal density function.} \tag{27}$$

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \text{ is the standard normal cumulative distribution function.} \tag{28}$$

We follow (Villez, 2018) and also introduce the following notation:

$$\mu_{ij} := \mu_i - \mu_j, \tag{29}$$

$$\sigma_{ij} := \sqrt{\sigma_i^2 + \sigma_j^2}, \tag{30}$$

$$A(\mu, \sigma) := 2\sigma\phi\left(\frac{\mu}{\sigma}\right) + \mu\left[2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right]. \tag{31}$$

We note that for $X \sim \mathcal{N}(\mu, \sigma)$ it can be shown that $\mathbb{E}\left[|X|\right] = A(\mu, \sigma)$.

CRPS for a Gaussian.

$$\mathrm{CRPS}\big(\mathcal{N}(\mu, \sigma), y\big) = \sigma\left(2\phi(z) + z\big(2\Phi(z) - 1\big) - \frac{1}{\sqrt{\pi}}\right), \tag{32}$$

where $z = \frac{y-\mu}{\sigma}$, $\phi(z)$ is PDF of the standard normal distribution, and $\Phi(z)$ is CDF of the standard normal distribution

Proof.

$$\begin{aligned}
\mathrm{CRPS}(\mathcal{N}(\mu, \sigma), y) &= \int_{-\infty}^{+\infty} (F_{\mathcal{N}}(t) - \mathbb{I}\{y \le t\})^2 dt \\
&= \int_{-\infty}^{y} (F_{\mathcal{N}}(t) - \mathbb{I}\{y \le t\})^2 dt + \int_{y}^{+\infty} (F_{\mathcal{N}}(t) - \mathbb{I}\{y \le t\})^2 dt \\
&= \int_{-\infty}^{y} (F_{\mathcal{N}}(t) - 0)^2 dt + \int_{y}^{+\infty} (F_{\mathcal{N}}(t) - 1)^2 dt \\
&= \int_{-\infty}^{y} F_{\mathcal{N}}^2(t) dt + \int_{y}^{+\infty} \big(1 - F_{\mathcal{N}}(t)\big)^2 dt = I_1 + I_2.
\end{aligned}$$

First we compute

$$I_1 = \int_{-\infty}^{y} F_{\mathcal{N}}^2(t) dt = \int_{-\infty}^{y} \Phi^2\left(\frac{t-\mu}{\sigma}\right) dt = \sigma \int_{-\infty}^{\frac{y-\mu}{\sigma}} \Phi^2(z) dz = \sigma \int_{-\infty}^{z_y} \Phi^2(z) dz,$$

where we used the substitution $z = \frac{t-\mu}{\sigma}$ and define $z_y = \frac{y-\mu}{\sigma}$. For the second part, we get

$$\begin{aligned}
I_2 &= \int_{y}^{+\infty} \big(1 - F_{\mathcal{N}}(t)\big)^2 dt = \int_{y}^{+\infty} \left(1 - \Phi\left(\frac{t-\mu}{\sigma}\right)\right)^2 dt \\
&= \sigma \int_{\frac{y-\mu}{\sigma}}^{+\infty} \big(1 - \Phi(z)\big)^2 dz = \sigma \int_{z_y}^{+\infty} \big(1 - \Phi(z)\big)^2 dz.
\end{aligned}$$

Let us calculate the first part:

$$\begin{aligned}
I_1 &= \sigma \int_{-\infty}^{z_y} \Phi^2(z) dz = \begin{bmatrix} u = \Phi(z)^2 & dv = dz \\ du = 2\Phi(z)\phi(z)dz & v = z \end{bmatrix} = \sigma\left[z\Phi^2(z)|_{-\infty}^{z_y} - \int_{-\infty}^{z_y} 2z\Phi(z)\phi(z)dz\right] \\
&= \sigma\left[z_y\Phi^2(z_y) - 2\int_{-\infty}^{z_y} z\Phi(z)\phi(z)dz\right] = \sigma\left[z_y\Phi^2(z_y) + 2\int_{-\infty}^{z_y} z\Phi(z)\, d\phi(z)\right],
\end{aligned}$$

where we have used the fact that $\big(\phi(z)\big)'_z = -z\phi(z)$.

Next,

$$\begin{aligned} I_2 &= \sigma \int_{z_y}^{+\infty} \big(1-\Phi(z)\big)^2 dz = \begin{bmatrix} u = \big(1-\Phi(z)\big)^2 & dv = dz \\ du = 2\big(\Phi(z)-1\big)\phi(z)dz & v = z \end{bmatrix} \\ &= \sigma \left[z\big(1-\Phi(z)\big)^2 |_{z_y}^{+\infty} - 2\int_{z_y}^{+\infty} z\big(\Phi(z)-1\big)\phi(z)dz \right] \\ &= \sigma \left[-z_y\big(1-\Phi(z_y)\big)^2 + 2\int_{z_y}^{+\infty} \big(\Phi(z)-1\big)d\phi(z) \right]. \end{aligned} \tag{33}$$

Finally,

$$\begin{aligned} \mathrm{CRPS}(\mathcal{N}(\mu,\sigma),y) = I_1 + I_2 &= \sigma \left[z_y\left(2\Phi(z_y)-1\right) + 2\int_{-\infty}^{+\infty} \Phi(z)\, d\phi(z) - 2\int_{z_y}^{+\infty} d\phi(z) \right] \\ &= \sigma \left[z_y\left(2\Phi(z_y)-1\right) + 2I_3 + 2\phi(z_y) \right] \\ &= \sigma \left[z_y\left(2\Phi(z_y)-1\right) - \frac{1}{\sqrt{\pi}} + 2\phi(z_y) \right]. \end{aligned} \tag{34}$$

We used the following identity:

$$I_3 = \int_{-\infty}^{+\infty} \Phi(z)\, d\phi(z) = \Phi(z)\phi(z)\big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \phi^2(z)dz = -\frac{1}{2\sqrt{\pi}}.$$

□

Entropy of CRPS for a Gaussian. If we use the alternative representation of CRPS, we can express it as the expected absolute value of a centered Gaussian:

$$\begin{aligned} X, X' \sim \mathcal{N}(\mu,\sigma^2) \implies X - X' \sim \mathcal{N}(0, 2\sigma^2) \\ \implies \mathbb{E}_{X,X'\sim P}\big|X-X'\big| = A(0,\sqrt{2}\sigma) = \sqrt{2}\sigma\sqrt{\frac{2}{\pi}} = \frac{2\sigma}{\sqrt{\pi}}. \end{aligned} \tag{35}$$

$$H(\mathcal{N}(\mu,\sigma^2)) = \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X-X'\big| = \frac{1}{2}\frac{2\sigma}{\sqrt{\pi}} = \frac{\sigma}{\sqrt{\pi}}. \tag{36}$$

Divergence function of CRPS for two Gaussians. Let $P = \mathcal{N}(\mu_i, \sigma_i^2)$ and $Q = \mathcal{N}(\mu_j, \sigma_j^2)$ be two Gaussian distributions. Recall that the divergence function of CRPS is given by:

$$d(P,Q) = \mathbb{E}_{X\sim P, Y\sim Q}\big|X-Y\big| - \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X-X'\big| - \frac{1}{2}\mathbb{E}_{Y,Y'\sim Q}\big|Y-Y'\big|. \tag{37}$$

Since X and Y are independent, $X - Y \sim \mathcal{N}(\mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)$. Using our notation, we get:

$$\mathbb{E}_{X\sim P, Y\sim Q}\big|X-Y\big| = A(\mu_{ij}, \sigma_{ij}). \tag{38}$$

We also note that $H(P) = \frac{\sigma_i}{\sqrt{\pi}}$ and $H(Q) = \frac{\sigma_j}{\sqrt{\pi}}$. Combining all these results, we can express the divergence function as follows:

$$d(P,Q) = A(\mu_{ij}, \sigma_{ij}) - \frac{\sigma_i + \sigma_j}{\sqrt{\pi}}. \tag{39}$$

Expected CRPS between two Gaussians.

$$\mathrm{CRPS}(P,Q) = \mathbb{E}_{X\sim P, Y\sim Q}\big|X-Y\big| - \frac{1}{2}\mathbb{E}_{X,X'\sim P}\big|X-X'\big|. \tag{40}$$

Let $P = \mathcal{N}(\mu_i, \sigma_i^2)$ and $Q = \mathcal{N}(\mu_j, \sigma_j^2)$ be two Gaussian distributions. Then the expected CRPS between these two distributions can be expressed as:

$$\mathrm{CRPS}\big(\mathcal{N}(\mu_i,\sigma_i), \mathcal{N}(\mu_j,\sigma_j)\big) = A(\mu_{ij},\sigma_{ij}) - \frac{\sigma_i}{\sqrt{\pi}}. \tag{41}$$

Entropy of CRPS for Gaussian mixture.

$$H(\widehat{P}_{\mathrm{ens}}) = \frac{1}{2}\mathbb{E}_{X,X'\sim\widehat{P}_{\mathrm{ens}}}\big|X-X'\big| = \frac{1}{2}\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\mathbb{E}_{X\sim\widehat{P}_i, X'\sim\widehat{P}_j}\big|X-X'\big| = \frac{1}{2}\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}A(\mu_{ij},\sigma_{ij}). \tag{42}$$

Expected CRPS between a Gaussian and a Gaussian mixture.

$$\mathrm{CRPS}(\widehat{P}_i, \widehat{P}_{\mathrm{ens}}) = \mathbb{E}_{X\sim\widehat{P}_i, Y\sim\widehat{P}_{\mathrm{ens}}}\big|X-Y\big| - \frac{1}{2}\mathbb{E}_{X,X'\sim\widehat{P}_i}\big|X-X'\big| = \frac{1}{M}\sum_{j=1}^{M}A(\mu_{ij},\sigma_{ij}) - \frac{\sigma_i}{\sqrt{\pi}}. \tag{43}$$

Divergence between a Gaussian and a Gaussian mixture.

$$d(\widehat{P}_i, \widehat{P}_{\mathrm{ens}}) = \mathrm{CRPS}(\widehat{P}_i, \widehat{P}_{\mathrm{ens}}) - H(\widehat{P}_{\mathrm{ens}}) = \frac{1}{M}\sum_{j=1}^{M}A(\mu_{ij},\sigma_{ij}) - \frac{\sigma_i}{\sqrt{\pi}} - \frac{1}{2}\frac{1}{M^2}\sum_{l=1}^{M}\sum_{j=1}^{M}A(\mu_{lj},\sigma_{lj}). \tag{44}$$

B.2 Log score

B.2.1 Entropy of Log score

Log score is associated with Shannon entropy:

$$H(P) = \mathbb{E}_{Y\sim P}\left[-\log p(Y)\right] = -\int_{\mathbb{R}} p(y)\log p(y)\,dy. \tag{45}$$

B.2.2 Divergence of Log score

Divergence function is the Kullback-Leibler divergence:

$$d(\widehat{P}, P) = D_{\mathrm{KL}}(P \,\|\, \widehat{P}) = \int_{\mathbb{R}} p(y)\log\frac{p(y)}{\widehat{p}(y)}\,dy. \tag{46}$$

B.2.3 Log score for Gaussians

Log score entropy for a Gaussian. Logarithm of the Gaussian density is a quadratic function, so we can easily express the entropy using the expression for Gaussian:

$$Y\sim P = \mathcal{N}(\mu,\sigma) \implies p(y) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}} \implies \mathrm{LS}(P,y) = -\log p(y) = \frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{(y-\mu)^2}{2\sigma^2}. \tag{47}$$

$$\begin{aligned} H\big(\mathcal{N}(\mu,\sigma)\big) &= \mathbb{E}_{Y\sim\mathcal{N}(\mu,\sigma)}\big[-\log p(Y)\big] = \mathbb{E}_{Y\sim\mathcal{N}(\mu,\sigma)}\left[\frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\mathbb{E}_{Y\sim\mathcal{N}(\mu,\sigma)}\left[(y-\mu)^2\right] = \frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2} = \frac{1}{2}\log\left(2\pi e\sigma^2\right). \end{aligned}$$

Log score divergence for two Gaussians. Here, the derivation is very similar, using the variance/second moment formulas.

$$P = \mathcal{N}(\mu_p, \sigma_p^2),\ Q = \mathcal{N}(\mu_q, \sigma_q^2) \implies p(y) = \frac{1}{\sqrt{2\pi}\sigma_p} e^{-\frac{(y-\mu_p)^2}{2\sigma_p^2}},\ q(y) = \frac{1}{\sqrt{2\pi}\sigma_q} e^{-\frac{(y-\mu_q)^2}{2\sigma_q^2}}. \tag{48}$$

$$\log p(y) = -\frac{1}{2}\log(2\pi\sigma_p^2) - \frac{(y-\mu_p)^2}{2\sigma_p^2},\ \log q(y) = -\frac{1}{2}\log(2\pi\sigma_q^2) - \frac{(y-\mu_q)^2}{2\sigma_q^2}. \tag{49}$$

$$\begin{aligned} d(P,Q) &= \mathbb{E}_{Y\sim Q}\left[-\log p(Y)\right] - H(Q) = D_{\mathrm{KL}}(Q \,\|\, P) \\ &= \mathbb{E}_{Y\sim Q}\left[\log q(Y) - \log p(Y)\right] = \mathbb{E}_{Y\sim Q}\left[\frac{1}{2}\log(2\pi\sigma_p^2) + \frac{(y-\mu_p)^2}{2\sigma_p^2} - \frac{1}{2}\log(2\pi\sigma_q^2) - \frac{(y-\mu_q)^2}{2\sigma_q^2}\right] \\ &= \frac{1}{2}\left[\log\frac{\sigma_p^2}{\sigma_q^2} - 1 + \frac{\sigma_q^2 + (\mu_p-\mu_q)^2}{\sigma_p^2}\right]. \end{aligned}$$

Expected Log score between two Gaussians.

$$\begin{aligned} \mathrm{LS}(P,Q) = d(P,Q) + H(Q) &= \frac{1}{2}\left[\log\frac{\sigma_p^2}{\sigma_q^2} - 1 + \frac{\sigma_q^2 + (\mu_p-\mu_q)^2}{\sigma_p^2}\right] + \frac{1}{2}\log\left(2\pi e\sigma_q^2\right) \\ &= \frac{1}{2}\left[\log\left(2\pi\sigma_p^2\right) + \frac{\sigma_q^2 + (\mu_p-\mu_q)^2}{\sigma_p^2}\right]. \end{aligned} \tag{50}$$

Pairwise divergences of a Gaussian ensemble.

$$\begin{aligned} \widehat{R}^{1,1}_{\mathrm{Exc}} &= \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} d(\widehat{P}_i, \widehat{P}_j) = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} D_{\mathrm{KL}}(\widehat{P}_j, \widehat{P}_i) \\ &= \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{1}{2}\left[\log\frac{\sigma_i^2}{\sigma_j^2} - 1 + \frac{\sigma_j^2 + (\mu_i-\mu_j)^2}{\sigma_i^2}\right] \\ &= \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left[\frac{\sigma_j^2 + (\mu_i-\mu_j)^2}{\sigma_i^2} - 1\right], \end{aligned} \tag{51}$$

because sum of all pairwise differences of logarithms is 0.

Divergence between a Gaussian mixture and one of its components. For the log score the divergence function is KL-divergence, it is not symmetric. This provides two possible ways to approximate the excess risk: assume the true distribution to be the full mixture and average over the components or the other way around. In the notation of (Kotelevskii et al., 2025) they correspond to Bregman Information (BI) and Reverse Bregman Information (RBI).

$$\begin{aligned} d(\widehat{P}_{\mathrm{ens}}, \widehat{P}_i) &= D_{\mathrm{KL}}(\widehat{P}_i \,\|\, \widehat{P}_{\mathrm{ens}}) \\ &= \mathrm{LS}(\widehat{P}_{\mathrm{ens}}, \widehat{P}_i) - H(\widehat{P}_i) = -\int_{\mathbb{R}} \widehat{p}_i(y)\log\big(\widehat{p}_{\mathrm{ens}}(y)\big)\,dy - \frac{1}{2}\log\left(2\pi e\sigma_i^2\right). \end{aligned}$$

$$
\begin{aligned}
d(\widehat{P}_i, \widehat{P}_{\text{ens}}) &= D_{\text{KL}}(\widehat{P}_{\text{ens}} \,\|\, \widehat{P}_i) = \text{LS}(\widehat{P}_i, \widehat{P}_{\text{ens}}) - H(\widehat{P}_{\text{ens}}) \\
&= -\int_{\mathbb{R}} \widehat{p}_{\text{ens}}(y) \log \widehat{p}_i(y)\, dy - H(\widehat{P}_{\text{ens}}) = \frac{1}{M}\sum_{j=1}^{M} \text{LS}(\widehat{P}_i, \widehat{P}_j) - H(\widehat{P}_{\text{ens}}) \\
&= \frac{1}{M}\sum_{j=1}^{M} \text{LS}(\widehat{P}_i, \widehat{P}_j) - H(\widehat{P}_{\text{ens}}).
\end{aligned}
$$

There is no more explicit analytical formula due to appearance of logarithm of the mixture density, which is a sum of the component densities. In the discrete case there are expressions involving the *LogSumExp* function.

Divergence between a Gaussian ensemble and its Gaussian approximation. $\widehat{P}_* = \mathcal{N}(\mu_*, \sigma_*)$, $\mu_* = \frac{1}{M}\sum_{i=1}^{M} \mu_i$, $\sigma_*^2 = \frac{1}{M}\sum_{i=1}^{M}(\sigma_i^2 + \mu_i^2) - \mu_*^2$.

$$
\begin{aligned}
\widehat{R}_{\text{Exc}}^{3a,1} &= \frac{1}{M}\sum_{i=1}^{M} d(\widehat{P}_*, \widehat{P}_i) = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\left[\log\frac{\sigma_*^2}{\sigma_i^2} - 1 + \frac{\sigma_i^2 + (\mu_* - \mu_i)^2}{\sigma_*^2}\right] \\
&= \frac{1}{2}\left[\log\left(\sigma_*^2\right) - \frac{1}{M}\sum_{i=1}^{M}\log\left(\sigma_i^2\right) - 1 + \frac{1}{\sigma_*^2}\cancelto{\sigma_*^2}{\frac{1}{M}\sum_{i=1}^{M}\sigma_i^2 + (\mu_* - \mu_i)^2}\right] \\
&= \frac{1}{2}\left[\log\left(\sigma_*^2\right) - \frac{1}{M}\sum_{i=1}^{M}\log\left(\sigma_i^2\right)\right].
\end{aligned}
$$

Divergence between a Gaussian ensemble and mean Gaussian approximation. $\bar{P}_* = \mathcal{N}(\mu_*, \bar{\sigma}_*^2)$, where $\mu_* = \frac{1}{M}\sum_{i=1}^{M} \mu_i$ and $\bar{\sigma}_*^2 = \frac{1}{M}\sum_{i=1}^{M} \sigma_i^2$.

$$
\begin{aligned}
\widehat{R}_{\text{Exc}}^{3b,1} &= \frac{1}{M}\sum_{i=1}^{M} d(\widehat{P}, \widehat{P}_i) = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{2}\left[\log\frac{\bar{\sigma}_*^2}{\sigma_i^2} - 1 + \frac{\sigma_i^2 + (\mu_* - \mu_i)^2}{\bar{\sigma}_*^2}\right] \\
&= \frac{1}{2}\left[\log\left(\bar{\sigma}_*^2\right) - \frac{1}{M}\sum_{i=1}^{M}\log\left(\sigma_i^2\right) - 1 + \cancelto{1}{\frac{1}{\bar{\sigma}_*^2 M}\sum_{i=1}^{M}\sigma_i^2} + \frac{1}{\bar{\sigma}_*^2 M}\sum_{i=1}^{M}(\mu_* - \mu_i)^2\right] \\
&= \frac{1}{2}\left[\log\left(\bar{\sigma}_*^2\right) - \frac{1}{M}\sum_{i=1}^{M}\log\left(\sigma_i^2\right) + \frac{1}{\bar{\sigma}_*^2 M}\sum_{i=1}^{M}(\mu_* - \mu_i)^2\right].
\end{aligned}
$$

B.3 Quadratic Score

Quadratic score is a generalization of the Brier score for continuous outcomes:

$$
\text{QS}(\widehat{P}, y) = -2\widehat{p}(y) + \int_{\mathbb{R}} \widehat{p}(t)^2 dt. \tag{52}
$$

It can also be viewed as a way to make the value of the probability density at the outcome y into a proper scoring rule. Simply taking $S(\widehat{P}, y) = -\widehat{p}(y)$ is not a proper scoring rule.

B.3.1 Quadratic Score Entropy

$$
\begin{aligned}
H(P) &= \mathbb{E}_{Y\sim P}\big[\text{QS}(P, Y)\big] = -2\mathbb{E}_{Y\sim P}\big[p(Y)\big] + \int_{\mathbb{R}} p(t)^2 dt \\
&= -2\int_{\mathbb{R}} p(y)p(y)dy + \int_{\mathbb{R}} p(t)^2 dt = -\int_{\mathbb{R}} p(y)^2 dy.
\end{aligned} \tag{53}
$$

B.3.2 Quadratic Score Entropy for a Gaussian

$$H\big(\mathcal{N}(\mu,\sigma)\big) = -\int_{\mathbb{R}} p(y)^2 dy = -\int_{\mathbb{R}} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\right)^2 dt = -\frac{1}{2\pi\sigma^2}\int_{\mathbb{R}} e^{-\frac{(t-\mu)^2}{\sigma^2}} dt$$
$$= \begin{bmatrix} u = \frac{t-\mu}{\sigma} \\ du = \frac{1}{\sigma}dt \end{bmatrix} = -\frac{1}{2\pi\sigma^2}\sigma\int_{\mathbb{R}} e^{-u^2} du = \frac{\sqrt{\pi}}{2\pi\sigma} = -\frac{1}{2\sqrt{\pi}\sigma}.$$

B.3.3 Expected Quadratic Score for a Gaussian Prediction and Gaussian Outcome

$$\mathbb{E}_{Y\sim Q}\big[\mathrm{QS}(P,Y)\big] = -2\mathbb{E}_{Y\sim Q}\big[p(Y)\big] + \int_{\mathbb{R}} p(t)^2 dt = -2\int_{\mathbb{R}} p(y)q(y)dy - H(P). \tag{54}$$

We will focus first on computing the integral. Since P and Q are Gaussian distributions, we can express the integral as follows:

$$\int_{\mathbb{R}} p(t)q(t)dt = \frac{1}{2\pi\sigma_P\sigma_Q}\int_{\mathbb{R}} e^{-\frac{(t-\mu_P)^2}{2\sigma_P^2} - \frac{(t-\mu_Q)^2}{2\sigma_Q^2}} dt = \frac{1}{2\pi\sigma_P\sigma_Q}\int_{\mathbb{R}} e^{-\frac{1}{2}\left(\frac{(t-\mu_P)^2}{\sigma_P^2} + \frac{(t-\mu_Q)^2}{\sigma_Q^2}\right)} dt. \tag{55}$$

Let us simplify the exponent:

$$\frac{(t-\mu_P)^2}{\sigma_P^2} + \frac{(t-\mu_Q)^2}{\sigma_Q^2} = \frac{t^2}{\sigma_P^2} - 2\frac{t\mu_P}{\sigma_P^2} + \frac{\mu_P}{\sigma_P^2} + \frac{t^2}{\sigma_Q^2} - 2\frac{t\mu_Q}{\sigma_Q^2} + \frac{\mu_Q}{\sigma_Q^2}$$
$$= \left(\frac{1}{\sigma_P^2} + \frac{1}{\sigma_Q^2}\right)t^2 - 2\left(\frac{\mu_P}{\sigma_P^2} + \frac{\mu_Q}{\sigma_Q^2}\right)t + \left(\frac{\mu_P^2}{\sigma_P^2} + \frac{\mu_Q^2}{\sigma_Q^2}\right) = At^2 - 2Bt + C = A\left(t - \frac{B}{A}\right)^2 - \frac{B^2}{A} + C, \tag{56}$$

where we denote:

$$A = \frac{1}{\sigma_P^2} + \frac{1}{\sigma_Q^2},\ B = \frac{\mu_P}{\sigma_P^2} + \frac{\mu_Q}{\sigma_Q^2},\ C = \frac{\mu_P^2}{\sigma_P^2} + \frac{\mu_Q^2}{\sigma_Q^2}. \tag{57}$$

Now we have:

$$\int_{\mathbb{R}} e^{-\frac{1}{2}\left(At^2 - 2Bt + C\right)} dt = e^{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)}\int_{\mathbb{R}} e^{-\frac{A}{2}\left(t - \frac{B}{A}\right)^2} dt. \tag{58}$$

The leftover integral is a Gaussian integral, which can be computed as follows:

$$\int_{\mathbb{R}} e^{-\frac{A}{2}\left(t - \frac{B}{A}\right)^2} dt = \sqrt{\frac{2\pi}{A}} = \sqrt{2\pi}\frac{\sigma_P\sigma_Q}{\sqrt{\sigma_P^2 + \sigma_Q^2}}. \tag{59}$$

Now, we simplify the exponent of the coefficient in front of that integral:

$$A = \frac{1}{\sigma_P^2} + \frac{1}{\sigma_Q^2} = \frac{\sigma_Q^2 + \sigma_P^2}{\sigma_P^2\sigma_Q^2},\ B^2 = \left(\frac{\mu_P}{\sigma_P^2} + \frac{\mu_Q}{\sigma_Q^2}\right)^2 = \frac{\left(\mu_P\sigma_Q^2 + \mu_Q\sigma_P^2\right)^2}{\sigma_P^4\sigma_Q^4},\ C = \frac{\mu_P^2\sigma_Q^2 + \mu_Q^2\sigma_P^2}{\sigma_P^2\sigma_Q^2}. \tag{60}$$

$$C - \frac{B^2}{A} = \frac{\mu_P^2\sigma_Q^2 + \mu_Q^2\sigma_P^2}{\sigma_P^2\sigma_Q^2} - \frac{\left(\mu_P\sigma_Q^2 + \mu_Q\sigma_P^2\right)^2}{\sigma_P^2\sigma_Q^2\left(\sigma_Q^2 + \sigma_P^2\right)} = \frac{\left(\mu_P^2\sigma_Q^2 + \mu_Q^2\sigma_P^2\right)\left(\sigma_Q^2 + \sigma_P^2\right) - \left(\mu_P\sigma_Q^2 + \mu_Q\sigma_P^2\right)^2}{\sigma_P^2\sigma_Q^2\left(\sigma_Q^2 + \sigma_P^2\right)}. \tag{61}$$

$$\left(\mu_P^2\sigma_Q^2 + \mu_Q^2\sigma_P^2\right)\left(\sigma_Q^2 + \sigma_P^2\right) - \left(\mu_P\sigma_Q^2 + \mu_Q\sigma_P^2\right)^2$$
$$= \cancel{\mu_P^2\sigma_Q^4} + \cancel{\mu_Q^2\sigma_P^4} + \mu_P\sigma_P^2\sigma_Q^2 + \mu_Q\sigma_P^2\sigma_Q^2 - \left(\cancel{\mu_P^2\sigma_Q^4} + 2\mu_P\mu_Q\sigma_P^2\sigma_Q^2 + \cancel{\mu_Q^2\sigma_P^4}\right)$$
$$= \sigma_P^2\sigma_Q^2\left(\mu_P - \mu_Q\right)^2. \tag{62}$$

Thus, we have:

$$C - \frac{B^2}{A} = \frac{\sigma_P^2 \sigma_Q^2 \left(\mu_P - \mu_Q\right)^2}{\sigma_P^2 \sigma_Q^2 \left(\sigma_Q^2 + \sigma_P^2\right)} = \frac{\left(\mu_P - \mu_Q\right)^2}{\sigma_Q^2 + \sigma_P^2}. \tag{63}$$

$$\int_{\mathbb{R}} p(t)\, q(t)\, dt = \frac{1}{2\pi\sigma_P\sigma_Q} \int_{\mathbb{R}} e^{-\frac{(t-\mu_P)^2}{2\sigma_P^2} - \frac{(t-\mu_Q)^2}{2\sigma_Q^2}}\, dt = \frac{1}{2\pi\sigma_P\sigma_Q}\sqrt{2\pi}\frac{\sigma_P\sigma_Q}{\sqrt{\sigma_P^2+\sigma_Q^2}} e^{-\frac{1}{2}\frac{\left(\mu_P-\mu_Q\right)^2}{\sigma_P^2+\sigma_Q^2}}$$
$$= \frac{1}{\sqrt{2\pi(\sigma_P^2+\sigma_Q^2)}} e^{-\frac{\left(\mu_P-\mu_Q\right)^2}{2\left(\sigma_P^2+\sigma_Q^2\right)}}.$$

We can notice that this is the density of a Gaussian distribution with mean μ_P and variance $\sigma_P^2+\sigma_Q^2$ evaluated at μ_Q. We will also use a well-known notation for this value: $\mathcal{N}(\mu_P \mid \mu_Q, \sigma_P^2+\sigma_Q^2)$. Here

$$\mathcal{N}(a \mid b, s^2) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(a-b)^2}{2s^2}}, \tag{64}$$

and

$$\mathrm{QS}(P,Q) = -2\,\mathcal{N}\left(\mu_P \mid \mu_Q, \sigma_P^2+\sigma_Q^2\right) + \frac{1}{2\sqrt{\pi}\sigma_P}. \tag{65}$$

Additionally, we present another useful identity involving the expected quadratic score for a Gaussian prediction and Gaussian outcome:

$$\int_{\mathbb{R}} p(t)\, q(t)\, dt = -\frac{\mathrm{QS}(P,Q) + H(P)}{2}. \tag{66}$$

B.3.4 Quadratic Score Divergence for Two Gaussians

$$d(P,Q) = \mathbb{E}_{Y\sim Q}\big[\mathrm{QS}(P,Y)\big] - H(Q) = -2\mathbb{E}_{Y\sim Q}\left[p(Y)\right] - H(P) - H(Q). \tag{67}$$

We can already compute these terms using the results from the previous sections. We have the following:

$$-2\mathbb{E}_{Y\sim Q}\big[p(Y)\big] = -2\mathcal{N}\left(\mu_P \mid \mu_Q, \sigma_P^2+\sigma_Q^2\right) = -\frac{2}{\sqrt{2\pi(\sigma_P^2+\sigma_Q^2)}} e^{-\frac{\left(\mu_P-\mu_Q\right)^2}{2\left(\sigma_P^2+\sigma_Q^2\right)}}. \tag{68}$$

$$H(P) = -\frac{1}{2\sqrt{\pi}\sigma_P}, \quad H(Q) = -\frac{1}{2\sqrt{\pi}\sigma_Q}. \tag{69}$$

Thus, we can express the quadratic score divergence for two Gaussians as follows:

$$d(P,Q) = -2\mathcal{N}\left(\mu_P \mid \mu_Q, \sigma_P^2+\sigma_Q^2\right) + \frac{1}{2\sqrt{\pi}\sigma_P} + \frac{1}{2\sqrt{\pi}\sigma_Q}. \tag{70}$$

Another convenient expression for the quadratic score divergence is:

$$d(P,Q) = \mathbb{E}_{Y\sim Q}\big[\mathrm{QS}(P,Y)\big] - H(Q) = -2\int_{\mathbb{R}} p(t)\, q(t)\, dt + \int_{\mathbb{R}} p(t)^2\, dt + \int_{\mathbb{R}} q(t)^2\, dt = \int_{\mathbb{R}} \big(p(t) - q(t)\big)^2\, dt.$$

This divergence is symmetric: $d(P,Q) = d(Q,P)$. It is similar to CRPS score divergence, but this time we use densities instead of cumulative distribution functions.

B.3.5 Entropy of Quadratic Score for a Gaussian Mixture

$$H(\widehat{P}_{\text{ens}}) = -\int_{\mathbb{R}} \left(\frac{1}{M} \sum_{i=1}^{M} \widehat{p}_i(y) \right)^2 dy = -\frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \mathcal{N}\left(\mu_i \mid \mu_j, \sigma_i^2 + \sigma_j^2\right). \tag{71}$$

Alternatively, we can express the entropy of quadratic score for a Gaussian mixture as follows:

$$H(\widehat{P}_{\text{ens}}) = \frac{1}{M^2} \left[\sum_{i=1}^{M} H(\widehat{P}_i) - 2 \sum_{j<i}^{M} \mathcal{N}\left(\mu_i \mid \mu_j, \sigma_i^2 + \sigma_j^2\right) \right] \tag{72}$$

$$= -\frac{1}{M^2} \left[\frac{1}{2\sqrt{\pi}} \sum_{i=1}^{M} \frac{1}{\sigma_i} + 2 \sum_{j<i}^{M} \mathcal{N}\left(\mu_i \mid \mu_j, \sigma_i^2 + \sigma_j^2\right) \right]. \tag{73}$$

Next we express the entropy of quadratic score for a Gaussian mixture in terms of the expected quadratic score and entropy of its components:

$$H(\widehat{P}_{\text{ens}}) = -\frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \int_{\mathbb{R}} \widehat{p}_i(y)\, \widehat{p}_j(y)\, dy = \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{\text{QS}(\widehat{P}_i, \widehat{P}_j) + H(\widehat{P}_i)}{2}$$

$$= \frac{1}{2M} \sum_{i=1}^{M} H(\widehat{P}_i) + \frac{1}{2M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \text{QS}(\widehat{P}_i, \widehat{P}_j) = \frac{M+1}{2M^2} \sum_{i=1}^{M} H(\widehat{P}_i) + \frac{1}{2M^2} \sum_{i \neq j}^{M} \text{QS}(\widehat{P}_i, \widehat{P}_j).$$

B.3.6 Quadratic Score Divergence between a Gaussian and a Gaussian Mixture

$$d(\widehat{P}_i, \widehat{P}_{\text{ens}}) = \mathbb{E}_{Y \sim \widehat{P}_{\text{ens}}} \left[\text{QS}(\widehat{P}_i, Y) \right] - H(\widehat{P}_{\text{ens}}) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}_{Y \sim \widehat{P}_j} \left[\text{QS}(\widehat{P}_i, Y) \right] - H(\widehat{P}_{\text{ens}})$$

$$= \frac{1}{M} \sum_{j=1}^{M} \text{QS}(\widehat{P}_i, \widehat{P}_j) - H(\widehat{P}_{\text{ens}}) = -\frac{2}{M} \sum_{j=1}^{M} \mathcal{N}\left(\mu_i \mid \mu_j, \sigma_i^2 + \sigma_j^2\right) - H(\widehat{P}_i) - H(\widehat{P}_{\text{ens}}). \tag{74}$$

C Proper Scoring Rules from Consistent Scoring Functions

A closely related concept to proper scoring rule is the *consistent scoring function*. In some situations a probabilistic forecast is not possible or not needed (e.g., a point prediction is required by some regulation); for this setup a similar theory was developed (Savage, 1971; Gneiting, 2011). There is one important difference when evaluating point forecasts: since we still follow the random data model (the true outcome y is stochastic, comes from distribution P), we need to select a functional $T(P)$ of the true data distribution that we want to estimate. You can refer to (Gneiting, 2011) for a detailed discussion, and in the next subsection we provide a brief overview of the main definitions and results.

C.1 Introduction

Definition C.1 (Definition 2.1 from (Gneiting, 2011))**.** Scoring function S is consistent for the functional T relative to the class of distributions $\mathcal{P}$ if

$$\mathbb{E}_{Y \sim P} \left[S(t, Y) \right] \leq \mathbb{E}_{Y \sim P} \left[S(x, Y) \right], \tag{75}$$

for all distributions $P \in \mathcal{P}$, all $t \in T(P)$ and all $x \in \text{dom}(Y)$.

One of the most famous results in this area is the following.

Theorem C.2 (Savage (Savage, 1971), 1971)**.** *Every scoring function that is consistent for the mean functional $T(P) = \mathbb{E}_P[Y]$ admits a representation as a Bregman divergence:*

$$S(x, y) = \varphi(y) - \varphi(x) - \varphi'(x)(y - x), \tag{76}$$

where φ is a convex function, and φ' is its subgradient. Such $S(x, y)$ are also called Bregman functions.

For example, the quadratic score $S(x, y) = (y - x)^2$ is a Bregman function with $\varphi(t) = t^2$.

Theorem C.3 (Theorem 2.2 from (Gneiting, 2011))**.** *Score S is consistent for T if and only if any $t \in T(P)$ is an optimal point prediction: $t \in y_{Bayes}$.*

Optimal point prediction in this setting is the Bayes act $\widehat{y}_{\text{Bayes}}$ and can be expressed as follows:

$$\widehat{y}_{\text{Bayes}} = \arg\min_x \mathbb{E}_{Y\sim P}\left[S(x, Y)\right]. \tag{77}$$

In the case of the mean functional, $y_{\text{Bayes}} = T(P) = \mathbb{E}_{Y\sim P}[Y]$ which also follows from the representation of the consistent score as a Bregman function. Now we are ready to introduce the main tool. If we have a scoring function that is consistent for some functional, then we can use it to construct a (rather simple) proper scoring rule:

Theorem C.4 (Theorem 2.3 from (Gneiting, 2011))**.** *If S_T is consistent for the functional $T(\cdot)$, then the following scoring rule is proper:*

$$S(\widehat{P}, y) = S_T\big(T(\widehat{P}), y\big).$$

C.2 SE Score from Quadratic Scoring Function

Based on the quadratic scoring function, which is consistent for the mean, we introduce the following proper scoring rule for continuous distributions:

$$\text{SE}(\widehat{P}, y) = \big(y - \mathbb{E}_{Y\sim\widehat{P}}[Y]\big)^2,$$

and call it the SE score. This score is proper, but not strictly proper, since any distribution with the same mean will have the same score.

C.3 Risk Components for SE Score

The risk components for this score are as follows:

- **Total risk**: $R_{\text{Tot}}(\widehat{P}, P) = \mathbb{E}_{Y\sim P}\big[\text{SE}(\widehat{P}, Y)\big] = \mathbb{E}_{Y\sim P}\left[(Y - \mathbb{E}_{X\sim\widehat{P}}[X])^2\right]$, where

$$\begin{aligned}\mathbb{E}_{Y\sim P}\left[(Y - \mathbb{E}_{X\sim\widehat{P}}[X])^2\right] &= \mathbb{E}_{Y\sim P}\left[Y^2\right] - 2\mathbb{E}_{Y\sim P}[Y]\cdot\mathbb{E}_{X\sim\widehat{P}}[X] + (\mathbb{E}_{X\sim\widehat{P}}[X])^2 \\ &= \left(\mathbb{E}_{X\sim\widehat{P}}[X] - \mathbb{E}_{Y\sim P}[Y]\right)^2 + \mathbb{E}_{Y\sim P}\left[Y^2\right] - (\mathbb{E}_{Y\sim P}[Y])^2 \\ &= \left(\mathbb{E}_{X\sim\widehat{P}}[X] - \mathbb{E}_{Y\sim P}[Y]\right)^2 + \text{Var}_{Y\sim P}[Y].\end{aligned}$$

  $R_{\text{Tot}}(\widehat{P}, P) = \left(\mathbb{E}_{X\sim\widehat{P}}[X] - \mathbb{E}_{Y\sim P}[Y]\right)^2 + \text{Var}_{Y\sim P}[Y]$.

- **Bayes risk**: $R_{\text{Bayes}}(P) = H_{\text{SE}}(P) = \mathbb{E}_{Y\sim P}\big[\text{SE}(P, Y)\big] = \text{Var}_{Y\sim P}[Y]$.

- **Excess risk**: $R_{\text{Exc}}(\widehat{P}, P) = d_{\text{SE}}(\widehat{P}, P) = R_{\text{Tot}}(\widehat{P}, P) - R_{\text{Bayes}}(P) = \left(\mathbb{E}_{X\sim\widehat{P}}[X] - \mathbb{E}_{Y\sim P}[Y]\right)^2$.

In other words, the SE score corresponds to the following entropy and divergence functions:

$$H_{\text{SE}}(P) = \text{Var}_{Y\sim P}[Y],\ d_{\text{SE}}(\widehat{P}, P) = \left(\mathbb{E}_{X\sim\widehat{P}}[X] - \mathbb{E}_{Y\sim P}[Y]\right)^2.$$

Table 10: Representations of some proper scoring rules

Score	**Entropy** $H(P)$	**Supergradient map** $h_P(y)$
CRPS	$\frac{1}{2}\mathbb{E}_{X,X'\sim P}\lvert X-X'\rvert$	$\mathbb{E}_{X\sim P}\lvert X-y\rvert$
LS	$-\int_{\mathbb{R}} p(t)\log p(t)dt$	$-\log p(y)$
QS	$-\int_{\mathbb{R}} p(t)^2 dt$	$-2p(y)$
SE	$\mathrm{Var}_{Y\sim P}[Y]$	$y^2-2y\mathbb{E}_{X\sim P}[X]$

D Characterization of Proper Scoring Rules and General Form of Risk Estimates

The following theorem provides a general form of a proper scoring rule using the notion of a convex function on the space of probability measures:

Definition D.1. A scoring rule S is *regular* if $H(P) = S(P,P)$ is finite and $S(P,Q) > -\infty$, $\forall P, Q \in \mathcal{P}$.

Theorem D.2 (Theorem 1 from (Gneiting & Raftery, 2007), notation from Theorem 12 in (Waghmare & Ziegel, 2026))**.** *A regular scoring rule is proper if and only if there exists a concave function $H\colon \mathcal{P} \to \mathbb{R}$ such that*

$$S(P,y) = H(P) + \langle h_P, \delta_y - P\rangle = H(P) + h_P(y) - \int h_P(t)\,dP(t) \tag{78}$$

for every $P \in \mathcal{P}$ and $y \in \mathcal{Y}$, where h_P is a supergradient of H at P and $\langle h_P, Q\rangle = \int h_P(t)\,dQ(t)$.

From this theorem we can obtain the following general expression of the divergence for a proper scoring rule:

$$d(P,Q) = H(P) - H(Q) - \langle h_P, P-Q\rangle. \tag{79}$$

The concave function of Theorem D.2 coincides with the corresponding score entropy H. In case of our ensemble setting, this leads to the following relation between the risk estimates:

$$\begin{aligned}\widehat{R}^1_{\text{Bayes}} &= \frac{1}{M}\sum_{i=1}^{M} H(\widehat{P}_i) \le H\left(\frac{1}{M}\sum_{i=1}^{M}\widehat{P}_i\right)\\ &= H(\widehat{P}_{\text{ens}}) = \widehat{R}^2_{\text{Bayes}}.\end{aligned} \tag{80}$$

For the Logarithmic score, we know that:

$$\text{LS}\colon \widehat{R}^3_{\text{Bayes}} = H(\widehat{P}_*) \ge H(\widehat{P}_{\text{ens}}) = \widehat{R}^2_{\text{Bayes}}, \tag{81}$$

since among the distributions with the same variance (in this case, both have variance σ_*^2), the Gaussian distribution has the highest (Shannon) entropy. Combining these results, we get:

$$\text{LS}\colon \widehat{R}^1_{\text{Bayes}} \le \widehat{R}^2_{\text{Bayes}} \le \widehat{R}^3_{\text{Bayes}}. \tag{82}$$

Whether this result holds for other scores or in general remains open. We think that it depends on the relationship between the particular score's entropy and the Gaussian distribution that we used for our ensemble estimate.

CRPS.

$$\widehat{R}^3_{\text{Bayes}} = \sqrt{\frac{\sigma_*^2}{\pi}} = \frac{1}{\sqrt{\pi}}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\sigma_i^2 + \mathrm{Var}\left[\mu_i\right])} \ge \frac{1}{\sqrt{\pi}}\sqrt{\frac{1}{M}\sum_{i=1}^{M}\sigma_i^2} \ge \frac{1}{\sqrt{\pi}}\frac{1}{M}\sum_{i=1}^{M}\sigma_i = \widehat{R}^1_{\text{Bayes}}.$$

QS.

$$\widehat{R}^3_{\text{Bayes}} = -\frac{1}{2\sqrt{\pi}\sigma_*} = -\frac{1}{2\sqrt{\pi}} \frac{1}{\sqrt{\frac{1}{M}\sum_{i=1}^M \sigma_i^2 + \text{Var}\left[\mu_i\right]}} \geq$$
$$= -\frac{1}{2\sqrt{\pi}} \frac{1}{\sqrt{\frac{1}{M}\sum_{i=1}^M \sigma_i^2}} \geq -\frac{1}{2\sqrt{\pi}} \frac{1}{M} \sum_{i=1}^M \frac{1}{\sigma_i} = \widehat{R}^1_{\text{Bayes}}.$$

Here we have applied Jensen's inequality to the function $f(x) = -\frac{1}{\sqrt{x}}$, $x_i = \sigma_i^2$.

SE.

$$\widehat{R}^3_{\text{Bayes}} = \sigma_*^2 = \frac{1}{M}\sum_{i=1}^M \sigma_i^2 + \text{Var}\left[\mu_i\right] \geq \frac{1}{M}\sum_{i=1}^M \sigma_i^2 = \widehat{R}^1_{\text{Bayes}}.$$

D.1 Additional Results for Excess Risk

$$\widehat{R}^{2,1}_{\text{Exc}} = \frac{1}{M}\sum_{i=1}^M d(\widehat{P}_{\text{ens}}, \widehat{P}_i) = \frac{1}{M}\sum_{i=1}^M \Big[H(\widehat{P}_{\text{ens}}) - H(\widehat{P}_i) - \langle h_{\widehat{P}_{\text{ens}}}, \widehat{P}_{\text{ens}} - \widehat{P}_i\rangle\Big] \tag{83}$$

$$= H(\widehat{P}_{\text{ens}}) - \frac{1}{M}\sum_{i=1}^M H(\widehat{P}_i) - \langle h_{\widehat{P}_{\text{ens}}}, \widehat{P}_{\text{ens}} - \cancelto{0}{\frac{1}{M}\sum_{i=1}^M \widehat{P}_i} \rangle = \widehat{R}^2_{\text{Bayes}} - \widehat{R}^1_{\text{Bayes}} \geq 0. \tag{84}$$

$$\begin{aligned}
\widehat{R}^{1,1}_{\text{Exc}} &= \frac{1}{M^2}\sum_{i=1}^M\sum_{j=1}^M d(\widehat{P}_i, \widehat{P}_j) = \frac{1}{M^2}\sum_{i=1}^M\sum_{j=1}^M \Big[H(\widehat{P}_i) - H(\widehat{P}_j) - \langle h_{\widehat{P}_i}, \widehat{P}_i - \widehat{P}_j\rangle\Big] \\
&= -\frac{1}{M^2}\sum_{i=1}^M\sum_{j=1}^M \langle h_{\widehat{P}_i}, \widehat{P}_i - \widehat{P}_j\rangle = -\frac{1}{M}\sum_{i=1}^M \langle h_{\widehat{P}_i}, \widehat{P}_i - \frac{1}{M}\sum_{j=1}^M \widehat{P}_j\rangle = \frac{1}{M}\sum_{i=1}^M \langle h_{\widehat{P}_i}, \widehat{P}_{\text{ens}} - \widehat{P}_i\rangle \qquad (85) \\
&\geq \frac{1}{M}\sum_{i=1}^M \Big[H(\widehat{P}_{\text{ens}}) - H(\widehat{P}_i)\Big] = H(\widehat{P}_{\text{ens}}) - \frac{1}{M}\sum_{i=1}^M H(\widehat{P}_i) = \widehat{R}^2_{\text{Bayes}} - \widehat{R}^1_{\text{Bayes}} = \widehat{R}^{2,1}_{\text{Exc}}.
\end{aligned}$$

$$\begin{aligned}
\widehat{R}^{1,2}_{\text{Exc}} &= \frac{1}{M}\sum_{i=1}^M d(\widehat{P}_i, \widehat{P}_{\text{ens}}) = \frac{1}{M}\sum_{i=1}^M \Big[H(\widehat{P}_i) - H(\widehat{P}_{\text{ens}}) - \langle h_{\widehat{P}_i}, \widehat{P}_i - \widehat{P}_{\text{ens}}\rangle\Big] \\
&= \frac{1}{M}\sum_{i=1}^M H(\widehat{P}_i) - H(\widehat{P}_{\text{ens}}) + \frac{1}{M}\sum_{i=1}^M \langle h_{\widehat{P}_i}, \widehat{P}_{\text{ens}} - \widehat{P}_i\rangle = \widehat{R}^{1,1}_{\text{Exc}} - \widehat{R}^{2,1}_{\text{Exc}} \qquad (86) \\
&\implies \widehat{R}^{1,1}_{\text{Exc}} = \widehat{R}^{2,1}_{\text{Exc}} + \widehat{R}^{1,2}_{\text{Exc}}.
\end{aligned}$$

E Experiments

The code for reproducing all the experiments in this paper is provided at TBD

E.1 Training Objective

In our experiments, we consider neural networks predicting the mean and variance of an output distribution $\mathcal{N}(\mu(x;w), \sigma^2(x;w))$ over y for a given input x. The standard way of optimizing a model with parameters w to predict a faithful output distribution is by minimizing the log-likelihood of the output distribution on the training set D (Nix & Weigend, 1994; Lakshminarayanan et al., 2017). Formulated as negative log-likelihood, the element-wise loss is given by

$$\mathcal{L}(w) = \frac{1}{2}\log\sigma^2(x;w) + \frac{(y-\mu(x;w))^2}{2\sigma^2(x;w)} + const. \tag{87}$$

However, training both mean and variance networks in conjunction is known to suffer from instability. Therefore, there have been multiple attempts to stabilize training and obtain both accurate mean predictions and calibrated variance (Skafte et al., 2019; Seitzer et al., 2022; Immer et al., 2023). We follow the approach of Immer et al. (2023) in using the natural parameterization of the Gaussian to reformulate the optimization problem and obtain more stable optimization properties. Instead of predicting mean μ and variance σ^2, in the natural parameterization $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = -\frac{1}{2\sigma^2}$ are predicted, where one needs to make sure that $\eta_2 < 0$. Then, the negative log-likelihood is given by

$$\mathcal{L}(w) = -\begin{pmatrix}\eta_1(x;w)\\ \eta_2(x;w)\end{pmatrix}^T \begin{pmatrix} y \\ y^2\end{pmatrix} - \frac{\eta_1(x;w)^2}{4\eta_2(x;w)} - \frac{1}{2}\log(-2\eta_2(x;w)) + const. \tag{88}$$

We compare the two losses on a simple regression dataset using a three layer neural networks consisting of feed-forward layers with hidden dimension 32 and ReLU activation. Data points are drawn according to a sine function with larger noise in the right open half-plane. While we observe in Figure 5 that using equation (88) needs more gradient steps than optimization with equation (87), training and validation losses are much more stable. Furthermore, the resulting behavior is arguable more favorable for the natural parameterization (see Figure 6). For the standard parameterization, the predicted variance collapses away from data, while it stays at least relatively constant for the natural parameterization. Additionally, for the standard parameterization the predicted mean follows a linear trend, while for the natural parameterization it resorts to a constant value. As the natural parameterization lead to much more stable results throughout all our experiments, we exclusively consider the natural parameterization as training objective in the reported results.

E.2 Synthetic Experiment

The ground-truth conditional is a heteroscedastic two-component mixture,

$$p^*(y \mid x) = \pi(x)P_1(y \mid x) + \big(1-\pi(x)\big)P_2(y \mid x),$$

with mixing weight and component means

$$\begin{aligned}\pi(x) &= \frac{1}{1+\exp(1.2x)},\\ \mu_1(x) &= \tfrac{x}{3} + 1.2\sin(0.8x),\\ \mu_2(x) &= \tfrac{x}{3} - 1.2\cos(0.8x),\end{aligned}$$

and input-dependent noise scale

$$\sigma(x) = 0.12 + 0.28\big(0.5 + 0.5\sin(0.7x)\big)^2.$$

We take the components to be Gaussian with shared heteroscedastic variance,

$$P_1(y \mid x) = \mathcal{N}\big(y; \mu_1(x), \sigma^2(x)\big), \qquad P_2(y \mid x) = \mathcal{N}\big(y; \mu_2(x), \sigma^2(x)\big),$$

equivalently $y = \mu_k(x) + \epsilon\sigma(x)$ with $\epsilon \sim \mathcal{N}(0, 1)$ and $k \in \{1, 2\}$ drawn according to $\pi(x)$. Figure 7 shows the ground-truth conditional and samples.

We draw $n = 1200$ training pairs from $p^*(y \mid x)$ and fit a regression network. The backbone is a fully connected MLP with two hidden layers of width 8 and SiLU (Hendrycks & Gimpel, 2016). On top of the backbone, we use two output heads that parameterize a Gaussian predictive distribution via its natural parameters (Immer et al., 2023), and we train by maximizing the corresponding Gaussian log-likelihood. We form an ensemble of 10 independently initialized models, each trained for 100 epochs with Adam (Kingma & Ba, 2015).

Given the trained ensemble, we compute our uncertainty scores. Figure 2 in the main part of the paper reports one instance of our framework using the logarithmic score: each dot shows the ensemble-averaged predictive mean, and the color intensity encodes the corresponding uncertainty (see color bar).

Below we extend the main-text visualization to additional scoring rules (CRPS, SE, and Quadratic). Across all rules, Bayes risk is highest in regions where the mapping is least determined, i.e., where the two generating curves are sampled with comparable probability, while excess risk rises for inputs outside the training data support, reflecting increased epistemic uncertainty. Although absolute magnitudes vary across rules (a score-specific effect), the spatial patterns are consistent. Extended plots are shown in Figures 8-9.

E.3 Selective Prediction

Datasets. We evaluate on seven datasets. First, we introduce `dots`, where inputs are 32×32 grayscale images containing non-overlapping dots and targets are the corresponding dot counts. Second, we introduce `arrow`, where inputs are 32×32 grayscale images of left-pointing arrows and targets are the arrow angles. Third, we derive five datasets from Cityscapes (Cordts et al., 2016) by tiling each image into 224×224 RGB crops and using the provided segmentation maps to define regression targets: the number of pixels belonging to a given class. The classes are `street`, `building`, `sky`, `car`, and `vegetation`.

For train–test splits, we generate new samples for `dots` and `arrow`, and use the official Cityscapes test split (Cordts et al., 2016) for the remaining datasets.

Training details. For `dots` and `arrow`, we use a small CNN with three convolutional layers followed by two fully connected layers and ReLU activations. Models are trained with Adam (Kingma & Ba, 2015), learning rate 10^{-3}, and batch size 64. For the Cityscapes-derived datasets, we train ResNet-18 (He et al., 2016) with the same optimizer settings. Unless stated otherwise, we report results from ensembles of 10 independently trained models per dataset.

Extended results. Selective prediction is quantified using the prediction–reject ratio (PRR; Malinin & Gales, 2021). Although introduced for classification, PRR extends directly to regression by fixing the performance metric used to compute areas; we use mean squared error (MSE). We consider retention rates (i.e., $1-$rejection) from 0.5 to 1. Example retention curves with PRR are shown in Figure 10 (insets), and full per-dataset results appear in Table 11.

E.4 Out-of-Distribution Detection

Datasets. We construct inputs as 64×64 grayscale mosaics of four MNIST digits (LeCun et al., 1998) arranged (upper-left, upper-right, bottom-left, bottom-right) to form a four-digit number; the target is this number. For out-of-distribution (OOD) data, we substitute each quadrant with images from CIFAR-10 (Krizhevsky, 2009), SVHN (Netzer et al., 2011), or Fashion-MNIST (Xiao et al., 2017), as well as a mixture of these three sources. We also perform a fine-grained analysis of positional effects by replacing exactly one quadrant at a time with EMNIST (Cohen et al., 2017) (upper-left, upper-right, bottom-left, or bottom-right), keeping the others from MNIST.

Table 11: Results for selective prediction (PRR↓) on different datasets $\mathcal{D}$ for different scoring rules (SR).

$\mathcal{D}$	SR	$\widehat{R}^{1,1}_{\text{Tot}}$	$\widehat{R}^{2,1}_{\text{Tot}}$	$\widehat{R}^{3a,1}_{\text{Tot}}$	$\widehat{R}^{3b,1}_{\text{Tot}}$	$\widehat{R}^{3a,2}_{\text{Tot}}$	$\widehat{R}^{3b,2}_{\text{Tot}}$	$\widehat{R}^{1}_{\text{Bayes}}$	$\widehat{R}^{2}_{\text{Bayes}}$	$\widehat{R}^{3a}_{\text{Bayes}}$	$\widehat{R}^{3b}_{\text{Bayes}}$	$\widehat{R}^{1,1}_{\text{Exc}}$	$\widehat{R}^{2,1}_{\text{Exc}}$	$\widehat{R}^{3a,1}_{\text{Exc}}$	$\widehat{R}^{3b,1}_{\text{Exc}}$	$\widehat{R}^{3a,2}_{\text{Exc}}$	$\widehat{R}^{3b,2}_{\text{Exc}}$
dots	CRPS	0.631 ±.002	0.631 ±.002	0.631 ±.002	0.644 ±.002	0.644 ±.002	0.664 ±.002	0.664 ±.002	0.644 ±.002	0.644 ±.002	0.664 ±.002	0.704 ±.001	0.704 ±.001	0.704 ±.001	0.705 ±.001	0.782 ±.002	0.716 ±.001
	Log	0.631 ±.002	0.631 ±.002	0.631 ±.002	0.643 ±.002	-	-	0.664 ±.002	-	0.644 ±.002	0.664 ±.002	0.729 ±.001	0.729 ±.001	0.729 ±.001	0.729 ±.001	-	-
	SE	0.632 ±.002	0.632 ±.002	0.632 ±.002	0.644 ±.002	0.644 ±.002	0.664 ±.002	0.664 ±.002	0.644 ±.002	0.644 ±.002	0.664 ±.002	0.685 ±.001	0.685 ±.001	0.685 ±.001	0.685 ±.001	-	-
	Quad.	0.631 ±.002	0.631 ±.002	0.631 ±.002	0.643 ±.002	0.644 ±.002	0.663 ±.002	0.664 ±.002	0.644 ±.002	0.644 ±.002	0.664 ±.002	0.760 ±.002	0.760 ±.002	0.760 ±.002	0.759 ±.002	0.805 ±.002	0.743 ±.001
arrow	CRPS	0.776 ±.017	0.776 ±.017	0.776 ±.017	0.785 ±.018	0.786 ±.018	0.806 ±.018	0.810 ±.017	0.786 ±.018	0.786 ±.018	0.810 ±.017	0.756 ±.012	0.756 ±.012	0.756 ±.012	0.756 ±.012	0.804 ±.014	0.762 ±.011
	Log	0.773 ±.017	0.773 ±.017	0.775 ±.017	0.783 ±.018	-	-	0.811 ±.017	-	0.786 ±.018	0.810 ±.017	0.765 ±.010	0.765 ±.010	0.765 ±.010	0.766 ±.011	-	-
	SE	0.777 ±.017	0.777 ±.017	0.777 ±.017	0.786 ±.018	0.786 ±.018	0.810 ±.017	0.810 ±.017	0.786 ±.018	0.786 ±.018	0.810 ±.017	0.753 ±.013	0.753 ±.013	0.753 ±.013	0.753 ±.013	-	-
	Quad.	0.774 ±.017	0.774 ±.017	0.774 ±.017	0.784 ±.018	0.785 ±.018	0.801 ±.018	0.811 ±.017	0.786 ±.018	0.786 ±.018	0.810 ±.017	0.783 ±.010	0.783 ±.010	0.783 ±.010	0.782 ±.010	0.815 ±.014	0.776 ±.010
street	CRPS	0.203 ±.013	0.203 ±.013	0.204 ±.013	0.209 ±.014	0.209 ±.014	0.232 ±.014	0.236 ±.014	0.209 ±.014	0.209 ±.014	0.237 ±.014	0.295 ±.012	0.295 ±.012	0.295 ±.012	0.298 ±.012	0.605 ±.016	0.407 ±.012
	Log	0.215 ±.013	0.215 ±.013	0.206 ±.013	0.209 ±.014	-	-	0.237 ±.014	-	0.209 ±.014	0.237 ±.014	0.479 ±.014	0.479 ±.014	0.488 ±.014	0.488 ±.014	-	-
	SE	0.204 ±.013	0.204 ±.013	0.204 ±.013	0.209 ±.014	0.209 ±.014	0.237 ±.014	0.237 ±.014	0.209 ±.014	0.209 ±.014	0.237 ±.014	0.238 ±.011	0.238 ±.011	0.238 ±.011	0.238 ±.011	-	-
	Quad.	0.206 ±.013	0.206 ±.013	0.211 ±.013	0.208 ±.014	0.210 ±.014	0.228 ±.014	0.238 ±.014	0.208 ±.014	0.209 ±.014	0.237 ±.014	0.896 ±.010	0.896 ±.010	0.907 ±.011	0.883 ±.011	1.063 ±.011	0.787 ±.013
building	CRPS	0.227 ±.012	0.227 ±.012	0.227 ±.012	0.235 ±.012	0.236 ±.012	0.264 ±.012	0.273 ±.012	0.237 ±.012	0.236 ±.012	0.270 ±.012	0.240 ±.011	0.240 ±.011	0.242 ±.011	0.242 ±.011	0.551 ±.012	0.318 ±.013
	Log	0.227 ±.012	0.227 ±.012	0.227 ±.012	0.232 ±.012	-	-	0.277 ±.012	-	0.236 ±.012	0.270 ±.012	0.336 ±.013	0.336 ±.013	0.338 ±.013	0.332 ±.013	-	-
	SE	0.228 ±.012	0.228 ±.012	0.228 ±.012	0.236 ±.012	0.236 ±.012	0.270 ±.012	0.270 ±.012	0.236 ±.012	0.236 ±.012	0.270 ±.012	0.229 ±.011	0.229 ±.011	0.229 ±.011	0.229 ±.011	-	-
	Quad.	0.226 ±.012	0.226 ±.012	0.227 ±.012	0.233 ±.013	0.235 ±.012	0.257 ±.013	0.282 ±.011	0.242 ±.013	0.236 ±.012	0.270 ±.012	0.517 ±.012	0.517 ±.012	0.538 ±.012	0.523 ±.012	0.888 ±.011	0.568 ±.013
sky	CRPS	0.043 ±.003	0.043 ±.003	0.044 ±.003	0.045 ±.004	0.045 ±.004	0.050 ±.006	0.051 ±.006	0.044 ±.004	0.045 ±.004	0.052 ±.006	0.044 ±.001	0.044 ±.001	0.045 ±.001	0.044 ±.001	0.099 ±.004	0.050 ±.001
	Log	0.046 ±.003	0.046 ±.003	0.045 ±.003	0.046 ±.004	-	-	0.050 ±.006	-	0.045 ±.004	0.052 ±.006	0.054 ±.001	0.054 ±.001	0.057 ±.001	0.056 ±.001	-	-
	SE	0.044 ±.004	0.044 ±.004	0.044 ±.004	0.045 ±.004	0.045 ±.004	0.052 ±.006	0.052 ±.006	0.045 ±.004	0.045 ±.004	0.052 ±.006	0.043 ±.002	0.043 ±.002	0.043 ±.002	0.043 ±.002	-	-
	Quad.	0.041 ±.003	0.041 ±.003	0.045 ±.003	0.044 ±.004	0.046 ±.004	0.048 ±.006	0.049 ±.006	0.042 ±.004	0.045 ±.004	0.052 ±.006	0.075 ±.003	0.075 ±.003	0.079 ±.003	0.073 ±.003	0.172 ±.004	0.072 ±.002
car	CRPS	0.149 ±.003	0.149 ±.003	0.154 ±.003	0.162 ±.003	0.162 ±.003	0.189 ±.004	0.186 ±.005	0.156 ±.003	0.161 ±.003	0.196 ±.005	0.151 ±.006	0.151 ±.006	0.154 ±.006	0.154 ±.006	0.347 ±.008	0.195 ±.008
	Log	0.176 ±.004	0.176 ±.004	0.163 ±.003	0.172 ±.003	-	-	0.177 ±.004	-	0.161 ±.003	0.196 ±.005	0.219 ±.007	0.219 ±.007	0.226 ±.008	0.219 ±.009	-	-
	SE	0.153 ±.003	0.153 ±.003	0.153 ±.003	0.161 ±.003	0.161 ±.003	0.196 ±.005	0.196 ±.005	0.161 ±.003	0.161 ±.003	0.196 ±.005	0.141 ±.005	0.141 ±.005	0.141 ±.005	0.141 ±.005	-	-
	Quad.	0.153 ±.005	0.153 ±.005	0.174 ±.004	0.174 ±.003	0.170 ±.003	0.187 ±.004	0.170 ±.004	0.148 ±.003	0.161 ±.003	0.196 ±.005	0.352 ±.016	0.352 ±.016	0.377 ±.014	0.342 ±.015	0.617 ±.012	0.334 ±.013
vegetation	CRPS	0.200 ±.011	0.200 ±.011	0.199 ±.011	0.209 ±.011	0.210 ±.011	0.260 ±.011	0.274 ±.012	0.211 ±.011	0.210 ±.011	0.272 ±.011	0.187 ±.009	0.187 ±.009	0.186 ±.009	0.187 ±.009	0.337 ±.005	0.199 ±.008
	Log	0.195 ±.011	0.195 ±.011	0.197 ±.011	0.203 ±.010	-	-	0.276 ±.012	-	0.210 ±.011	0.272 ±.011	0.195 ±.008	0.195 ±.008	0.195 ±.008	0.196 ±.008	-	-
	SE	0.200 ±.011	0.200 ±.011	0.200 ±.011	0.210 ±.011	0.210 ±.011	0.272 ±.011	0.272 ±.011	0.210 ±.011	0.210 ±.011	0.272 ±.011	0.187 ±.011	0.187 ±.011	0.187 ±.011	0.187 ±.011	-	-
	Quad.	0.199 ±.012	0.199 ±.012	0.196 ±.011	0.204 ±.011	0.209 ±.011	0.247 ±.011	0.278 ±.012	0.214 ±.012	0.210 ±.011	0.272 ±.011	0.217 ±.008	0.217 ±.008	0.216 ±.008	0.216 ±.008	0.462 ±.005	0.241 ±.008

Training details. Across experiments, we use the same small CNN as in the selective prediction setup, with a larger first linear layer to accommodate the increased input size. We keep hyperparameters identical to the selective prediction setting, and train ensembles of 10 models on the in-distribution MNIST-mosaic task.

Extended results. For evaluation, we generate matched in-distribution and OOD sets of equal size and compute AUROC using uncertainty scores to discriminate OOD from in-distribution samples. Full per-dataset results are listed in Table 12. Replacing all four quadrants with another dataset yields similar detection performance across sources (first four rows), with SVHN—despite also containing digits—detected comparably well. In the positional analysis (last four rows), replacing the upper-left digit (the most significant position in the four-digit number) is detected most reliably, with performance decreasing as the replaced digit's positional significance decreases, which aligns with its influence on the target value.

E.5 Active Learning

Datasets. For the active learning experiments, we considered the following six tabular datasets from the UCI repository. YearPredictionMSD (`YMSD`; Bertin-Mahieux et al., 2011), SGEMM GPU kernel performance (`SGEMM`; Ballester-Ripoll et al., 2019), Combined Cycle Power Plant (`CCPP`; Tüfekci, 2014), Physicochemical Properties of Protein Tertiary Structure (`CASP`; Rana, 2013), Online News Popularity (`NEWS`; Fernandes et al., 2015) and BlogFeedback (`BLOG`; Buza, 2013). For all datasets, we split 20% into a test set for evaluation. The remaining set was further split into validation and pool set. We selected around 200 samples (exact sizes per dataset in code, which was selected based initial performance compared to performance on full pool set) as a training set to start the active learning experiments.

Table 12: Out-of-distribution detection results (AUROC↑). The ID dataset is MNIST mosaic, where four MNIST digits are tiled in a grid with each corresponding to a digit of the target value. OOD datasets use the corresponding datasets as sources for the tiles. The EMNIST variations replace just one tile with a letter from the EMNIST dataset.

$\mathcal{D}_{\text{ood}}$	SR	$\widehat{R}^{1,1}_{\text{Tot}}$	$\widehat{R}^{2,1}_{\text{Tot}}$	$\widehat{R}^{3a,1}_{\text{Tot}}$	$\widehat{R}^{3b,1}_{\text{Tot}}$	$\widehat{R}^{3a,2}_{\text{Tot}}$	$\widehat{R}^{3b,2}_{\text{Tot}}$	$\widehat{R}^{1}_{\text{Bayes}}$	$\widehat{R}^{2}_{\text{Bayes}}$	$\widehat{R}^{3a}_{\text{Bayes}}$	$\widehat{R}^{3b}_{\text{Bayes}}$	$\widehat{R}^{1,1}_{\text{Exc}}$	$\widehat{R}^{2,1}_{\text{Exc}}$	$\widehat{R}^{3a,1}_{\text{Exc}}$	$\widehat{R}^{3b,1}_{\text{Exc}}$	$\widehat{R}^{3a,2}_{\text{Exc}}$	$\widehat{R}^{3b,2}_{\text{Exc}}$
CIFAR-10	CRPS	0.969 ±.018	0.969 ±.018	0.969 ±.018	0.953 ±.022	0.949 ±.021	0.864 ±.006	0.831 ±.015	0.949 ±.021	0.949 ±.021	0.832 ±.015	0.984 ±.011	0.984 ±.011	0.984 ±.011	0.985 ±.011	0.963 ±.016	0.984 ±.012
	Log	0.977 ±.016	0.977 ±.016	0.972 ±.017	0.962 ±.022	-	-	0.830 ±.015	-	0.949 ±.021	0.832 ±.015	0.984 ±.012	0.984 ±.012	0.984 ±.013	0.984 ±.013	-	-
	SE	0.966 ±.018	0.966 ±.018	0.966 ±.018	0.949 ±.021	0.949 ±.021	0.832 ±.015	0.832 ±.015	0.949 ±.021	0.949 ±.021	0.832 ±.015	0.982 ±.011	0.982 ±.011	0.982 ±.011	0.982 ±.011	-	-
	Quad.	0.976 ±.017	0.976 ±.017	0.976 ±.017	0.962 ±.022	0.950 ±.021	0.893 ±.018	0.829 ±.016	0.949 ±.021	0.949 ±.021	0.832 ±.015	0.972 ±.022	0.972 ±.022	0.972 ±.022	0.974 ±.021	0.950 ±.019	0.980 ±.017
SVHN	CRPS	0.984 ±.011	0.984 ±.011	0.983 ±.011	0.978 ±.013	0.976 ±.012	0.938 ±.000	0.919 ±.012	0.976 ±.013	0.976 ±.012	0.920 ±.012	0.986 ±.012	0.986 ±.012	0.986 ±.012	0.986 ±.012	0.964 ±.022	0.983 ±.016
	Log	0.986 ±.011	0.986 ±.011	0.985 ±.011	0.981 ±.013	-	-	0.919 ±.013	-	0.976 ±.012	0.920 ±.012	0.980 ±.018	0.980 ±.018	0.980 ±.019	0.979 ±.019	-	-
	SE	0.982 ±.011	0.982 ±.011	0.982 ±.011	0.976 ±.012	0.976 ±.012	0.920 ±.012	0.920 ±.012	0.976 ±.012	0.976 ±.012	0.920 ±.012	0.987 ±.010	0.987 ±.010	0.987 ±.010	0.987 ±.010	-	-
	Quad.	0.986 ±.011	0.986 ±.011	0.986 ±.011	0.981 ±.014	0.976 ±.013	0.953 ±.010	0.918 ±.013	0.976 ±.013	0.976 ±.012	0.920 ±.012	0.950 ±.045	0.950 ±.045	0.950 ±.045	0.954 ±.043	0.943 ±.034	0.968 ±.031
Fashion-MNIST	CRPS	0.936 ±.014	0.936 ±.014	0.936 ±.014	0.911 ±.017	0.908 ±.018	0.825 ±.025	0.804 ±.028	0.908 ±.017	0.907 ±.018	0.805 ±.028	0.971 ±.010	0.971 ±.010	0.971 ±.010	0.972 ±.010	0.936 ±.005	0.973 ±.010
	Log	0.950 ±.012	0.950 ±.012	0.941 ±.013	0.922 ±.016	-	-	0.804 ±.028	-	0.907 ±.018	0.805 ±.028	0.974 ±.011	0.974 ±.011	0.974 ±.011	0.973 ±.011	-	-
	SE	0.932 ±.014	0.932 ±.014	0.932 ±.014	0.907 ±.018	0.907 ±.018	0.805 ±.028	0.805 ±.028	0.907 ±.018	0.907 ±.018	0.805 ±.028	0.966 ±.010	0.966 ±.010	0.966 ±.010	0.966 ±.010	-	-
	Quad.	0.946 ±.012	0.946 ±.012	0.947 ±.013	0.921 ±.016	0.908 ±.017	0.844 ±.023	0.803 ±.028	0.907 ±.017	0.907 ±.018	0.805 ±.028	0.963 ±.017	0.963 ±.017	0.963 ±.017	0.964 ±.017	0.922 ±.006	0.970 ±.013
Mixture (C,S,F)	CRPS	0.963 ±.013	0.963 ±.013	0.963 ±.013	0.947 ±.015	0.944 ±.015	0.875 ±.010	0.851 ±.017	0.944 ±.015	0.944 ±.015	0.852 ±.017	0.981 ±.010	0.981 ±.010	0.981 ±.010	0.981 ±.010	0.955 ±.014	0.980 ±.012
	Log	0.971 ±.012	0.971 ±.012	0.966 ±.013	0.955 ±.015	-	-	0.850 ±.017	-	0.944 ±.015	0.852 ±.017	0.979 ±.013	0.979 ±.013	0.979 ±.014	0.979 ±.014	-	-
	SE	0.960 ±.013	0.960 ±.013	0.960 ±.013	0.944 ±.015	0.944 ±.015	0.852 ±.017	0.852 ±.017	0.944 ±.015	0.944 ±.015	0.852 ±.017	0.979 ±.010	0.979 ±.010	0.979 ±.010	0.979 ±.010	-	-
	Quad.	0.969 ±.012	0.969 ±.012	0.970 ±.012	0.954 ±.015	0.945 ±.015	0.896 ±.013	0.849 ±.017	0.944 ±.015	0.944 ±.015	0.852 ±.017	0.962 ±.028	0.962 ±.028	0.962 ±.028	0.964 ±.027	0.938 ±.019	0.973 ±.020
EMNIST (tl)	CRPS	0.836 ±.012	0.836 ±.012	0.836 ±.012	0.800 ±.012	0.797 ±.012	0.718 ±.013	0.709 ±.014	0.797 ±.012	0.796 ±.012	0.709 ±.014	0.915 ±.012	0.915 ±.012	0.915 ±.012	0.915 ±.012	0.846 ±.020	0.919 ±.012
	Log	0.855 ±.012	0.855 ±.012	0.842 ±.012	0.811 ±.012	-	-	0.709 ±.014	-	0.796 ±.012	0.709 ±.014	0.922 ±.011	0.922 ±.011	0.922 ±.011	0.921 ±.011	-	-
	SE	0.831 ±.012	0.831 ±.012	0.831 ±.012	0.796 ±.012	0.796 ±.012	0.709 ±.014	0.709 ±.014	0.796 ±.012	0.796 ±.012	0.709 ±.014	0.903 ±.012	0.903 ±.012	0.903 ±.012	0.903 ±.012	-	-
	Quad.	0.850 ±.012	0.850 ±.012	0.850 ±.012	0.809 ±.012	0.797 ±.012	0.731 ±.013	0.708 ±.014	0.797 ±.012	0.796 ±.012	0.709 ±.014	0.910 ±.011	0.910 ±.011	0.910 ±.011	0.911 ±.011	0.827 ±.026	0.919 ±.012
EMNIST (tr)	CRPS	0.559 ±.010	0.559 ±.010	0.559 ±.010	0.550 ±.009	0.550 ±.009	0.534 ±.007	0.533 ±.007	0.550 ±.009	0.550 ±.009	0.533 ±.007	0.594 ±.014	0.594 ±.014	0.594 ±.014	0.594 ±.014	0.580 ±.010	0.595 ±.013
	Log	0.563 ±.011	0.563 ±.011	0.561 ±.010	0.552 ±.010	-	-	0.533 ±.007	-	0.550 ±.009	0.533 ±.007	0.600 ±.016	0.600 ±.016	0.600 ±.016	0.599 ±.016	-	-
	SE	0.558 ±.010	0.558 ±.010	0.558 ±.010	0.550 ±.009	0.550 ±.009	0.533 ±.007	0.533 ±.007	0.550 ±.009	0.550 ±.009	0.533 ±.007	0.589 ±.012	0.589 ±.012	0.589 ±.012	0.589 ±.012	-	-
	Quad.	0.562 ±.010	0.562 ±.010	0.562 ±.010	0.552 ±.009	0.550 ±.009	0.536 ±.007	0.533 ±.007	0.550 ±.009	0.550 ±.009	0.533 ±.007	0.598 ±.016	0.598 ±.016	0.599 ±.016	0.598 ±.016	0.578 ±.009	0.597 ±.014
EMNIST (bl)	CRPS	0.557 ±.022	0.557 ±.022	0.557 ±.022	0.550 ±.021	0.549 ±.021	0.535 ±.019	0.534 ±.018	0.549 ±.021	0.549 ±.021	0.534 ±.018	0.591 ±.022	0.591 ±.022	0.591 ±.022	0.591 ±.022	0.561 ±.013	0.593 ±.021
	Log	0.560 ±.022	0.560 ±.022	0.558 ±.022	0.551 ±.021	-	-	0.534 ±.018	-	0.549 ±.021	0.534 ±.018	0.595 ±.021	0.595 ±.021	0.594 ±.021	0.594 ±.021	-	-
	SE	0.556 ±.021	0.556 ±.021	0.556 ±.021	0.549 ±.021	0.549 ±.021	0.534 ±.018	0.534 ±.018	0.549 ±.021	0.549 ±.021	0.534 ±.018	0.586 ±.022	0.586 ±.022	0.586 ±.022	0.586 ±.022	-	-
	Quad.	0.560 ±.022	0.560 ±.022	0.560 ±.022	0.551 ±.021	0.549 ±.021	0.537 ±.019	0.534 ±.018	0.549 ±.021	0.549 ±.021	0.534 ±.018	0.593 ±.018	0.593 ±.018	0.593 ±.018	0.593 ±.018	0.557 ±.014	0.595 ±.019
EMNIST (br)	CRPS	0.548 ±.013	0.548 ±.013	0.548 ±.013	0.541 ±.014	0.541 ±.013	0.529 ±.013	0.528 ±.013	0.541 ±.013	0.541 ±.013	0.528 ±.013	0.577 ±.010	0.577 ±.010	0.577 ±.010	0.577 ±.010	0.553 ±.019	0.577 ±.009
	Log	0.551 ±.013	0.551 ±.013	0.549 ±.013	0.542 ±.014	-	-	0.528 ±.013	-	0.541 ±.013	0.528 ±.013	0.581 ±.009	0.581 ±.009	0.581 ±.009	0.581 ±.009	-	-
	SE	0.547 ±.013	0.547 ±.013	0.547 ±.013	0.541 ±.013	0.541 ±.013	0.528 ±.013	0.528 ±.013	0.541 ±.013	0.541 ±.013	0.528 ±.013	0.573 ±.011	0.573 ±.011	0.573 ±.011	0.573 ±.011	-	-
	Quad.	0.550 ±.013	0.550 ±.013	0.550 ±.013	0.542 ±.014	0.541 ±.013	0.530 ±.013	0.528 ±.013	0.541 ±.013	0.541 ±.013	0.528 ±.013	0.579 ±.008	0.579 ±.008	0.579 ±.008	0.579 ±.008	0.552 ±.020	0.578 ±.008

Training details. In each acquisition iteration, we selected a number of additional samples from the pool to augment the training set. We do not strictly select based on the score induced by a given uncertainty measure, but obtain a categorical distribution based on applying the softmax over the uncertainty score for all samples in the pool set. We then sample without replacement according to this distribution. This avoids selecting very similar samples within an acquisition batch, which is a known issue in practice (Kirsch et al., 2019). For these experiments, we utilized ensembles of standard three-layer MLPs with ReLU activations with hidden sizes of 100.

Results. The full results are provided in Figure 11. We observe that for most settings, using the uncertainty scores as acquisition function improves upon random sampling. Notable exceptions are total and Bayes risks for the quadratic score, which perform rather poorly.

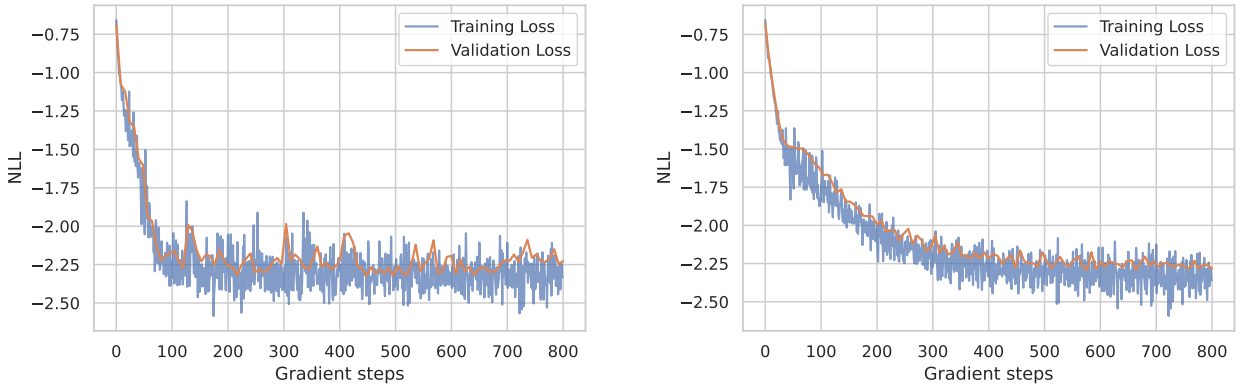


Figure 5: Training and validation loss when minimizing the Gaussian NLL on a synthetic regression example. **Left:** Training with the standard parameterization (equation (87)). **Right:** Training with the natural parameterization (equation (88)).

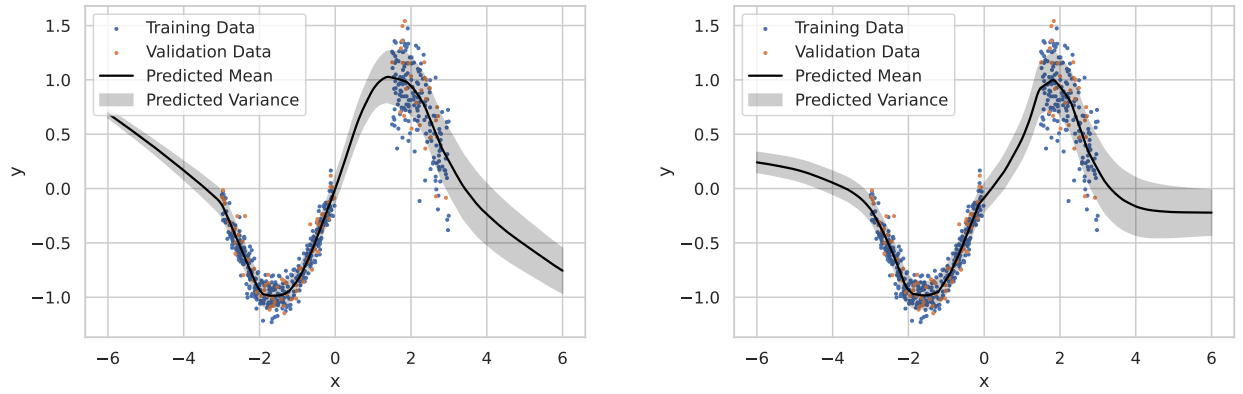


Figure 6: Best model when minimizing the Gaussian NLL on a synthetic regression example. **Left:** Predicting with the standard parameterization (equation (87)). **Right:** Predicting with the natural parameterization (equation (88)).

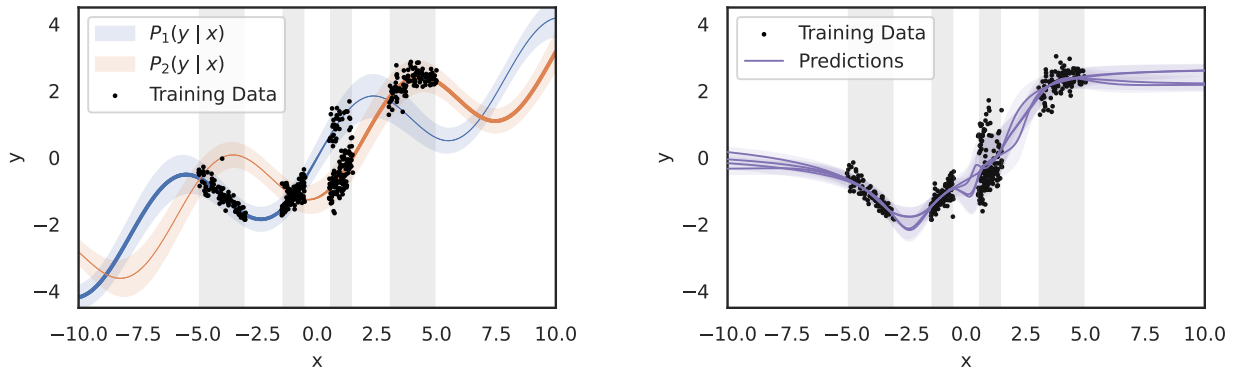
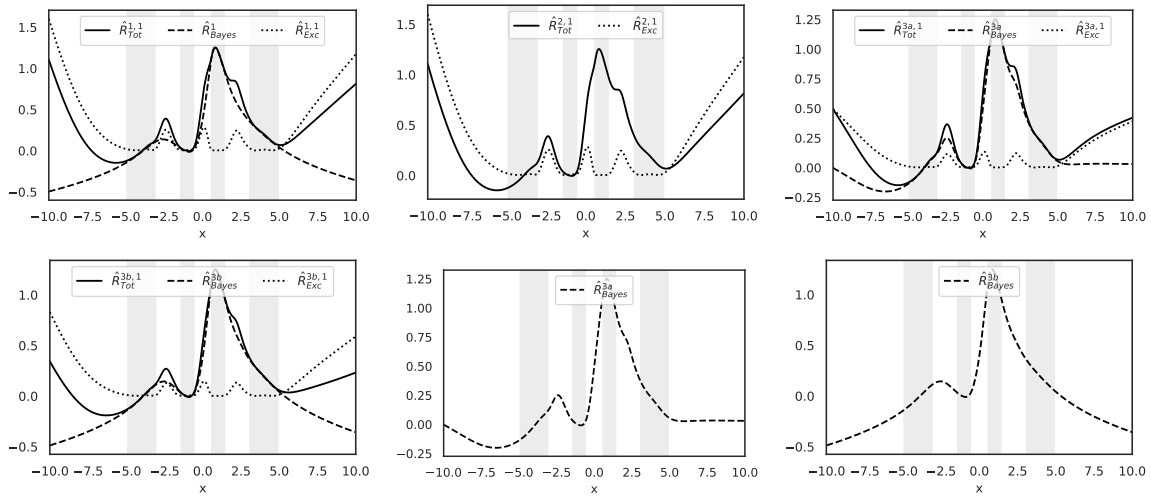
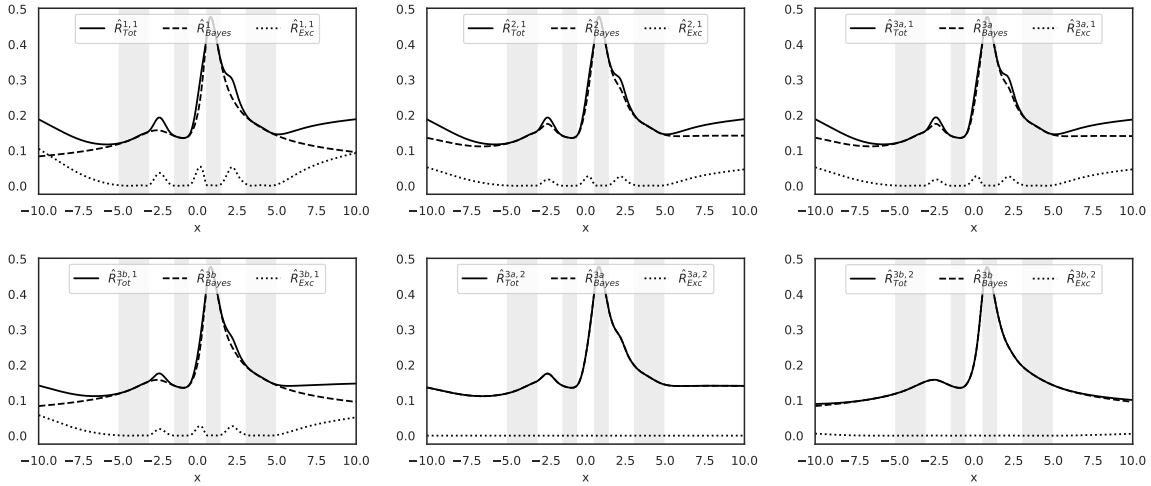


Figure 7: Synthetic regression setup. **Left:** Mixture components P_1 and P_2 of the ground-truth conditional $p^*(y \mid x)$; curve thickness encodes the mixing weights $\pi(x)$ and $1 - \pi(x)$, respectively. **Right:** Training samples drawn from p^* and predictive distributions of individual ensemble members.

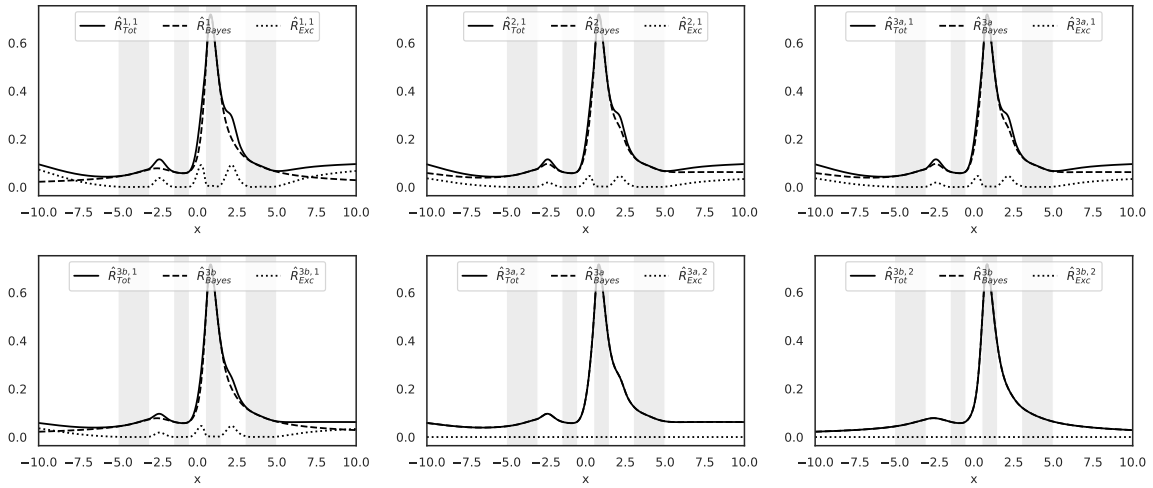


(a) Log score

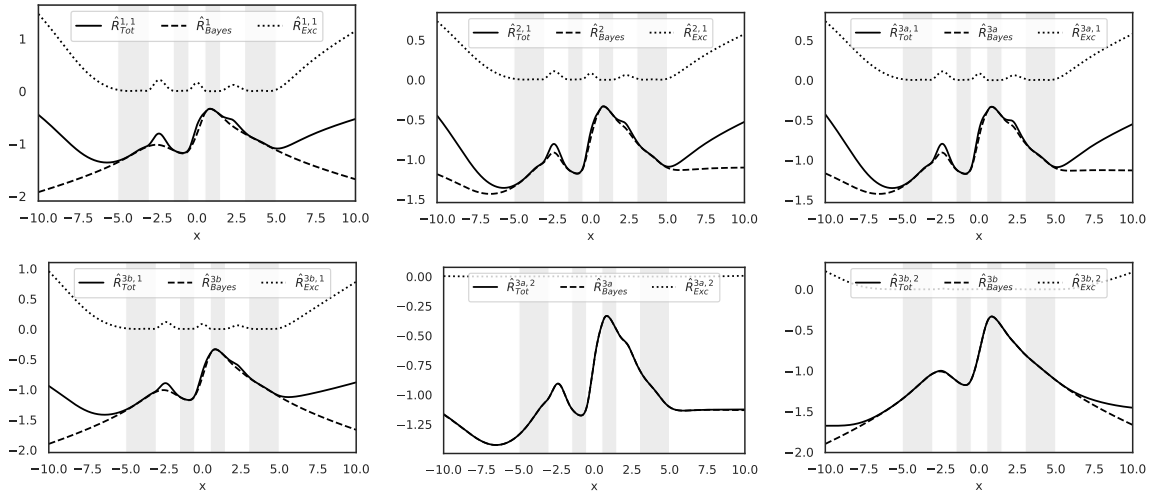


(b) CRPS score

Figure 8: Synthetic 1D experiment (part 1): Log and CRPS score.



(a) SE score



(b) Quadratic score

Figure 9: Synthetic 1D experiment (part 2): SE and Quadratic score.

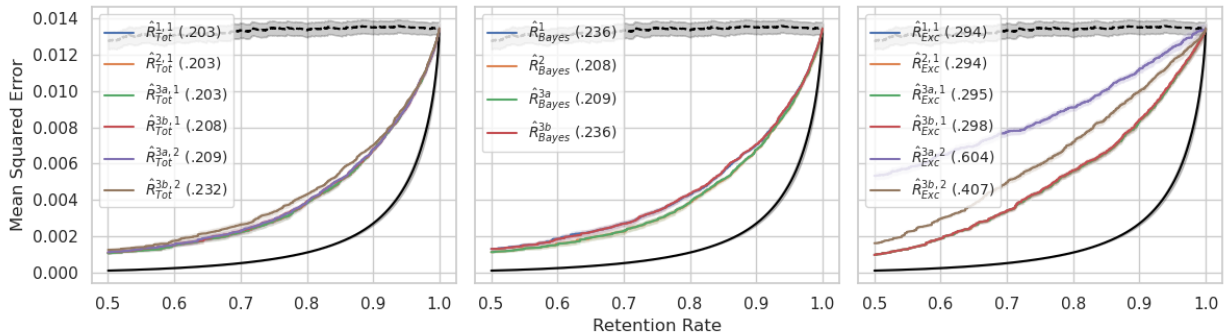


Figure 10: Exemplary retention curve for `street` dataset, PRRs reported in the inset. Black solid line is the optimal, oracle retention curved sorted by the actual MSE of each prediction. The black dashed line is the random baseline.

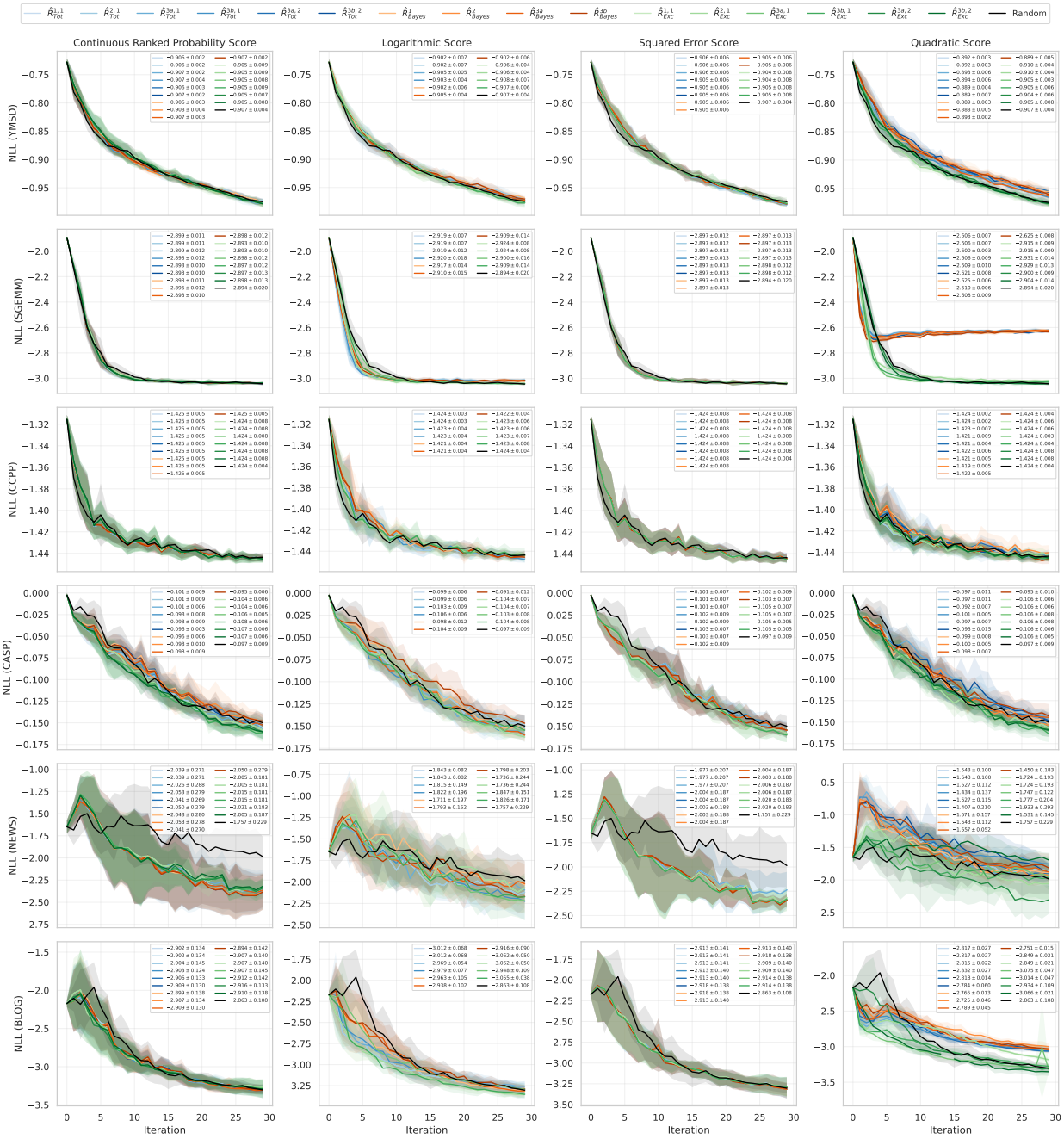


Figure 11: Detailed results on active learning. We report the negative log-likelihood (NLL) of the ensemble after training for each acquisition iteration. In the insets are the average NLLs over the iterations as summary statistic to compare different acquisition functions. Statistics are obtained over five runs.