Is Epistemic Uncertainty Faithfully Represented by Evidential Deep Learning Methods?

Mira Jürgens [1] **Nis Meinert** [2] **Viktor Bengs** [3] **Eyke Hüllermeier** [3] **Willem Waegeman** [1]

Abstract

Trustworthy ML systems should not only return accurate predictions, but also a reliable representation of their uncertainty. Bayesian methods are commonly used to quantify both aleatoric and epistemic uncertainty, but alternative approaches, such as evidential deep learning methods, have become popular in recent years. The latter group of methods in essence extends empirical risk minimization (ERM) for predicting second-order probability distributions over outcomes, from which measures of epistemic (and aleatoric) uncertainty can be extracted. This paper presents novel theoretical insights of evidential deep learning, highlighting the difficulties in optimizing second-order loss functions and interpreting the resulting epistemic uncertainty measures. With a systematic setup that covers a wide range of approaches for classification, regression and counts, it provides novel insights into issues of identifiability and convergence in second-order loss minimization, and the relative (rather than absolute) nature of epistemic uncertainty measures.

1. Introduction

Machine learning systems are capable of making predictions based on patterns found in data. However, these patterns may be incomplete, noisy, or biased, which can lead to errors and inaccuracies in the system's predictions. To address this challenge, uncertainty quantification has emerged as a technique to measure the reliability of a model's predictions and the confidence that can be placed in its results. In supervised learning, uncertainty is often classified into two categories: aleatoric and epistemic (Kendall and Gal, 2017; Hüllermeier and Waegeman, 2021). Since outcomes cannot be predicted deterministically, we assume for classical supervised learning methods a conditional probability $p(y \mid \boldsymbol{x})$ on $\mathcal{Y}$ as the ground truth and train a probabilistic predictor to produce estimates $\hat{p}(y \mid \boldsymbol{x})$. These estimates capture the aleatoric uncertainty about the actual outcome $y \in \mathcal{Y}$, which refers to the inherent randomness that the learner cannot eliminate. However, this approach does not allow the learner to express its epistemic uncertainty, which is the lack of knowledge about how accurately $\hat{p}$ approximates p.

To capture both types of uncertainty, a popular choice is to consider second-order distributions, which map a query instance $\boldsymbol{x}$ to a probability distribution over all possible candidate distributions p. This concept forms one of the core concepts behind Bayesian methods, such as Bayesian neural networks, together with the incorporation of domain knowledge via a prior distribution (Depeweg et al., 2018). However, exact computation of the posterior distribution can be computationally intractable, particularly when using deep learning methods. As a result, approximate techniques such as variational inference and sampling-based methods have been used extensively (Graves, 2011).

An alternative approach, known as evidential deep learning, that adopts a direct uncertainty quantification by means of second-order risk minimization, is currently gaining a lot of traction (see Table 1 for an overview of the most important methods). This approach takes a frequentist statistics perspective and introduces specific loss functions to estimate a second-order distribution over $\hat{p}(y \mid \boldsymbol{x})$, without the need for specifying a prior, unlike Bayesian methods. In the last years, evidential deep learning has been widely used in applications of supervised learning, and many authors claim that it is state-of-the-art in downstream ML tasks that benefit from epistemic uncertainty quantification, such as out-of-distribution detection (Yang et al., 2022), robustness to adversarial attacks (Kopetzki et al., 2021) or active learning (Park et al., 2023). However, in all these tasks, one only needs *relative* epistemic uncertainty measures, e.g., when an example attains a higher epistemic uncertainty than a second example, it is more likely to be out-of-distribution. For specific optimization problems, some authors have expressed

[1]Department of Data Analysis and Mathematical Modeling, Ghent University, Belgium [2]Institue of Communications and Navigation, German Aerospace Center (DLR), Neustrelitz, Germany [3]Department of Informatics, University of Munich (LMU), Germany. Correspondence to: Mira Jürgens <mira.juergens@ugent.be>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Table 1. Overview of most important methods analyzed in this paper. See text for a further explanation of the column headers (NLL = negative log-likelihood, MAE = mean absolute error, OOD = out-of-distribution, KL = Kullbeck-Leibler divergence).

METHOD	OPTIMIZATION PROBLEM	FIRST-ORDER LOSS L_1	SETTING	REGULARIZER $R(\mathbf{\Phi})$	QUANTIFICATION OF EPISTEMIC UNCERTAINTY
SENSOY ET AL. (2018)	OUTER	BRIER	CLASSIFICATION	ENTROPY	PSEUDO-COUNTS
MALININ AND GALES (2018)	OUTER	KL	CLASSIFICATION	OOD KL	MUTUAL INFORMATION
MALININ AND GALES (2019)	OUTER	REVERSE KL	CLASSIFICATION	OOD KL	MUTUAL INFORMATION
CHARPENTIER ET AL. (2020)	OUTER	NLL	CLASSIFICATION	ENTROPY	PSEUDO-COUNTS
HUSELJIC ET AL. (2020)	INNER	NLL	CLASSIFICATION	OOD KL	PSEUDO-COUNTS
MALININ ET AL. (2020)	OUTER	NLL	REGRESSION	ENTROPY	MUTUAL INFORMATION
AMINI ET AL. (2020)	INNER	NLL	REGRESSION	SPECIFIC	VARIANCE
MA ET AL. (2021)	INNER	NLL	REGRESSION	SEPCIFIC	VARIANCE
TSILIGKARIDIS (2021)	OUTER	L_p-NORM	CLASSIFICATION	RÉNYI DIVERGENCE	MUTUAL INFORMATION
KOPETZKI ET AL. (2021)	OUTER	MULTIPLE	CLASSIFICATION	ENTROPY	PSEUDO-COUNTS
BAO ET AL. (2021)	OUTER	BRIER	CLASSIFICATION	SPECIFIC	PSEUDO-COUNTS
LIU ET AL. (2021)	INNER	NLL+MAE	REGRESSION	PSEUDO-COUNTS	VARIANCE
CHARPENTIER ET AL. (2022)	OUTER	NLL	CLASS./REG./COUNTS	ENTROPY	PSEUDO-COUNTS
OH AND SHIN (2022)	INNER	SQUARED LOSS	REGRESSION	GRADIENT-BASED	VARIANCE
PANDEY AND YU (2023)	INNER	NLL	REGRESSION	DISTANCE TO OOD	VARIANCE
KOTELEVSKII ET AL. (2023)	OUTER	SPECIFIC	CLASSIFICATION	ENTROPY	ENTROPY

Table 2. Overview of the key points covered in previous work that critically scrutinizes evidential deep learning (EDL). Detailed differences are discussed in Section 3.

REFERENCE	CRITIQUE	CLASSIFICATION	REGRESSION	COUNT	EXPERIMENTS
BENGS ET AL. (2022)	CONVERGENCE ISSUES OF EDL	✓	✗	✗	✓
MEINERT ET AL. (2023)	CONVERGENCE ISSUES OF EDL	✗	✓	✗	✓
BENGS ET AL. (2023)	LACK OF PROPERNESS OF SECOND-ORDER LOSS FUNCTIONS	✓	✓	✗	✗
OUR WORK	NO COHERENCE WITH A SUITABLE REFERENCE DISTRIBUTION	✓	✓	✓	✓

concerns in different ways as to whether such measures also have a quantitative interpretation. Nonetheless, previous papers have not covered all possible aspects, see Table 2 for a general overview.

This paper presents novel theoretical results that put previous papers in a unifying perspective by analyzing a broad class of exponential family models that cover classification, regression and count data. More specifically, we intend to answer the following two questions: (a) What properties are needed for second-order risk minimization methods to represent epistemic uncertainty in a quantitatively faithful manner? and (b) Do such methods represent epistemic uncertainty in a faithful manner? For the first research question, we will introduce in Section 3.1 the notion of a *reference distribution*, and we will argue that second-order methods should provide an estimate of this distribution if they intend to model epistemic uncertainty in a frequentist manner. For the second research question, we will present in Sections 3.2-3.4 and Section 4 novel theoretical and experimental results, which show that the answer to this question is negative, i.e., evidential deep learning methods do not represent epistemic uncertainty in a faithful manner.

2. Problem definition and literature review

Consider a classification or regression task with instance space $\mathcal{X}$ and label space $\mathcal{Y}$. Without loss of generality, we assume $\mathcal{X} = \mathbb{R}^D$ with D the dimensionality of the data. Similarly, we assume $\mathcal{Y} = \mathbb{R}^{D'}$ in the case of regression, $\mathcal{Y} = \{1, \dots, K\}$ in K-class classification, and $\mathcal{Y} = \mathbb{N}$ in the case of counts. Furthermore, we have training data of the form $\mathcal{D} = \left\{ \left(\boldsymbol{x}_i, y_i \right) \right\}_{i=1}^N$, in which each $(\boldsymbol{x}_i, y_i)$ is i.i.d. sampled via a joint probability measure P on $\mathcal{X} \times \mathcal{Y}$. Correspondingly, each feature vector $\boldsymbol{x} \in \mathcal{X}$ is associated with a conditional distribution $p(\cdot \,|\, \boldsymbol{x})$ on $\mathcal{Y}$, such that $p(y \,|\, \boldsymbol{x})$ is the probability to observe label y as an outcome given $\boldsymbol{x}$. More generally, $\mathbb{P}(A)$ will denote the set of all probability distributions on the set A.

2.1. First-order risk minimization

The classical statistics and machine learning literature usually considers first-order distributions during risk minimization. More specifically, one has a hypothesis space $\mathcal{H}_1$ containing $\Theta \to \mathbb{P}(\mathcal{Y})$ mappings with Θ a parameter space. In fact, $\mathcal{H}_1$ defines probability density (or mass) functions $p(y \,|\, \boldsymbol{\theta})$ from a parametric family with parameter vector $\boldsymbol{\theta} \in \Theta$, for continuous (or discrete) random variables. In a supervised learning context, $\boldsymbol{\theta}$ is a function of the features: we will use the notation $\boldsymbol{\theta}(\boldsymbol{x}; \mathbf{\Phi})$, being an $\mathcal{X} \to \Theta$ function that is parameterized by a set of parameters $\mathbf{\Phi}$. Risk minimization is then usually defined as

$$\min_{\mathbf{\Phi}} \sum_{i=1}^{N} L_1(y_i, p(y \,|\, \boldsymbol{\theta}(\boldsymbol{x}_i; \mathbf{\Phi}))) \,. \tag{1}$$

As a loss function $L_1 : \mathcal{Y} \times \mathbb{P}(\mathcal{Y}) \to \mathbb{R}$, often the negative log-likelihood is chosen:

$$L_1(y, p(y \mid \boldsymbol{\theta}(\boldsymbol{x}; \boldsymbol{\Phi}))) = -\log(p(y \mid \boldsymbol{\theta}(\boldsymbol{x}; \boldsymbol{\Phi}))\,.$$

For classical machine learning models, usually a regularization term that penalizes $\boldsymbol{\Phi}$ is incorporated in (1) as well, but this is less common for modern deep learning models.

2.2. Second-order risk minimization

In second-order risk minimization, one defines a second hypothesis space $\mathcal{H}_2$ with $\mathcal{M} \to \mathbb{P}(\Theta)$ mappings with parameter space $\mathcal{M}$. Thus, $\mathcal{H}_2$ contains probability density functions $p(\boldsymbol{\theta} \mid \boldsymbol{m})$ from a parametric family with parameter vector $\boldsymbol{m}(\boldsymbol{x}; \boldsymbol{\Phi}) \in \mathcal{M}$ that is a function of $\boldsymbol{x}$. The function $\boldsymbol{m}(\boldsymbol{x}; \boldsymbol{\Phi})$, in turn, will also be parameterized by a set of parameters $\boldsymbol{\Phi}$, which will be estimated using a training dataset. Below we will show concretely how $\boldsymbol{m}(\boldsymbol{x}; \boldsymbol{\Phi}) \in \mathcal{M}$ can be parameterized in the case of neural networks.

As shown in Table 1, one considers two different types of optimization problems in second-order risk minimization. In a first branch of papers, which will be further referred to as *inner loss minimization*, one computes an expectation over $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))$ inside the loss function. Let $L_1 : \mathcal{Y} \times \mathbb{P}(\mathcal{Y})$ be a traditional loss function, and let $R(\boldsymbol{\Phi})$ be a penalty term with regularization parameter $\lambda \in \mathbb{R}_{\geq 0}$, then the empirical risk is minimized as follows:

$$\min_{\boldsymbol{\Phi}} \sum_{i=1}^{N} L_1\big(y_i, \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))}[p(y \mid \boldsymbol{\theta})]\big) + \lambda R(\boldsymbol{\Phi}). \quad (2)$$

Most of the papers with this formulation focus on the regression setting, adopting various mechanisms for the penalty term $R(\boldsymbol{\Phi})$.

In a second branch of papers, which will be further referred to as *outer loss minimization*, one computes an expectation over $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))$ outside the loss function. The risk is then defined and minimized in a very similar way:

$$\min_{\boldsymbol{\Phi}} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))}[L_1(y_i, p(y \mid \boldsymbol{\theta}))] + \lambda R(\boldsymbol{\Phi})\,. \quad (3)$$

Papers with this formulation mainly focus on the classification setting. Here, too, one constructs various mechanisms for the penalty term $R(\boldsymbol{\Phi})$.

Second-order risk minimization is often presented as an alternative for Bayesian inference in neural networks. Given a training dataset $\mathcal{D}$ and a prior $p(\boldsymbol{\Phi})$, Bayesian methods maintain a posterior distribution $p(\boldsymbol{\Phi} \mid \mathcal{D})$ over the network parameters $\boldsymbol{\Phi}$. From the posterior one can also obtain a distribution over the parameter vector $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \int \boldsymbol{\theta}(\boldsymbol{x}; \boldsymbol{\Phi})\, p(\boldsymbol{\Phi} \mid \mathcal{D})\, d\boldsymbol{\Phi} \quad \text{for all } \boldsymbol{x} \in \mathcal{X}.$$

For Bayesian neural networks, computing the posterior is usually very challenging, but deep evidential learning methods avoid the usage of sophisticated inference algorithms (Ulmer et al., 2023): Instead of defining a prior over the weights of the neural network, one can also define a prior over the parameters $\boldsymbol{\theta}$ of the first-order probability distribution (Bissiri et al., 2016). That is why in evidential deep learning the conjugate prior of $p(y \mid \boldsymbol{\theta})$ is often picked as a hypothesis space for $p(\boldsymbol{\theta} \mid \boldsymbol{m})$.

2.3. The exponential family

All previous studies consider exponential family distributions as the underlying hypothesis space for $\mathcal{H}_1$ and $\mathcal{H}_2$. The exponential family can be used to represent classification, regression and count data using Categorical, Normal, and Poisson distributions, respectively. For clarity we briefly review the most popular parameterizations.

In multi-class classification, $\mathcal{H}_1$ contains all categorical distributions with parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \triangle^K$ for $\triangle^K$ being the probability $(K-1)$-simplex. As a conjugate prior, $\mathcal{H}_2$ contains all Dirichlet distributions with parameter vector $\boldsymbol{m} = (m_1, \ldots, m_K) \in \mathbb{R}_+^K$. With neural networks it is natural to consider $m_k(\boldsymbol{x}; \boldsymbol{\Phi}) = \exp\{\boldsymbol{w}_k^\top \boldsymbol{\Psi}(\boldsymbol{x})\}$ with $\boldsymbol{w}_k$ a parameter vector per class, and $\boldsymbol{\Psi} : \mathbb{R}^D \to \mathbb{R}^C$ a function that maps the input to an embedding via multiple feature learning layers. The parameter set $\boldsymbol{\Phi}$ is then given by $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K\}$ and all parameters learned in the embedding $\boldsymbol{\Psi}$. Sometimes the architecture is slightly changed, so that $m_1, ..., m_K$ are lower bounded by one (Kopetzki et al., 2021).

In regression, $\mathcal{H}_1$ typically represents the set of Gaussian distributions with parameters $\boldsymbol{\theta} = (\mu, \sigma^2)$. $\mathcal{H}_2$ contains all normal-inverse Gamma (NIG) distributions with parameters $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta)$. Second-order risk minimization is typically realized via a neural network architecture with one or multiple feature learning layers, represented by a mapping $\boldsymbol{\Psi} : \mathbb{R}^D \to \mathbb{R}^C$, followed by four output neurons with functions $\gamma(\boldsymbol{x}; \boldsymbol{\Psi}) = \boldsymbol{w}_1^\top \boldsymbol{\Psi}(\boldsymbol{x})$, $\nu(\boldsymbol{x}; \boldsymbol{\Psi}) = \exp\{\boldsymbol{w}_2^\top \boldsymbol{\Psi}(\boldsymbol{x})\}$, $\alpha(\boldsymbol{x}; \boldsymbol{\Psi}) = \exp\{\boldsymbol{w}_3^\top \boldsymbol{\Psi}(\boldsymbol{x})\}$ and, $\beta(\boldsymbol{x}; \boldsymbol{\Psi}) = \exp\{\boldsymbol{w}_4^\top \boldsymbol{\Psi}(\boldsymbol{x})\}$. Here the exponential function guarantees that the parameters ν, α and β are positive. Sometimes the softplus function is used as an alternative (Amini et al., 2020). The parameter set $\boldsymbol{\Phi}$ is hence given by $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \boldsymbol{w}_4\}$ and all the parameters of the feature learning layers.

For binary classification problems, $\mathcal{H}_1$ and $\mathcal{H}_2$ simplifiy to Bernoulli and Beta distributions. For multivariate regression problems, $\mathcal{H}_1$ and $\mathcal{H}_2$ generalize to multivariate Gaussian and normal-inverse Wishart distributions, respectively (Meinert and Lavin, 2022; Bramlage et al., 2023). For count data, Charpentier et al. (2022) assume that $\mathcal{H}_1$ contains all Poisson distributions with rate parameter θ, while $\mathcal{H}_2$ con-

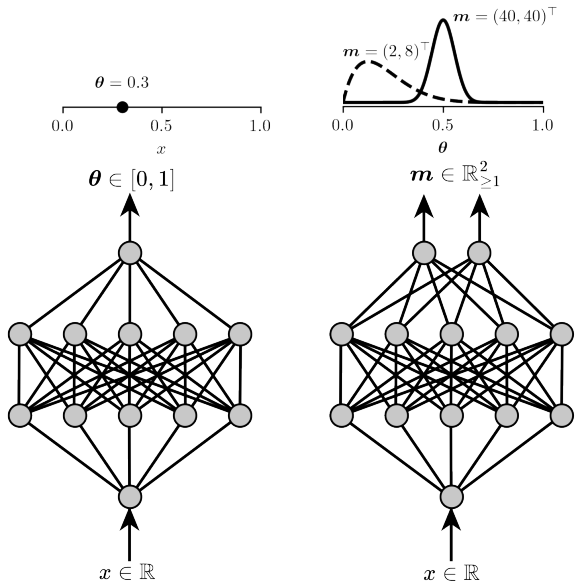


Figure 1. General overview of a neural network architecture for first-order (left) and second-order (right) risk minimization. While the first-order model learns the Bernoulli parameter θ of the data generating distribution directly, the second-order model predicts the parameters $\boldsymbol{m} = (\alpha, \beta)^\top$ of a Beta distribution, which defines a distribution with density $f(\theta|x, \alpha, \beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ over θ. The example is aligned with our experimental setup, where the feature space is one-dimensional.

tains all Gamma distributions with parameters $\boldsymbol{m} = (\alpha, \beta)$. In all these cases, the parameter vector $\boldsymbol{m}$ can be learned as the output neurons of a neural network. Figure 1 shows an example of a potential neural network architecture for the binary classification case.

2.4. Quantifying epistemic uncertainty

In the literature, there is no consensus on how to quantify epistemic uncertainty with second-order risk minimization methods, because various metrics are used. As shown in Table 1, the number of pseudo-counts, defined by $\sum_{k=1}^{K} m_k$, is a popular metric for classification problems. In regression papers it is defined differently, referred to as *evidence*, and only used as a regularizer (Amini et al., 2020). In the latter case, the number of pseudo-counts is given by ν and 2α (Gelman et al., 2013; Meinert et al., 2023), but only ν is used for estimating the epistemic uncertainty. This feels unnatural, since both parameters are necessary to capture the spread of the second-order distribution $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}; \boldsymbol{\Phi}))$.

Mutual information is also often used to quantify epistemic uncertainty of second-order loss minimization methods—as well as Bayesian neural networks (Depeweg et al., 2018). It can be used for classification and regression, and it is defined as the difference between total uncertainty and aleatoric uncertainty:

$$I(\boldsymbol{m}) = \mathrm{H}(\mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))}[p(y \mid \boldsymbol{\theta})]) - \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))}[\mathrm{H}(p(y \mid \boldsymbol{\theta})])\,,$$

with H being the entropy. Recently, for mutual information, Wimmer et al. (2023) raised the question whether an additive decomposition of total uncertainty in aleatoric and epistemic components makes sense, whereupon a slight modification has been suggested by Schweighofer et al. (2023).

Papers that focus on regression typically use the variance of μ for quantifying epistemic uncertainty. For NIG distributions the measure becomes: $\mathrm{Var}(\mu) = \frac{\beta}{(\alpha-1)\nu}$. Some recent works have also investigated variance-based approaches for classification (Sale et al., 2023a; Duan et al., 2024).

An alternative scalar quantity for epistemic uncertainty is the entropy of $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))$. As far as we know, this measure has only been considered in one recent paper (Kotelevskii et al., 2023), which is surprising, because it is often used as a regularizer — see Section 3.4 for a more in-depth discussion. This last paper is also the only paper that consistently chooses the same measure for regularization and quantification of epistemic uncertainty. Roughly speaking, one can say that entropy analyzes the uncertainty of $\boldsymbol{\theta}$ under $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))$, whereas the covariance matrix rather quantifies the spread around the mean of $\boldsymbol{\theta}$. In Appendix A we provide entropy formulas for NIG, Dirichlet and Gamma distributions.

Finally, based on distances and reference quantities, another proposal to quantify the different types of uncertainties was recently made (Sale et al., 2023b). For the concrete choice of the Wasserstein metric, these measures were explicitly determined and calculated for the Dirichlet distribution as an example.

3. Theoretical analysis

We now return to the two research questions that were formulated in the introduction: (a) What properties are needed for second-order risk minimization methods to represent epistemic uncertainty in a quantitatively faithful manner? and (b) Do such methods represent epistemic uncertainty in a faithful manner? The first question will be answered in Section 3.1, and the second one in Sections 3.2—3.4.

3.1. Reference second-order distribution

Most authors use epistemic uncertainty measures as scoring functions to obtain state-of-the-art performance on downstream tasks such as OOD detection, adversarial robustness and active learning, but they do not intend to further interpret those measures. Yet, there are a few notable exceptions (see Table 2). For the outer maximization case, Bengs et al.

(2022) define two theoretical properties that second-order probability distribution should obey: (1) monotonicity in epistemic uncertainty when the sample size increases, and (2) convergence of the second-order distribution to a Dirac delta function when the sample size grows to infinity. In another paper, Bengs et al. (2023) extend classical properties for proper scoring rules to second-order risk minimization and show that existing second-order losses do not lead to proper scores. For regularized inner loss minimization, Meinert et al. (2023) give theoretical and experimental arguments to advocate that deep evidential regression methods do not represent epistemic uncertainty in a faithful manner.

The question regarding the "right" representation of epistemic uncertainty is a difficult one, akin to the fundamental question of how to do statistical inference in the right way (Martin, 2023). Needless to say, a definite answer to this question cannot be expected. For the purpose of our analysis, we opt for a pragmatic approach specifically tailored to our setting, in which the second-order model is essentially meant to capture the uncertainty of a first-order predictor learned from the training data. The latter is the empirical minimizer of (1), which is extended to (2) or (3) by the second-order learner. The randomness of this minimizer is due to the randomness of the training data $\mathcal{D}$.[1] Theoretically, the uncertainty is therefore reflected by the distribution of the first-order predictor induced by the sampling process that generates the training data. This leads us to the notion of a *reference distribution*, and the idea that the epistemic uncertainty predicted by the second-order learner should ideally be represented by this distribution.

Definition 3.1. Let $\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N})$ denote the minimizer of (1) for a training set $\mathcal{D}_N$ of size N, where $\mathcal{D}_N \sim P^N$. For a given $\boldsymbol{x} \in \mathcal{X}$, we define the reference second-order distribution as the conditional distribution induced by $\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N})$:

$$q_N(\boldsymbol{\theta} \mid \boldsymbol{x}) := \mathbb{P}\big(\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N}) = \boldsymbol{\theta}\big) = \int_{(\mathcal{X}\times\mathcal{Y})^N} \mathbb{I}\big[\boldsymbol{\theta}_{\mathcal{D}_N}(\boldsymbol{x}; \boldsymbol{\Phi}_{\mathcal{D}_N}) = \boldsymbol{\theta}\big]\, dP^N\,, \quad (4)$$

where $\mathbb{I}[\cdot]$ is the indicator function. In words, $q_N(\boldsymbol{\theta} \mid \boldsymbol{x})$ is the probability (density) of obtaining the first-order prediction $\boldsymbol{\theta}$ when training on a random data set of size N and using the induced model for predicting on $\boldsymbol{x}$.

Intuitively, this definition follows a frequentist statistics viewpoint, in which a ground-truth model is estimated using limited training data, and epistemic uncertainty is caused by the sum of squared bias and variance of the estimate. The bias-variance decomposition is frequentist in nature, because it assumes a ground-truth model that needs to be estimated using a training dataset of limited size. Typically, the assumption is made that a training dataset of a given size can be resampled to quantify bias (by comparison of the mean to the ground truth) and variance in an objective way. More precisely, one can resample the training datasets d times from the underlying distribution, resulting in d empirical risk minimisers $\{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_d\}$, which then allow to compute the empirical (cumulative) distribution function:

$$\hat{Q}_N(\boldsymbol{\theta}|\boldsymbol{x}) := \frac{1}{d}\sum_{i=1}^{d} \mathbb{I}(\hat{\boldsymbol{\theta}}_i(\boldsymbol{x}, \boldsymbol{\Phi}) \leq \boldsymbol{\theta}).$$

Moreover, note that the reference second-order distribution depends on various design choices for first-order risk minimization, such as the hypothesis space, loss function and optimization algorithm. As such, we believe that the reference should not necessarily be interpreted as a *ground-truth* epistemic uncertainty representation, but it defines the distribution that might be expected when compatible hypothesis spaces, loss functions and optimization algorithms are used for first-order and second-order risk minimization (i.e. for second-order risk minimization one could use the same neural network architecture and optimizer as for first-order risk minimization, and the likelihood of the former might correspond to the conjugate prior of the latter). The second-order prediction $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}; \boldsymbol{\Phi}))$ should ideally be close to the reference $q_N(\boldsymbol{\theta} \mid \boldsymbol{x})$, which can be quantified with any distance metric on probability distributions (Rachev et al., 2013).

3.2. Unregularized inner loss minimization

Let us now focus on the second research question. We aim to investigate whether second-order risk minimization methods represent epistemic uncertainty in a faithful manner by analyzing inner and outer loss minimization more in detail. We start by analyzing (2), and we first consider the case where the regularization parameter λ is zero to simplify the analysis. Similar to Bayesian model averaging, the two hypothesis spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ can be combined to a joint hypothesis space $\mathcal{H}_J$ containing $\mathcal{M} \to \mathbb{P}(\mathcal{Y})$ operators that are defined by

$$p(y \mid \boldsymbol{m}(\boldsymbol{x})) = \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}; \boldsymbol{\Phi}))}[p(y \mid \boldsymbol{\theta})]\,.$$

This distribution is known as the predictive distribution in Bayesian statistics, and for exponential family members it can be computed analytically—see Appendix B. As a result, optimization problem (2) without penalty term can be simplified to

$$\hat{\boldsymbol{\Phi}} = \arg\min_{\boldsymbol{\Phi}} \sum_{i=1}^{N} L_1(y_i, p(y \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi})))\,. \quad (5)$$

The following theorem, in which we analyse the predictive distribution for various types of models, sheds more light

[1] In practice, additional randomness might be caused by the learning algorithm, which might not be guaranteed to deliver the true minimizer.

on epistemic uncertainty quantification using optimization problem (5).

Theorem 3.2. *Let $\mathcal{C}_1$ and $\mathcal{C}_J$ be the co-domains of $\mathcal{H}_1$ and $\mathcal{H}_J$. The following properties hold when $\mathcal{H}_2$ consists of*

- *Dirichlet distributions: $\mathcal{H}_J$ is not injective, and $\mathcal{C}_1 = \mathcal{C}_J$;*
- *NIG distributions: $\mathcal{H}_J$ is not injective, and $\mathcal{C}_1 \subset \mathcal{C}_J$;*
- *Gamma distributions: $\mathcal{H}_J$ is injective, and $\mathcal{C}_1 \subset \mathcal{C}_J$.*

In the proof (see Appendix B) we have to compute the predictive distribution for the three different cases, but let us rather discuss the implications of Theorem 3.2. For the first two cases, we find that for all $p(y \mid \boldsymbol{m})$ in the codomain, multiple $\boldsymbol{m}$ exist that generate $p(y \mid \boldsymbol{m})$. Consequently, also for the risk minimizer $\bar{\boldsymbol{m}}$ induced by $\boldsymbol{\Phi}$ in (5), there exists another $\boldsymbol{m} \in \mathcal{M}$ that results in the same $p(y \mid \boldsymbol{m})$. Thus, the risk minimizer is not uniquely defined. When using the entropy of $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))$ to quantify epistemic uncertainty, one can see that a wide spectrum of values is covered, so different degrees of epistemic uncertainty can be obtained. For NIG distributions, considering second-order risk minimization brings increased flexibility, because $\mathcal{H}_1$ maps Gaussian distributions, whereas $\mathcal{H}_J$ models Student's t-distributions. For Dirichlet distributions, no flexibility gains are observed, because $\mathcal{H}_1$ and $\mathcal{H}_J$ both map Categorical distributions. Moreover, in Appendix C we show that NIG and Dirichlet distributions also differ in a second way: for the latter type of models, the optimization problem is convex when the feature mapping $\boldsymbol{\Psi}$ is a linear transformation and L_1 the negative log-likelihood, but convexity is not preserved for the former type of models.

For Gamma distributions, $\mathcal{H}_J$ is injective, so this case again differs from the other two cases. Due to the injectiveness, an identifiability problem does not appear. Second-order risk minimization has as main advantage that a more flexible class of distributions is considered for the data: negative binomial distributions instead of Poisson distributions. However, there is a one-to-one mapping between the parameters of the Gamma distributions and the parameters of the negative binomial distribution. Thus, epistemic uncertainty quantification is also arbitrary in this case, because the risk minimizer yields a Gamma distribution that results in the best negative binomial fit. Consequently, no guarantee can be given that the returned second-order distribution is close to the reference distribution, which was defined in Def. 3.1.

3.3. Unregularized outer loss minimization

We now analyze optimization problem (3) with $\lambda = 0$:

$$\hat{\boldsymbol{\Phi}} = \arg\min_{\boldsymbol{\Phi}} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))} [L_1(y_i, p(y \mid \boldsymbol{\theta}))] \; . \tag{6}$$

Similarly as for inner loss minimization, closed-form expressions exist for the second-order loss (see Appendix D and Charpentier et al. (2022)). Yet, it turns out that optimization problem (6) has different characteristics than optimization problem (5). Here comes our main result for this subsection:

Theorem 3.3. *Let $L_1 : \mathcal{Y} \times \mathbb{P}(\mathcal{Y})$ be convex in its second argument, and let $\mathcal{H}_2$ be a hypothesis space that contains universal approximators, that is, for all $p : \mathcal{X} \times \Theta \to [0, 1]$ and $\epsilon > 0$ there exists $\boldsymbol{\Phi}$ such that*

$$\sup_{\boldsymbol{\theta} \in \Theta, \, \boldsymbol{x} \in \mathcal{X}} \| p(\boldsymbol{\theta}(\boldsymbol{x})) - p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi})) \| \leq \epsilon \, .$$

A parameter set $\boldsymbol{\Phi}$ that yields $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \hat{\boldsymbol{\Phi}})) = \delta(\boldsymbol{\theta}(\boldsymbol{x}_i))$ for all $i \in \{1, \ldots, N\}$ is a minimizer of (6), where δ is the Dirac delta function on $\boldsymbol{\theta}(\boldsymbol{x}_i) \in \Theta$.

The proof can be found in Appendix E. The theorem implies that, for universal approximators such as deep neural networks, one of the risk minimizers is a Dirac distribution. Such a distribution cannot correspond to the reference distribution in Def. 3.1, since a single configuration $\boldsymbol{\theta}$ becomes the only candidate to represent $p(y \mid \boldsymbol{\theta})$. This is unwanted, because any practical machine learning setting deals with training datasets of limited size. Thus, the epistemic uncertainty should not disappear.

Theorem 3.3 differs from Theorem 3.2 in the sense that non-uniqueness of the risk minimizer is not shown explicitly. Theorem 3.3 rather indicates that one of the risk minimizers leads to unwanted Dirac distributions, and by means of experiments we will show that such distributions are indeed found with optimization routines. Let us remark that the theorem generalizes Theorem 1 by Bengs et al. (2022) in two ways: (1) it analyzes all exponential family members instead of only Dirichlet distributions, which makes the theorem applicable to a wider range of settings, and (2) it considers universal approximators that map a feature vector to a second-order distribution using $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))$, which makes the theorem practically more relevant.

3.4. Regularized risk minimization

We now take a look at the regularized risk minimization problems that were presented in (2) and (3). Needless to say, regularization is helpful to remedy some of the issues that were raised so far, but some complications arise, because different authors use different regularizers (see Table 1). We restrict our analysis to a popular group of regularizers, which enforce $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))$ to look similar to a predefined distribution $p(\boldsymbol{\theta} \mid \boldsymbol{m}_0)$:

$$R(\boldsymbol{\Phi}) = \sum_{i=1}^{N} \mathrm{d}_{\mathrm{KL}}(p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi})), p(\boldsymbol{\theta} \mid \boldsymbol{m}_0)) \, , \tag{7}$$

where d_{KL} denotes the KL-divergence between the two distributions. When $\boldsymbol{m}_0$ parameterizes the uniform distribution

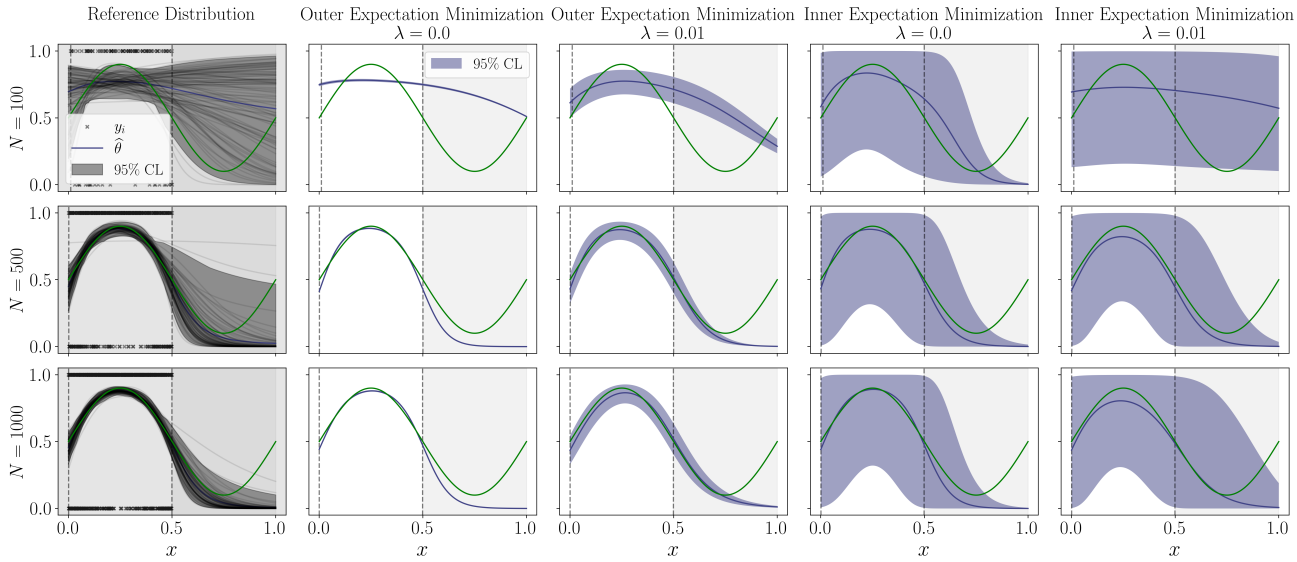


Figure 2. Binary classification experiments for training sample size $N \in \{100, 500, 1000\}$. The true θ as a function of x is shown in green. The mean estimated θ for the reference distribution and the second-order models is given in blue. Confidence bounds of the reference distribution (visualized in grey) are obtained by resampling the training data 100 times. Confidence bounds for the models trained via second-order risk minimization (visualized in purple) are obtained by the confidence intervals of the learned Beta distribution defined by its 2.5% and 97.5% quantile. The black dots denote one training dataset. See main text for experimental setup.

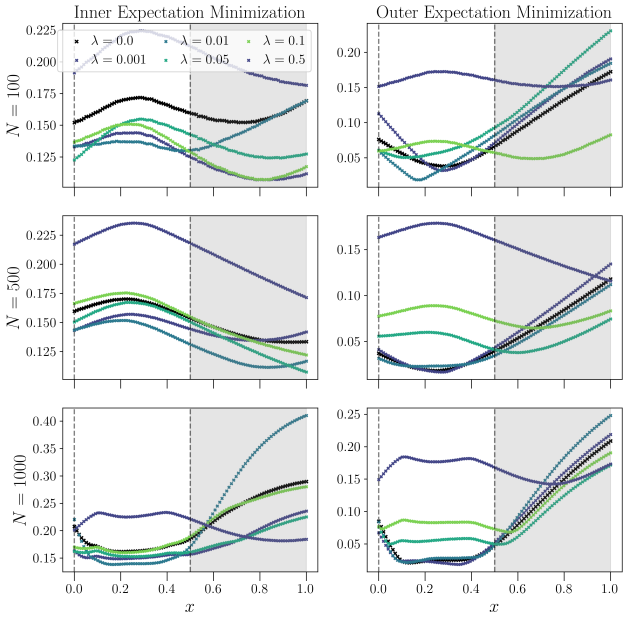


Figure 3. Empirical Wasserstein-1 distance between the reference and the estimated second-order distributions for $\lambda \in [0.0, 0.001, 0.01, 0.05, 0.1, 0.5]$. The distance is calculated for 100 equidistant points in $\mathcal{X} = [0, 1]$, by evaluating empirical and second-order distributions in the parameter space and calculating the average L_1 distance.

(on first-order distributions), minimizing KL-divergence corresponds to maximizing the entropy of $p(\boldsymbol{\theta} \mid \boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\Phi}))$, so in fact most regularizers considered in Table 1 are considered in our analysis. The main result for regularized second-order loss minimization is presented in the following theorem.

Theorem 3.4. *Consider optimization problems* (2) *and* (3) *with the regularizer defined by* (7)*. Then, for any data-generating distribution P, there exist $\lambda \geq 0$ and $\boldsymbol{x} \in \mathcal{X}$ for which the second-order distribution differs from the reference distribution in Def. 3.1.*

The proof (see Appendix F) is based on an equivalent formulation of (2) and (3) as constrained optimization problems. This reformulation reveals the role of the regularizer: it penalizes predictions with a low epistemic uncertainty. More specifically, the estimated epistemic uncertainty for all training data points is summed up, and this sum is fixed to a certain level determined by λ. In other words, the constraint in the two optimization problems defines an *epistemic uncertainty budget* that cannot be exceeded for the training data points. Now, in light of Def. 3.1 we consider for all $\boldsymbol{x} \in \mathcal{X}$ a probability distribution defined over the empirical risk minimizer $\hat{\boldsymbol{\theta}}$ and epistemic uncertainty quantified by its entropy. If λ is chosen sufficiently small, the allowed epistemic uncertainty budget will be too small to represent the epistemic uncertainty of the reference distribution, expressed by entropy or any other epistemic uncertainty measure.

The main role of the regularizer is to overcome the issues implied by Theorems 3.2 and 3.3. Let us take a look first at inner loss minimization. Theorem 3.2 implies that the two terms in (2) are not conflicting with each other when Dirichlet or NIG distributions are considered for $\mathcal{H}_2$. Among the minimizers of (5), we can simply choose the one that minimizes the regularization term. For Gamma distributions, the situation is slightly different: the two objectives will be conflicting with each other, thus adding a regularization term will yield a higher value for the initial loss function. For outer loss minimization, the Dirac delta of Theorem 3.3 will no longer be the risk minimizer.

4. Experimental analysis

The main goal of the experiments is to illustrate that our theoretical observations also arise in practice. As argued before, the faithfulness of an epistemic uncertainty representation w.r.t. Def. 3.1 can only be empirically verified via simulation studies that assume access to the reference distribution P. We performed two simulation studies, covering the classification and the regression case. For the sake of brevity, we only describe the classification case here, but the results for the regression case also confirm the theoretical results (see Appendix G). The code for reproducing the experiments is available on GitHub.[2]

4.1. Setup

We consider a binary classification setup, because the reference distribution and the second-order fit can be easily visualized in that case. We generate synthetic training data $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ of size N, with $x_i \sim \text{Unif}([0, 0.5])$, and $y_i \sim \text{Bern}\big(\theta(x_i)\big)$. The Bernoulli parameter θ is described as a function of the one-dimensional features x_i: $\theta(x_i) = 0.5 + 0.4 \sin 2\pi x_i$, thus $\theta(x_i) \in [0.5, 0.9]$. Intuitively, this looks like a very simple problem, but the aleatoric uncertainty is very high for most x, so one needs a lot of training data to estimate $\theta(x)$ with high precision.

We use the same model architecture for first and second-order risk minimization, consisting of 32 neurons and 2 fully-connected hidden layers. The two approaches only differ in the number of output neurons, which is 1 for the first-order learner, and 2 for the second-order learner. The models are trained using the Adam optimizer with a learning rate of 0.0005 and 5000 training epochs. For estimating the reference distribution in (4), we resample the training set 100 times, estimating each time the first-order risk minimizer $\hat{\theta}_k$ for $k = 1, \ldots, 100$. In this way the reference distribution of Def. 3.1 can be empirically approximated.

To train the first-order model in (1), negative log-likelihood is used as a loss function (i.e., logistic loss). The second-order models predict the parameters $\boldsymbol{m}(x) = (\alpha(x), \beta(x))^\top$ of a Beta distribution and are trained using the loss functions derived from outer and inner expectation minimization, respectively. In both cases, we use closed-form expressions for the loss functions, which avoids the need for sampling—see Appendices B and D for derivations. For the regularized models, we use negative entropy of the predicted distribution with $\lambda = 0.01$.

4.2. Results

Figure 2 shows the mean prediction as well as the 95% confidence levels for the reference model and the fitted second-order models, trained using different sample sizes. For outer loss minimization without entropy regularization, the conclusions of Theorem 3.3 are nicely reflected in the results: The narrow confidence bounds around the mean prediction indicate a distribution very close to a Dirac delta. When regularization is added, the confidence bounds become wider, but still deviate substantially from those obtained for the reference distribution, especially in the part of the input space where no data has been observed. Hence the epistemic uncertainty quantification by the second-order models is arbitrary and tightly bound to the choice of the regularization strength λ, and this phenomenon occurs over various sample sizes. For inner loss minimization, a higher epistemic uncertainty is estimated for the example run that is shown, but here, too, the estimated second-order distributions deviate substantially from the reference. However, in accordance with Theorem 3.2, we would like to emphasize that the parameters of the second-order distribution are unidentifiable for inner loss minimization. For a different run of the optimization algorithm on the same training data, we observed that the results may look very different.

In order to analytically evaluate the distance between the reference distribution and the second-order distributions, the empirical Wasserstein-1 distance was calculated per instance and for different λ values. Figure 3 shows the results: Both in regions with and without training data, a relatively high distance to the reference distribution is observed. Using different orders of magnitude for λ does not seem to significantly change that pattern.

To analyze some of these observations more deeply, we perform a second experiment in which we train the models with more epochs (up to 15 000) and 40 runs on the same training dataset. The variation between different runs comes from the uniform initialization of the network parameters. In Figure 4 the estimated second-order parameters $\hat{\alpha}$ and $\hat{\beta}$ as a function of the number of epochs for different runs are displayed. For outer loss minimization, one can see that no convergence occurred after 15 000 epochs. The estimates of α and β, averaged over the instance space, keep growing

[2] https://github.com/mkjuergens/EpistemicUncertaintyAnalysis

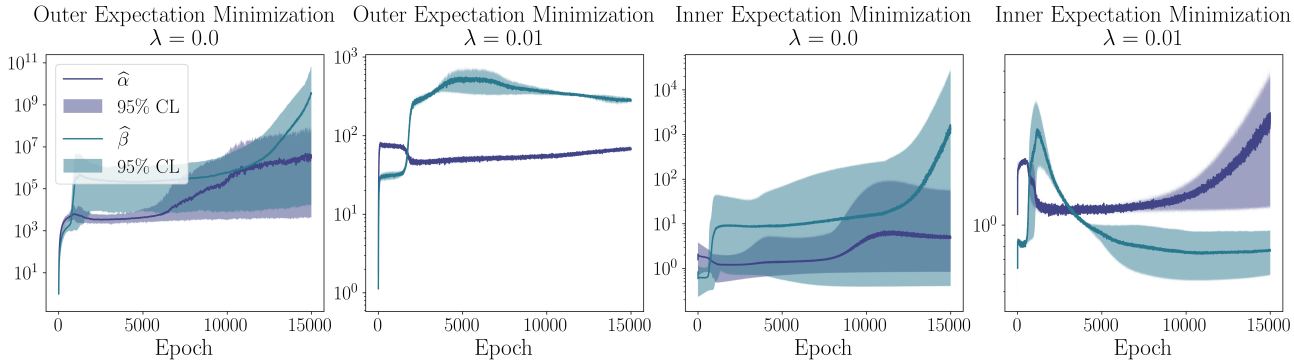


Figure 4. Behavior of the average value of the predicted parameters $\widehat{\alpha}$ and $\widehat{\beta}$, obtained by the second-order learners as a function of the number of training epochs. To obtain the mean and the confidence bounds, 40 models were trained on the same $N = 1000$ instances with different random weight intializations. The average is first taken over the instance space and individual runs are plotted. In addition, the average over different runs is also shown.

in magnitude, which is in accordance with Theorem 3.3. When regularization is added, α and β grow in a much smaller rate, thereby also solving the convergence issues, as expected due to Theorem 3.4. For inner loss minimization, one can observe large standard deviations in the estimates for α and β when the Adam optimizer is running multiple times on the same training dataset. This is expected because of Theorem 3.2. Adding some regularization substantially decreases this variation.

5. Conclusion

In this paper, we proposed the comparison with a reference distribution as an objective procedure to assess whether evidential deep learning methods represent epistemic uncertainty in a faithful manner. Using this procedure, we showed that epistemic uncertainty is in general not faithfully represented. More specifically, the regularizers that characterize evidential deep learning methods are mainly needed to stabilize the corresponding optimization problems, thereby yielding an "uncertainty budget" that cannot be exceeded. Conversely, in first-order risk minimization, regularization is mainly used to prevent overfitting by controlling the bias-variance trade-off. The resulting measures of epistemic uncertainty cannot be interpreted in a quantitative manner. In many downstream tasks, such as out-of-distribution detection, only relative interpretations of epistemic uncertainty are needed. This probably explains the strong performance of evidential deep learning methods on such tasks. We hypothesize that the OOD scores reported by evidential deep learning methods could be interpreted as density estimates in feature space. This observation was made by Grathwohl et al. (2020); Charpentier et al. (2022), but we believe that more research is needed to draw final conclusions of that kind.

Impact statement

Machine learning systems are increasingly prevalent in high-stake application domains with a direct impact on humans, such as medical diagnostics, legal decisions, self-driving cars, fraud detection, natural disaster forecasting, etc. The profound consequences of inaccurate predictions in these areas emphasize the necessity of assessing the certainty of machine-learned forecasts. The current upswing in research on uncertainty quantification in the machine learning community underscores the recognition of this crucial aspect.

Taking center stage is Evidential Deep Learning (EDL), a method gaining prominence in safety-critical domains. However, despite its growing utilization, the operational principles of EDL remain less than fully understood. Rather than solely focusing on developing superior techniques for specific datasets, we advocate for a meticulous examination of existing methods. Unraveling their limitations is not just an academic exercise, but it serves as a practical guide for decision-makers. Understanding when to employ these methods and when to exercise caution is paramount in navigating the intricate terrain of high-stakes predictions.

Acknowledgements

M.J. and W.W. received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

References

A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Proc. NIPS, 30th Advances in Neural Information Processing Systems*, pages 5574–5584, 2017.

E. Hüllermeier and W. 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.

S. Depeweg, J.M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proc. ICML, 35th International Conference on Machine Learning*, pages 1184–1193, 2018.

Alex Graves. Practical variational inference for neural networks. In *Proc. NIPS, 24th Advances in Neural Information Processing Systems*, page 2348–2356, 2011.

M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proc. NeurIPS, 31st Conference on Neural Information Processing Systems*, pages 3183–3193, 2018.

A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. In *Proc. NeurIPS, 31st Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

A. Malinin and M. Gales. Reverse KL-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Proc. NeurIPS, 32nd Advances in Neural Information Processing Systems*, pages 14520–14531, 2019.

B. Charpentier, D. Zügner, and S. Günnemann. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. In *Proc. NeurIPS, 33rd Neural Information Processing Systems*, volume 33, pages 1356–1367, 2020.

D. Huseljic, B. Sick, M. Herde, and D. Kottke. Separation of aleatoric and epistemic uncertainty in deterministic deep neural networks. In *Proc. ICPR, 25th International Conference on Pattern Recognition*, pages 9172–9179. IEEE, 2020.

Andrey Malinin, Sergey Chervontsev, Ivan Provilkov, and Mark Gales. Regression prior networks. *arXiv preprint arXiv:2006.11590*, 2020.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Proc. NeurIPS, 33rd Advances in Neural Information Processing Systems*, volume 33, pages 14927–14937, 2020.

Huan Ma, Zongbo Han, Changqing Zhang, H. Fu, Joey Tianyi Zhou, and Qinghua Hu. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. In *Proc. NeurIPS, 34th Advances in Neural Information Processing Systems*, pages 6881–6893, 2021.

Theodoros Tsiligkaridis. Information aware max-norm Dirichlet networks for predictive uncertainty estimation. *Neural Networks*, 135:105–114, 2021.

A.-Kathrin Kopetzki, B. Charpentier, D. Zügner, S. Giri, and S. Günnemann. Evaluating robustness of predictive uncertainty estimation: Are Dirichlet-based models reliable? In *Proc. ICML, 38th International Conference on Machine Learning*, pages 5707–5718, 2021.

Wentao Bao, Qi Yu, and Yu Kong. Evidential deep learning for open set action recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13329–13338, 2021.

Zhijian Liu, Alexander Amini, Sibo Zhu, Sertac Karaman, Song Han, and Daniela L. Rus. Efficient and robust lidar-based end-to-end navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, page 13247–13254. IEEE Press, 2021.

Bertrand Charpentier, Oliver Borchert, Daniel Zugner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian predictive uncertainty for exponential family distributions. In *International Conference on Learning Representations*, 2022.

Dongpin Oh and Bonggun Shin. Improving evidential deep learning via multi-task learning. In *Proc. AAAI, 36th Conference on Artificial Intelligence*, number 7, pages 7895–7903, 2022.

Deep Shankar Pandey and Qi Yu. Evidential conditional neural processes. In *Proc. AAAI, 37th Conference on Artificial Intelligence*, pages 9389–9397, 2023.

Nikita Kotelevskii, Samuel Horváth, Karthik Nandakumar, Martin Takáč, and Maxim Panov. Dirichlet-based uncertainty quantification for personalized federated learning with improved posterior networks. *arXiv preprint arXiv:2312.11230*, 2023.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In *Proc. NeurIPS, 35th Advances in Neural Information Processing Systems*, 2022.

Nis Meinert, Jakob Gawlikowski, and Alexander Lavin. The unreasonable effectiveness of deep evidential regression. In *Proc. AAAI, 37th Conference on Artificial Intelligence*, pages 9134–9142, 2023.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *Proc. ICML, 40th International Conference on Machine Learning*, 2023.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Proc. NeurIPS, 35th Advances in Neural Information Processing Systems*, pages 32598–32611, 2022.

Younghyun Park, Wonjeong Choi, Soyeong Kim, Dong-Jun Han, and Jaekyun Moon. Active learning for object detection with evidential deep learning and hierarchical uncertainty aggregation. In *International Conference on Learning Representations*, 2023.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions of Machine Learning Research*, 2023.

P.G. Bissiri, C.C. Holmes, and S.G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

Nis Meinert and Alexander Lavin. Multivariate deep evidential regression. *arXiv preprint arXiv:2104.06135*, 2022.

Lennart Bramlage, Michelle Karg, and Cristóbal Curio. Plausible uncertainties for human pose regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15133–15142, 2023.

A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013. ISBN 9781439840955.

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 *Proc. UAI, 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2282–2292, 2023.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, and Sepp Hochreiter. Introducing an improved information-theoretic measure of predictive uncertainty. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.

Yusuf Sale, Paul Hofman, Lisa Wimmer, Eyke Hüllermeier, and Thomas Nagler. Second-order uncertainty quantification: Variance-based measures. *arXiv preprint arXiv:2401.00276*, 2023a.

Ruxiao Duan, Brian Caffo, Harrison X Bai, Haris I Sair, and Craig Jones. Evidential uncertainty quantification: A variance-based perspective. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2132–2141, 2024.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-order uncertainty quantification: A distance-based approach. *arXiv preprint arXiv:2312.00995*, 2023b.

R. Martin. Fisher's underworld and the behavioral-statistical reliability balance in scientific inference. *arXiv preprint arXiv:2312.14912*, 2023.

Svetlozar T Rachev, Lev B Klebanov, Stoyan V Stoyanov, and Frank Fabozzi. *The Methods of Distances in the Theory of Probability and Statistics*, volume 10. Springer, 2013.

Will Grathwohl, Kuan-Chieh Jackson Wang, Jörn-Henrik Jacobsen, David Kristjanson Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model, and you should treat it like one. In *International Conference on Learning Representations*, 2020.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787.

A. Basic properties of exponential family members

In this appendix we review some important formulas for exponential family members.

Dirichlet distribution

Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ be the parameters of the Categorical distribution and let $\boldsymbol{m} = (m_1, \ldots, m_K) \in \mathbb{R}_+^K$ be the parameters of the Dirichlet distribution with probability density function:

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{m}) = \frac{1}{B(\boldsymbol{m})} \prod_{k=1}^{K} \theta_k^{m_k - 1} \tag{8}$$

with $B(\boldsymbol{m})$ the Beta function, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K) \in \triangle^K$ and $\triangle^K$ the probability $(K-1)$-simplex.

The entropy of the Dirichlet distribution is given by:

$$\mathrm{H}(p(\boldsymbol{\theta} \,|\, \boldsymbol{m})) = \log B(\boldsymbol{m}) + (m_0 - K)\psi(m_0) - \sum_{k=1}^{K} (m_k - K)\psi(m_k)\,,$$

with ψ the digamma function and $m_0 = \sum_{k=1}^{K} m_k$.

Normal-inverse Gamma distribution

Let $\boldsymbol{\theta} = (\mu, \sigma^2)$ be the parameters of the Gaussian distribution and let $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta) \in \mathbb{R} \times \mathbb{R}_+ \times [1, \infty) \times \mathbb{R}_+$ be the parameters of the normal-inverse-Gamma (NIG) distribution with probability density function:

$$p(\mu, \sigma^2 \,|\, \gamma, \nu, \alpha, \beta) = \frac{\sqrt{\nu}}{\sqrt{2\pi\sigma^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma^2}\right)^{(\alpha+1)} \exp\left\{-\frac{2\beta + \nu(\mu - \gamma)^2}{2\sigma^2}\right\}, \tag{9}$$

The entropy of the normal-inverse-Gamma distribution is given by:

$$\mathrm{H}(p(\mu, \sigma^2 \,|\, \gamma, \nu, \alpha, \beta)) = \frac{1}{2} + \log\left(\sqrt{2\pi}\beta^{\frac{3}{2}}\Gamma(\alpha)\right) - \frac{\log \nu}{2} + \alpha - \left(\alpha + \frac{3}{2}\right)\psi(\alpha)\,,$$

with Γ the Gamma function.

Gamma distribution

Let θ be the parameter of a Poisson distribution and let $\boldsymbol{m} = (\alpha, \beta) \in \mathbb{R}_+^2$ be the shape and rate parameters of the Gamma distribution with probability density function:

$$p(\theta \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}\,.$$

The entropy of the Gamma distribution is given by:

$$\mathrm{H}(p(\theta \,|\, \alpha, \beta)) = \alpha - \ln\beta + \ln\Gamma(\alpha) + (1 - \alpha)\psi(\alpha)\,.$$

B. Proof of Theorem 3.2

In this proof we derive

$$p(y \,|\, \boldsymbol{m}(\boldsymbol{x})) \quad = \quad \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \,|\, \boldsymbol{m}(\boldsymbol{x}; \boldsymbol{\Phi}))}\, p(y \,|\, \boldsymbol{\theta}) = \int p(y \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \boldsymbol{m}(\boldsymbol{x}))\, d\boldsymbol{\theta} \tag{10}$$

for different members of the exponential family.

When Dirichlet distributions are chosen for $\mathcal{H}_2$, the predictive distribution is given by[3]:

$$\begin{aligned}
p(y=k \,|\, \boldsymbol{m}=(m_1,...,m_k)) &= \int_{\triangle^K} p(y=k \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \boldsymbol{m})\, d\boldsymbol{\theta} \\
&= \int_{\triangle^K} \frac{\Gamma(\sum_{j=1}^K m_j)}{\prod_{j=1}^K \Gamma(m_j)} \prod_{j=1}^K \theta_j^{m_j-1} \theta_k \, d\boldsymbol{\theta} \\
&= \frac{\Gamma(\sum_{j=1}^K m_j)}{\prod_{j=1}^K \Gamma(m_j)} \int_{\triangle^K} \prod_{j=1}^K \theta_j^{m_j-1} \theta_k \, d\boldsymbol{\theta} \\
&= \frac{\Gamma(\sum_{j=1}^K m_j)}{\prod_{j=1}^K \Gamma(m_j)} \int_{\triangle^K} \mathrm{d}\boldsymbol{\theta} \prod_{j=1}^K \theta_j^{\mathbb{I}(j=k)+m_j-1} \\
&= \frac{\Gamma(\sum_{j=1}^K m_j)}{\prod_{j=1}^K \Gamma(m_j)} \frac{\prod_{j=1}^K \Gamma(\mathbb{I}(j=k)+m_j)}{\Gamma(1+\sum_{j=1}^K m_j)} \\
&= \frac{m_k}{\sum_{j=1}^K m_j} = \vartheta_k \,.
\end{aligned}$$

This is the probability mass function of the Categorical distribution with parameters $\vartheta_1, \vartheta_2, ..., \vartheta_k$. In this case $\mathcal{H}_J$ does not contain injective $\mathcal{M} \to \mathbb{P}(\mathcal{Y})$ operators, and $\mathcal{C}_1 = \mathcal{C}_J$ because the two hypothesis spaces contain Categorical distributions.

When normal-inverse Gamma distributions are chosen for $\mathcal{H}_2$, the predictive distribution is derived by Amini et al. (2020):

$$\begin{aligned}
p(y \,|\, \boldsymbol{m}=(\gamma,\nu,\alpha,\beta)) &= \int p(y \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \boldsymbol{m})\, d\boldsymbol{\theta} \\
&= \int_{\sigma^2=0}^{\infty} \int_{\mu=-\infty}^{\infty} p(y \,|\, \mu, \sigma^2)\, p(\mu, \sigma^2 \,|\, \gamma, \nu, \alpha, \beta)\, d\sigma^2\, d\mu \\
&= \cdots \\
&= \frac{\Gamma(1/2+\alpha)}{\Gamma(\alpha)} \sqrt{\frac{\nu}{\pi}} (2\beta(1+\nu))^{\alpha} (\nu(y-\gamma^2) + 2\beta(1+\nu))^{(-\frac{1}{2}+\alpha)} \\
&= \mathrm{St}\left(y;\, \gamma,\, \frac{\beta(1+\nu)}{\nu\alpha}, 2\alpha\right) ,
\end{aligned}$$

where $\mathrm{St}(y; \mu_{\mathrm{St}}, \sigma^2_{\mathrm{St}}, \nu_{\mathrm{St}})$ stands for Student's-t distribution with location μ_{St}, scale σ^2_{St} and ν_{St} degrees of freedom. $\mathcal{H}_J$ does not contain injective $\mathcal{M} \to \mathbb{P}(\mathcal{Y})$ operators, and $\mathcal{C}_1 \subset \mathcal{C}_J$, because the Gaussian distribution is a special case of the Student's-t distribution.

When Gamma distributions are chosen for $\mathcal{H}_2$, the predictive distribution is given by[4]:

$$\begin{aligned}
p(y=k \,|\, \boldsymbol{m}=(\alpha,\beta)) &= \int p(y=k \,|\, \theta)\, p(\theta \,|\, \boldsymbol{m})\, d\theta \\
&= \int \left[\frac{\theta^k e^{-\theta}}{k!}\right] \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}\right] d\theta \\
&= \frac{\beta^\alpha}{\Gamma(k+1)\,\Gamma(\alpha)} \int \theta^{\alpha-1+k} e^{-(\beta+1)\theta}\, d\theta \\
&= \frac{\beta^\alpha}{\Gamma(k+1)\,\Gamma(\alpha)} \frac{\Gamma(k+\alpha)}{(\beta+1)^{k+\alpha}} \\
&= \frac{\Gamma(k+\alpha)}{\Gamma(k+1)\,\Gamma(\alpha)} \left[\frac{\beta}{\beta+1}\right]^\alpha \left[\frac{1}{\beta+1}\right]^k \\
&= \binom{k+\alpha-1}{k} \left[\frac{\beta}{\beta+1}\right]^\alpha \left[\frac{1}{\beta+1}\right]^k .
\end{aligned}$$

[3]For this derivation we were inspired by https://stephentu.github.io/writeups/dirichlet-conjugate-prior.pdf

[4]For this derivation we were inspired by https://people.stat.sc.edu/Hitchcock/stat535slidesday18.pdf

This is the probability mass function of a negative binomial distribution with number of successes $r = \alpha$ and success probability $p = \beta/(\beta+1)$. The p.m.f. is an injective $\mathcal{M} \to \mathbb{P}(\mathcal{Y})$ operator, and $\mathcal{C}_1 \subset \mathcal{C}_J$ because the Poisson distribution is a special case of the negative binomial distribution.

C. Additional theoretical results for inner loss minimization

In this appendix we further analyze the convexity properties of inner loss minimization without regularization, using the negative log-likelihood as first-order loss function. Let us first recall the standard definition of convexity.

Definition C.1. A function $f : \mathbb{R}^D \to \mathbb{R}$ is convex if

$$f(\rho \boldsymbol{x} + (1-\rho)\boldsymbol{z}) \leq \rho f(\boldsymbol{x}) + (1-\rho) f(\boldsymbol{z}) \tag{11}$$

for all $\boldsymbol{x}, \boldsymbol{z}$ in the domain of f and all $\rho \in [0,1]$. Furthermore, if (11) holds as a strict inequality for all $\boldsymbol{x}, \boldsymbol{z}$ in the domain of f and all $\rho \in [0,1]$, then f is called strictly convex.

The following theorem presents formal results for inner expectation maximization using linear models as underlying model class for the neural network.

Theorem C.2. *The following statements hold concerning the convexity of the risk minimization problem in* (5)*.*

- *Let $\mathcal{H}_2$ consist of Dirichlet distributions, let L_1 be the negative log-likelihood, and let $\boldsymbol{\Psi}$ be a linear transformation, then* (5) *is convex.*
- *Let $\mathcal{H}_2$ consist of normal-inverse Gamma distributions, let L_1 be the negative log-likelihood, and let $\boldsymbol{\Psi}$ be a linear transformation, then* (5) *is not convex.*

Proof. Let us start with proving the first statement. With Dirichlet distributions for $\mathcal{H}_2$ and $\boldsymbol{\Psi}$ a linear transformation, (3.2) becomes

$$p(y = k \,|\, \boldsymbol{m}) = \frac{\mathrm{e}^{\boldsymbol{w}_k^\top \boldsymbol{\Psi}(\boldsymbol{x})}}{\sum_{j=1}^K \mathrm{e}^{\boldsymbol{w}_j^\top \boldsymbol{\Psi}(\boldsymbol{x})}} \,. \tag{12}$$

This is nothing more than the class of functions considered in multinomial logistic regression. The per-instance risk becomes

$$L_1(y, p(y \,|\, \boldsymbol{m}(\boldsymbol{x}_i))) = -\log\left(\frac{\mathrm{e}^{\boldsymbol{w}_y^\top \boldsymbol{\Psi}(\boldsymbol{x})}}{\sum_{j=1}^K \mathrm{e}^{\boldsymbol{w}_j^\top \boldsymbol{\Psi}(\boldsymbol{x})}}\right) = \log\left(\sum_{j=1}^K \mathrm{e}^{\boldsymbol{w}_j^\top \boldsymbol{\Psi}(\boldsymbol{x})}\right) - \boldsymbol{w}_y^\top \boldsymbol{\Psi}(\boldsymbol{x}) \,.$$

A function of type log-sum-exponential is a convex function (Boyd and Vandenberghe, 2004). The second part of the risk is linear, so the two components together are convex.

Let us now prove the second statement by evaluating (11) for

$$f(\gamma, \nu, \alpha, \beta) = -\log \mathrm{St}(y; \gamma, \sigma^2_{\mathrm{St}}(\nu, \alpha, \beta), 2\alpha).$$

Since $\exists\, \{(\nu, \beta), (\nu', \beta')\}$ with $(\nu, \beta) \neq (\nu', \beta')$ and $\sigma^2_{\mathrm{St}}(\nu, \alpha, \beta) = \sigma^2_{\mathrm{St}}(\nu', \alpha, \beta')\ \forall \alpha$ for these points the r.h.s. of (11) can be written as:

$$\rho f(\gamma, \nu, \alpha, \beta) + (1-\rho) f(\gamma, \nu', \alpha, \beta') = -\log \mathrm{St}(y; \gamma, \sigma^2_{\mathrm{St}}(\nu, \alpha, \beta), 2\alpha) \,.$$

Next, we note that $\hat{\sigma}^2_{\mathrm{St}} = \arg\max_{\sigma^2_{\mathrm{St}}} \mathrm{St}(y; \mu_{\mathrm{St}}, \sigma^2_{\mathrm{St}}, \nu_{\mathrm{St}}) = (x - \mu_{\mathrm{St}})^2$. Let $\hat{\sigma}^2_{\mathrm{St}} = \sigma^2_{\mathrm{St}}(\hat{\nu}, \hat{\alpha}, \hat{\beta}) = \sigma^2_{\mathrm{St}}(\hat{\nu}', \hat{\alpha}, \hat{\beta}')$, then the l.h.s. of (11) is always greater or equal than the r.h.s. for these points,

$$-\log \mathrm{St}\Big(y; \gamma, \sigma^2_{\mathrm{St}}(\rho\hat{\nu} + (1-\rho)\hat{\nu}', \hat{\alpha}, \rho\hat{\beta} + (1-\rho)\hat{\beta}'), 2\hat{\alpha}\Big) \geq -\log \mathrm{St}\big(y; \gamma, \hat{\sigma}^2_{\mathrm{St}}, 2\hat{\alpha}\big) \,.$$

If σ^2_{St} is an affine function both sides are equal. However, if $\boldsymbol{\Psi}$ is a linear transformation and $(\gamma, \nu, \alpha, \beta)$ are defined as output neurons with linear and exponential activations (see Section 2) or their related formulations with the softplus function, the inequality becomes strict. Since (11) has to hold for any points in the domain of f, we conclude that $-\log \mathrm{St}$ is not convex if $\boldsymbol{\Psi}$ is a linear transformation. □

The proof highlighted that the predictive distribution corresponds to a traditional multinomial logistic regression model when Dirichlet distributions are considered as model class, linear transformations as feature learning layers, and the negative log-likelihood as L_1. It is well known that risk minimization in multinomial logistic regression is not strictly convex[5]. Concretely, subtracting a fixed vector $\boldsymbol{w}_0$ from each $\boldsymbol{w}_k$ does not affect (12):

$$p(y = k \,|\, \boldsymbol{m}) = \frac{\mathrm{e}^{(\boldsymbol{w}_k - \boldsymbol{w}_0)^\top \boldsymbol{x}}}{\sum_{j=1}^K \mathrm{e}^{(\boldsymbol{w}_j - \boldsymbol{w}_0)^\top \boldsymbol{x}}} = \frac{\mathrm{e}^{\boldsymbol{w}_k^\top \boldsymbol{x}} \mathrm{e}^{-\boldsymbol{w}_0^\top \boldsymbol{x}}}{\sum_{j=1}^K \mathrm{e}^{\boldsymbol{w}_j^\top \boldsymbol{x}} \mathrm{e}^{-\boldsymbol{w}_0^\top \boldsymbol{x}}} = \frac{\mathrm{e}^{\boldsymbol{w}_k^\top \boldsymbol{x}}}{\sum_{j=1}^K \mathrm{e}^{\boldsymbol{w}_j^\top \boldsymbol{x}}} \,.$$

In other words, subtracting $\boldsymbol{w}_0$ from every $\boldsymbol{w}_k$ does not affect our hypothesis space at all. The hypothesis space is overparameterized, in the sense that there are multiple parameter settings that give rise to exactly the same hypothesis function, which maps $\boldsymbol{x}$ to the predictions. If the risk is minimized for some values of $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k\}$, then it is also minimized for $\{\boldsymbol{w}_1 - \boldsymbol{w}_0, \ldots, \boldsymbol{w}_k - \boldsymbol{w}_0\}$. Thus, the risk minimizer is not unique. However, (5) is still convex, and thus gradient descent will not run into local optima problems. But the Hessian is non-invertible, which causes a straightforward implementation of Newton's method to run into numerical problems.

In multinomial logistic regression, this overparameterization is avoided by eliminating one of the vectors $\boldsymbol{w}_k$ (typically $\boldsymbol{w}_K$, but any k could be chosen). One then optimizes over the remaining $K - 1$ parameter vectors. However, for the setup considered in this paper, such a simplification is not possible, because it would fix epistemic uncertainty to a predefined level.

D. Closed-form outer loss minimization for exponential family members

In this appendix we show how the empirical risk in outer loss minimization can be simplified. For exponential family members, the log-likelihood can be written as

$$\log p(y \,|\, \boldsymbol{\theta}) = \log h(y) + \boldsymbol{\theta}^\top \boldsymbol{u}(y) - A(\boldsymbol{\theta}) \,,$$

where $\boldsymbol{\theta} \in \mathbb{R}^L$ is the natural parameter, $h : \mathbb{R} \to \mathbb{R}$ is the carrier or base measure, $A : \mathbb{R}^L \to \mathbb{R}$ the log-normalizer and $u : \mathbb{R} \to \mathbb{R}^L$ the sufficient statistics.

Its expectation under the prior distribution then becomes

$$\mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \,|\, \boldsymbol{m})} \log p(y \,|\, \boldsymbol{\theta}) = \log h(y) + \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \,|\, \boldsymbol{m})} \Big[\boldsymbol{\theta}^\top \boldsymbol{u}(y) \Big] - \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \,|\, \boldsymbol{m})} A(\boldsymbol{\theta}) \,,$$

with $\mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \,|\, \boldsymbol{m})} \boldsymbol{\theta} = \boldsymbol{r}$.

Thus, when $\mathcal{H}_2$ contains Dirichlet distributions, the expected log-likelihood function becomes

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \,|\, \boldsymbol{m})} \log p(y \,|\, \boldsymbol{\theta}) &= 0 + \mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{m})} \left[\sum_{i=1}^K 1_{[y=i]} \log(\theta_i) \right] - 0 \\
&= \sum_{i=1}^K \mathbb{I}_{[y=i]} \, \mathbb{E}_{\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{m})} \left[\log(\theta_i) \right] \\
&= \sum_{i=1}^K \mathbb{I}_{[y=i]} \left(\psi(m_i) - \psi\left(\sum_{j=1}^K m_j \right) \right) , \\
&= \psi(m_y) - \psi\left(\sum_{j=1}^K m_j \right) , \qquad (13)
\end{aligned}$$

with $\mathbb{I}$ the indicator function, which returns one when its argument is true and zero otherwise. We used for the last inequality that the logarithmic moments[6] of a Dirichlet distribution are $\psi(m_i) - \psi(\sum_{j=1}^K m_j)$. In a similar vain, Charpentier et al. (2022) derived that the expected log-likelihood is given by

$$\mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \,|\, \boldsymbol{m})} \log p(y \,|\, \boldsymbol{\theta}) = \frac{1}{2} \left(-\frac{\alpha}{\beta}(y - \gamma)^2 - \frac{1}{\nu} + \psi(\alpha) - \log \beta - \log 2\pi \right) \qquad (14)$$

[5] http://deeplearning.stanford.edu/tutorial/supervised/SoftmaxRegression/

[6] https://en.wikipedia.org/wiki/Dirichlet_distribution#Entropy

when $\mathcal{H}_2$ contains inverse-normal Gamma distributions.

For Gamma distributions Charpentier et al. (2022) find

$$\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m})} \log p(y\,|\,\boldsymbol{\theta}) = (\psi(\alpha) - \log\beta)y - \frac{\alpha}{\beta} - \sum_{k=1}^{y} \log k\,.$$

E. Proof of Theorem 3.2

The empirical risk is given by

$$\mathcal{R}(\boldsymbol{\Phi}) = \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\boldsymbol{\Phi}))} L_1(y_i, p(y\,|\,\boldsymbol{\theta}))\,.$$

When the first-order loss L_1 is convex, it follows from Jensen's inequality that

$$\mathcal{R}(\boldsymbol{\Phi}) \geq \sum_{i=1}^{N} L_1(y_i, \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\Phi))} p(y\,|\,\boldsymbol{\theta}))\,.$$

Let $\widetilde{\boldsymbol{\Phi}}$ be the minimizer of the right-hand side, then it holds for all $\boldsymbol{\Phi}$ that

$$\mathcal{R}(\boldsymbol{\Phi}) \geq \sum_{i=1}^{N} L_1(y_i, \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\Phi))} p(y\,|\,\boldsymbol{\theta})) \tag{15}$$

$$\geq \sum_{i=1}^{N} L_1(y_i, \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\tilde{\boldsymbol{\Phi}}))} p(y\,|\,\boldsymbol{\theta}))\,. \tag{16}$$

In what follows let

$$\tilde{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\widetilde{\Phi}))} \boldsymbol{\theta}\,.$$

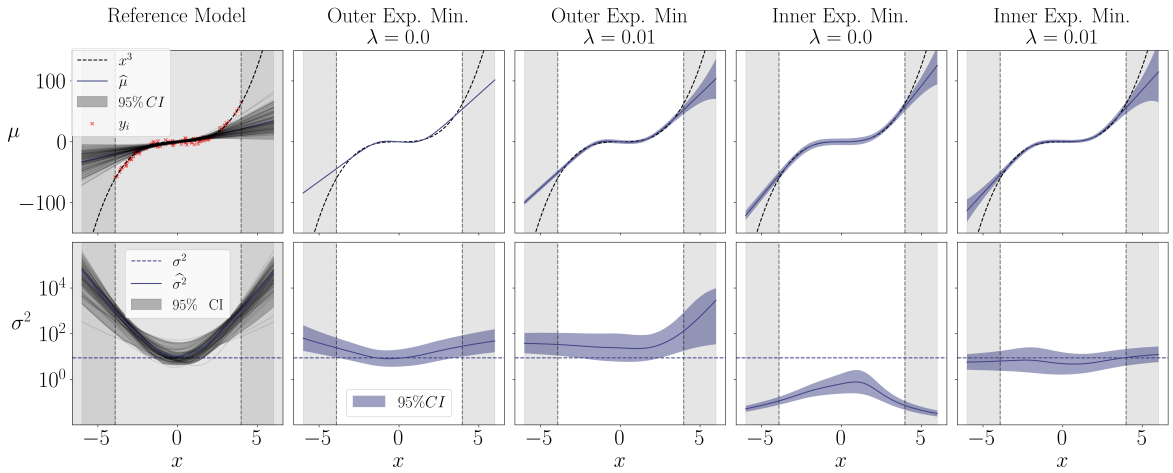
For all $i \in \{1, \ldots, N\}$, let $\boldsymbol{\Phi}'$ denote the parameterization that corresponds to $p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\boldsymbol{\Phi}')) = \delta(\tilde{\boldsymbol{\theta}}(\boldsymbol{x}_i))$, where δ is the Dirac delta function. Such a parameterization always exists, because we assumed a hypothesis space of universal approximators. Furthermore, note that

$$\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\boldsymbol{\Phi}'))} \boldsymbol{\theta} \;=\; \tilde{\boldsymbol{\theta}}(\boldsymbol{x}_i)\,.$$

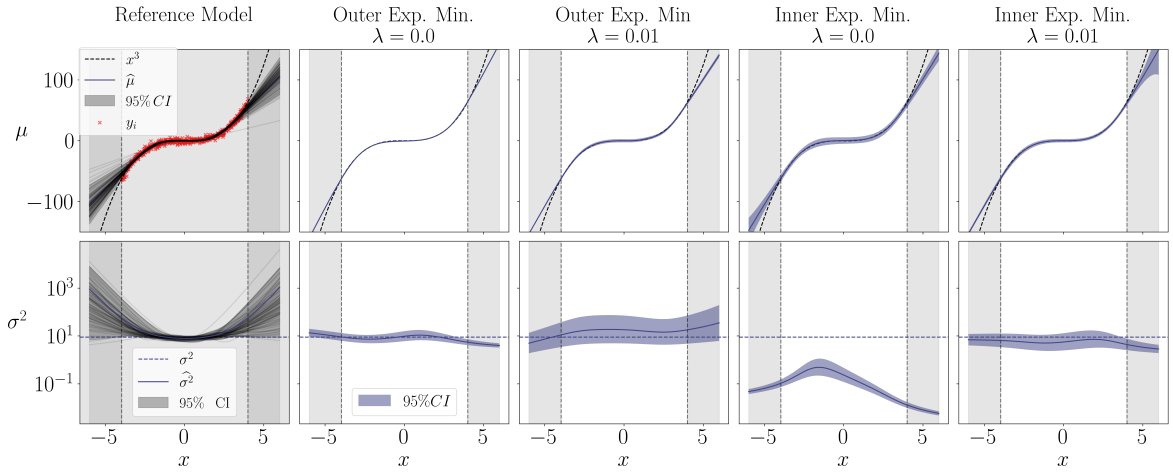
Then,

$$\begin{aligned}
\mathcal{R}(\boldsymbol{\Phi}') &= \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\boldsymbol{\Phi}'))} L_1\left(y_i, p(y\,|\,\boldsymbol{\theta})\right) \\
&= \sum_{i=1}^{N} L_1\left(y_i, \tilde{\boldsymbol{\theta}}(\boldsymbol{x}_i)\right) \\
&= \sum_{i=1}^{N} L_1\left(y_i, \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\tilde{\boldsymbol{\Phi}}))} p(y\,|\,\boldsymbol{\theta})\right).
\end{aligned}$$

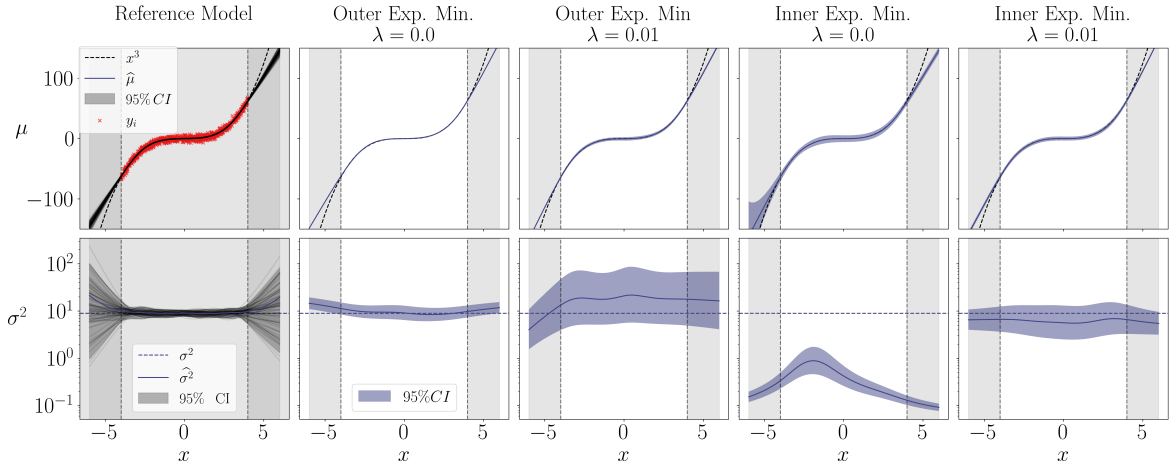
Combining this last observation with inequality (15) lets us conclude that for all $\boldsymbol{\Phi}$ it holds that $\mathcal{R}(\boldsymbol{\Phi}) \geq \mathcal{R}(\boldsymbol{\Phi}')$. Thus, $\boldsymbol{\Phi}'$ must be a solution of (6).



(a) $N = 100$



(b) $N = 500$



(c) $N = 1000$

Figure 5. Regression experiments for training sample size $N \in \{100, 500, 1000\}$. The reference model learns the parameters $\theta = (\mu, \sigma)$ of the underlying normal distribution. The true underlying mean $\mu = x^3$ and variance $\sigma^2 = 9$, are visualised in separate rows together with the mean predictions (in blue) and obtained confidence intervals. Confidence bounds for the predicted parameters by the reference model are obtained by resampling the training data 100 times. Confidence bounds for μ and σ^2 learned by second-order risk minimization (visualized in purple) are obtained by the quantiles of the normal distribution $\mathcal{N}(\widehat{\mu}, \frac{\widehat{\beta}}{(\widehat{\alpha}-1)\widehat{\nu}})$, and of the Inverse-Gamma distribution $\Gamma^{-1}(\widehat{\alpha}, \widehat{\beta})$, respectively. The red dots denote one training dataset. See main text for experimental setup.

F. Proof of Theorem 3.3

The proof is based on the fact that (2) and (3) can be reformulated as constrained optimization problems:

$$\begin{aligned}\hat{\boldsymbol{\Phi}} &= \arg\min_{\boldsymbol{\Phi}} \sum_{i=1}^{N} L_1\big(y_i, \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\boldsymbol{\Phi}))}[p(y\,|\,\boldsymbol{\theta})]\big) \\ &\quad \text{s.t. } R(\boldsymbol{\Phi}) \leq t \\ \hat{\boldsymbol{\Phi}} &= \arg\min_{\boldsymbol{\Phi}} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}\,|\,\boldsymbol{m}(\boldsymbol{x}_i;\boldsymbol{\Phi}))}[L_1(y_i, p(y\,|\,\boldsymbol{\theta}))] \\ &\quad \text{s.t. } R(\boldsymbol{\Phi}) \leq t\end{aligned}$$

with t being a bijection of λ.

To prove this equivalence[7], let us introduce the shorthand notation $\mathcal{R}(\boldsymbol{\Phi})$ for the empirical risk. In essence we intend to show equivalence between the *penalty problem*

$$\arg\min_{\boldsymbol{\Phi}} \mathcal{R}(\boldsymbol{\Phi}) + \lambda R(\boldsymbol{\Phi})$$

and the *constrained problem*

$$\arg\min_{\boldsymbol{\Phi}} \mathcal{R}(\boldsymbol{\Phi}) \qquad \text{s.t. } R(\boldsymbol{\Phi}) \leq t\,.$$

First, assume that $\hat{\boldsymbol{\Phi}}$ is a solution of the penalty problem, which implies for all $\boldsymbol{\Phi}$ that

$$\mathcal{R}(\boldsymbol{\Phi}) + \lambda R(\boldsymbol{\Phi}) \geq \mathcal{R}(\hat{\boldsymbol{\Phi}}) + \lambda R(\hat{\boldsymbol{\Phi}})\,. \tag{17}$$

Hence for all $\boldsymbol{\Phi}$ such that $R(\boldsymbol{\Phi}) \leq t = R(\hat{\boldsymbol{\Phi}})$ it holds that $\mathcal{R}(\boldsymbol{\Phi}) \geq \mathcal{R}(\hat{\boldsymbol{\Phi}})$. Thus, $\hat{\boldsymbol{\Phi}}$ is also a solution of the constrained problem when $t = R(\hat{\boldsymbol{\Phi}})$.

Conversely, assume that $\boldsymbol{\Phi}^{**}$ is a solution of the constrained problem with $t = R(\hat{\boldsymbol{\Phi}})$, which implies

$$R(\boldsymbol{\Phi}^{**}) \leq t = R(\hat{\boldsymbol{\Phi}}) \qquad \text{and} \qquad \mathcal{R}(\boldsymbol{\Phi}^{**}) \leq \mathcal{R}(\hat{\boldsymbol{\Phi}})\,.$$

Since $\hat{\boldsymbol{\Phi}}$ is a solution of the penalty problem, inequality (17) holds. Combining those two inequalities leads to

$$\mathcal{R}(\boldsymbol{\Phi}) + \lambda R(\boldsymbol{\Phi}) \geq \mathcal{R}(\boldsymbol{\Phi}^{**}) + \lambda R(\boldsymbol{\Phi}^{**})\,,$$

for all $\boldsymbol{\Phi}$. This allows us to say that $\boldsymbol{\Phi}^{**}$ is also a solution of the penalty problem. As a result, we have proven that $\hat{\boldsymbol{\Phi}} = \boldsymbol{\Phi}^{**}$, so the two optimization problems are equivalent.

The reformulation as a constrained optimization problem allows us to interpret the regularizer $R(\boldsymbol{\Phi})$ as an epistemic uncertainty budget that cannot be exceeded. As discussed in the main paper, when t or λ are chosen inadequately, the obtained solution will differ substantially from the reference distribution.

G. Additional experiments

For the regression experiments we use the same setup as introduced in (Amini et al., 2020), and critically analyzed by Meinert and Lavin (2022). That is, we generate datasets $\{(x_i, x_i^3 + \epsilon)\}_{i=1}^{N}$ of different sample sizes, where the instances $x_i \in U([-4, 4])$ are uniformly distributed and $\epsilon \sim \mathcal{N}(0, \sigma^2 = 9)$. Even though the setup is homoscedastic, we use negative log-likelihood loss for the first-order learner, while assuming an underlying normal distribution, making σ as well as μ learnable parameters. This choice is made because the second-order risk minimization methods also assume heteroskedastic noise, so in this way the comparison is more fair.

The second-order learner predicts the parameters $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta)$ of a normal-inverse-Gamma distribution, trained with the loss functions obtained by outer and inner expectation minimization, respectively. Identically as in the experiment of

[7] For the proof we found inspiration here: https://math.stackexchange.com/questions/416099/lasso-constraint-form-equivalent-to-penalty-form

Section 4, we use two neural networks of the same architecture, with one fully-connected hidden layer of 32 neurons. The first-order learner needs 2 output neurons to predict $\boldsymbol{\theta} = (\mu, \sigma)$, while the second-order learners uses 4 neurons for the parameters of the NIG distribution. The models are again trained for 5000 epochs using the Adam optimizer and a learning rate of 0.0001. Negative entropy of the NIG distribution is used as a regularization term in the loss function, where we set the weighting factor to $\lambda = 0.01$.

Figure 5 displays the confidence bounds and mean prediction of the estimated reference distribution and the fitted second-order models. The 95% confidence bound for $\widehat{\mu}$ and $\widehat{\sigma^2}$ are empirically computed for the reference model, and analytically for the second-order models, using the fact that $\mu \sim \mathcal{N}(\mu, \frac{\beta}{(\alpha-1)\nu})$ and $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$ for the normal-inverse-Gamma distribution with parameters $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta)$. We see that also in this setup, the mean prediction of the second-order learner remarkably well approximates the one of the reference model, however, the epistemic uncertainty is not well captured. For the outer expectation minimization, the convergence of the distribution of $\widehat{\mu}$ towards a Dirac delta distribution is clearly visible, while regularization does not improve the learning of the underlying epistemic uncertainty. The learned distribution of $\widehat{\sigma^2}$ is more difficult to interpret. Convergence to the Dirac delta distribution has not (yet) taken place. However, one has to keep in mind that all parameters are learned simultaneously by the neural network, and perhaps more is to be gained by focusing on μ. Similarly as for the classification case, Figure 5 should be cautiously interpreted for the inner loss minimization method, because the figure only visualizes one run.

Figure 6 shows the learned parameters as a function of the number of epochs (up to 10000) for 40 runs with random initialization of the neural network. For the unregularized case we see that the network learns to increase ν over time, thereby decreasing the variance of $\widehat{\mu}$. For the inner expectation minimization, we obtain similar results as in (Amini et al., 2020). Note here that for comparability to the first experiment we use negative entropy instead of the proposed evidence based regularizor. For the unregularized case, the distribution of $\widehat{\sigma^2}$, whose expectation is an indicator of the aleatoric uncertainty, is more peaked in the center of region where the model sees data, which is in accordance to the results of (Meinert and Lavin, 2022). Again the issue of non-identifiability of the learned parameters arises, and we obtained a high variability in the results for different runs.

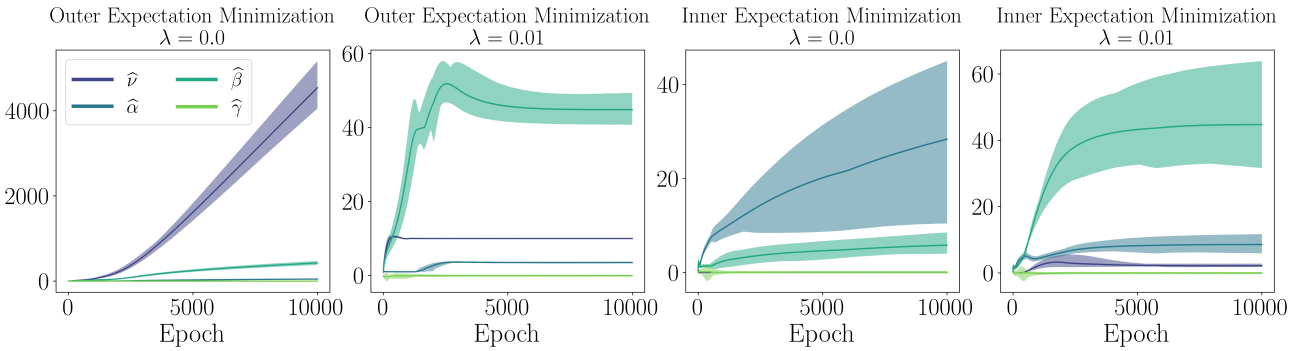


Figure 6. Behavior of the average value of the predicted parameters $\widehat{\boldsymbol{m}} = (\widehat{\gamma}, \widehat{\nu}, \widehat{\alpha}, \widehat{\beta})$, obtained by the second-order learners as a function of the number of training epochs. To obtain the mean and the confidence bounds, 40 models were trained on the same $N = 1000$ training instances, differing only in their random weight initializations. The shaded areas are the 95% empirical confidence bounds of the predictions, averaged over the instance space. In addition, the average over different runs is shown.