Uncertainty Regularized Evidential Regression

Kai Ye[1], Tiejin Chen[2], Hua Wei[2*], Liang Zhan[1*]

[1]University of Pittsburgh, Pittsburgh, PA, 15260, USA
[2]Arizona State University, Tempe, AZ, 85281, USA
hua.wei@asu.edu, liang.zhan@pitt.edu

Abstract

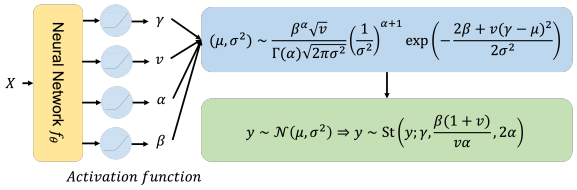
The Evidential Regression Network (ERN) represents a novel approach that integrates deep learning with Dempster-Shafer's theory to predict a target and quantify the associated uncertainty. Guided by the underlying theory, specific activation functions must be employed to enforce non-negative values, which is a constraint that compromises model performance by limiting its ability to learn from all samples. This paper provides a theoretical analysis of this limitation and introduces an improvement to overcome it. Initially, we define the region where the models can't effectively learn from the samples. Following this, we thoroughly analyze the ERN and investigate this constraint. Leveraging the insights from our analysis, we address the limitation by introducing a novel regularization term that empowers the ERN to learn from the whole training set. Our extensive experiments substantiate our theoretical findings and demonstrate the effectiveness of the proposed solution.

Introduction

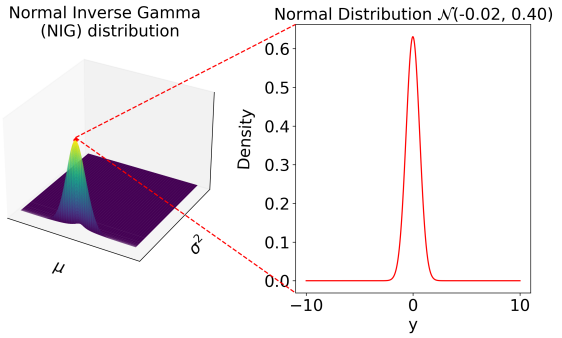
Deep learning methods have been successful in a broad spectrum of real-world tasks, including computer vision (Godard, Mac Aodha, and Brostow 2017; He et al. 2016; Dai et al. 2024), natural language processing (Zhao et al. 2023; Devlin et al. 2018; Vaswani et al. 2017), and medical domain (Ye et al. 2023; Tang et al. 2023). In these scenarios, evaluating model uncertainty becomes a crucial element. Within the realm of deep learning, uncertainty is generally categorized into two primary groups: the intrinsic randomness inherent in data, referred to as the *aleatoric uncertainty*, and the uncertainty associated with model parameters, known as the *epistemic uncertainty* (Gal and Ghahramani 2016; Guo et al. 2017).

Among these, accurately quantifying the uncertainty linked to the model's parameters proves to be a particularly demanding task, due to the inherent complexity involved. To tackle this, strategies such as Ensemble-based methods (Pearce, Leibfried, and Brintrup 2020; Lakshminarayanan, Pritzel, and Blundell 2017) and Bayesian neural networks (BNNs) (Gal and Ghahramani 2016; Wilson and Izmailov 2020; Blundell et al. 2015) have been proposed to measure epistemic uncertainty. Nonetheless, these methods either demand substantial computational resources or encounter challenges in scalability. In response to these limitations, the concept of evidential deep learning techniques (Sensoy, Kaplan, and Kandemir 2018; Amini et al. 2020; Malinin and Gales 2018) has emerged. These methods are formulated to handle uncertainty estimation by producing distribution parameters as their output.

The Evidential Regression Network (ERN) (Amini et al. 2020) introduces a novel deep-learning regression approach that incorporates Dempster-Shafer theory (Shafer 1976) to quantify model uncertainty, resulting in impressive achievements. Within the ERN framework, the training process is conceptualized as an evidence acquisition process, which



(a) Evidential Regression Network (ERN) architecture



(b) Normal Inverse-Gamma (NIG) distribution.

Figure 1: (a) An overview of the Evidential Regression Network (ERN) with (b) illustration of Normal Inverse-Gamma (NIG) distribution. ERN outputs four predictions as distribution parameters, with activation functions like Relu or Softplus to constrain the output to meet the non-negative requirements of distribution parameters.

*Corresponding author.

is inspired by evidential models for classification (Malinin and Gales 2018, 2019; Biloš, Charpentier, and Günnemann 2019; Haußmann, Gerwinn, and Kandemir 2019; Malinin, Mlodozeniec, and Gales 2019). During the training phase, ERN establishes prior distributions over the likelihood function, and each training sample contributes to the formation of a higher-order evidential distribution from which the likelihood function is drawn. During the inference phase, ERN produces the hyperparameters of the evidential distribution, facilitating both prediction and uncertainty estimation without the necessity for sampling. This approach was subsequently extended to multivariate regression tasks by Meinert and Lavin using different prior distributions.

Previous ERN methods (Amini et al. 2020; Malinin et al. 2020; Charpentier et al. 2021; Oh and Shin 2022; Feng et al. 2023; Mei et al. 2023) use specific activation functions like ReLU to ensure non-negative values for parameters of the evidential distribution, such as the variance. Nevertheless, the utilization of such activation functions may inadvertently hinder ERN models' capacity to learn effectively from training samples, thereby impairing overall model performance (Pandey and Yu 2023). Furthermore, in classification tasks, evidential models have underperformed because of the existence of "zero confidence regions" within the evidential space (Pandey and Yu 2023).

However, there is a notable lack of convergence analysis for evidential models in the context of regression tasks. In this paper, we explore the existence of zero confidence regions, which result in high uncertainty areas (HUA) during the training process of ERN models for regression tasks. Building upon the insights derived from our analysis, we propose a novel regularization term that enables the ERN to bypass the HUA and effectively learn from the zero-confidence regions. We also show that the proposed regularization can be generalized to various ERN variants. We conduct experiments on both synthetic and real-world data and show the effectiveness of the proposed method[1]. The main contributions of our work are summarized as follows:

- We revealed the existence of HUA in the learning process of ERN methods with theoretical analysis. The existence of HUA impedes the learning ability of evidential regression models, particularly in regions where ERN exhibits low confidence.
- We propose a novel uncertainty regularization term designed to handle this HUA in evidential regression models and provide theoretical proof of its effectiveness.
- Extensive experiments across multiple datasets and tasks are conducted to validate our theoretical findings and demonstrate the effectiveness of our proposed solution.

Background

Problem Setup

In the context of our study, we consider a regression task derived from a dataset $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, where $X_i \in \mathbb{R}^d$ denotes an independently and identically distributed (i.i.d.) input vector with d dimensions. Corresponding to each input X_i, we have a real-valued target $y_i \in \mathbb{R}$. Our dataset comprises N samples and the task is to predict the targets based on the input data points. We tackle the regression task by modeling the probabilistic distribution of the target variable y, which is formulated as $p\left(y \mid f_{\boldsymbol{\theta}}(X)\right)$, where f refers to a neural network, and $\boldsymbol{\theta}$ denotes its parameters. For simplicity, we omit the subscript i.

Evidential Regression Network

As is illustrated in Figure 1, Evidential Regression Network (ERN) (Amini et al. 2020) introduces a Gaussian distribution $\mathcal{N}\left(\mu, \sigma^2\right)$ with unknown mean μ and variance σ for modeling the regression problem. It is generally assumed that a target value y is drawn i.i.d. from the Gaussian distribution, and that the unknown parameters μ and σ follow a Normal Inverse-Gamma (NIG) distribution:

$$\begin{gathered} y \sim \mathcal{N}\left(\mu, \sigma^2\right) \\ \mu \sim \mathcal{N}\left(\gamma, \sigma^2 v^{-1}\right) \quad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta) \\ \left(\mu, \sigma^2\right) \sim \mathrm{NIG}(\gamma, v, \alpha, \beta) \end{gathered} \tag{1}$$

where $\Gamma(\cdot)$ is the gamma function, parameters $\boldsymbol{m} = (\gamma, v, \alpha, \beta)$, and $\gamma \in \mathbb{R}$, $v > 0$, $\alpha > 1$, $\beta > 0$. The parameters of NIG distribution $\boldsymbol{m}$ is modeled by the output of a neural network $f_{\boldsymbol{\theta}}(\cdot)$, where $\boldsymbol{\theta}$ is the trainable parameters of such neural network. To enforce constraints on (v, α, β), a $\mathrm{SoftPlus}$ activation is applied (additional +1 added to α). Linear activation is used for $\gamma \in \mathbb{R}$. Considering the NIG distribution in Eq 1, the prediction, aleatoric uncertainty, and epistemic uncertainty can be calculated as the following:

$$\underbrace{\mathbb{E}[\mu] = \gamma}_{\text{prediction}} \quad \underbrace{\mathbb{E}\left[\sigma^2\right] = \frac{\beta}{\alpha - 1}}_{\text{aleatoric}} \quad \underbrace{\mathrm{Var}[\mu] = \frac{\beta}{v(\alpha - 1)}}_{\text{epistemic}} \tag{2}$$

Therefore, we can use $\mathbb{E}[\mu] = \gamma$ as the prediction of ERN, $\mathbb{E}\left[\sigma^2\right] = \frac{\beta}{\alpha - 1}$ and $\mathrm{Var}[\mu] = \frac{\beta}{v(\alpha - 1)}$ as the uncertainty estimation of ERN.

The likelihood of an observation y given $\boldsymbol{m}$ is computed by marginalizing over μ and σ^2:

$$p\left(y \mid \boldsymbol{m}\right) = \mathrm{St}\left(y; \gamma, \frac{\beta(1 + v)}{v\alpha}, 2\alpha\right) \tag{3}$$

where $\mathrm{St}\left(y; \mu_{\mathrm{St}}, \sigma_{\mathrm{St}}^2, v_{St}\right)$ is the Student-t distribution with location μ_{St}, scale σ_{St}^2 and degrees of freedom v_{St}.

Training Objective of ERN The parameters $\boldsymbol{\theta}$ of ERN are trained by maximizing the marginal likelihood in Eq. 3. The training objective is to minimize the negative logarithm of $p\left(y \mid \boldsymbol{m}\right)$, therefore the negative log likelihood (NLL) loss function is formulated as:

$$\begin{aligned} \mathcal{L}_{\theta}^{\mathrm{NLL}} = & \frac{1}{2} \log\left(\frac{\pi}{v}\right) - \alpha \log(\Omega) \\ & + \left(\alpha + \frac{1}{2}\right) \log\left(\left(y - \gamma\right)^2 v + \Omega\right) \\ & + \log\left(\frac{\Gamma(\alpha)}{\Gamma\left(\alpha + \frac{1}{2}\right)}\right) \end{aligned} \tag{4}$$

[1]Code is at https://github.com/FlynnYe/UR-ERN

where $\Omega = 2\beta(1+v)$.

To minimize the evidence on errors, the regularization term $\mathcal{L}_\theta^{\mathrm{R}} = |y-\gamma| \cdot (2v+\alpha)$ is proposed to minimize evidence on incorrect predictions. Therefore, the loss function of ERN is:

$$\mathcal{L}_\theta^{\mathrm{ERN}} = \mathcal{L}_\theta^{\mathrm{NLL}} + \lambda \mathcal{L}_\theta^{\mathrm{R}} \tag{5}$$

where λ is a settable hyperparameter. For simplicity, we omit θ in the following sections.

Variants of ERN

ERN is for univariate regression and has been extended to multivariate regression with a different prior distribution normal-inverse-Wishart (NIW) distribution (Meinert and Lavin 2021). Multivariate ERN employs an NIW distribution and, similar to ERN, formulates the loss function as (see Meinert and Lavin for details):

$$\begin{aligned}\mathcal{L}^{\mathrm{MERN}} \equiv \mathcal{L}^{\mathrm{NLL}} &= \log\Gamma\left(\frac{\nu-n+1}{2}\right) - \log\Gamma\left(\frac{\nu+1}{2}\right) \\ &+ \frac{n}{2}\log\left(r+\nu\right) - \nu\sum_j \ell_j \\ &+ \frac{\nu+1}{2}\log\left|\boldsymbol{L}\boldsymbol{L}^\top + \frac{1}{r+\nu}\left(\vec{y}-\vec{\mu}_0\right)\left(\vec{y}-\vec{\mu}_0\right)^\top\right| \\ &+ \text{ const.}\end{aligned} \tag{6}$$

And estimation of the prediction as well as uncertainties as:

$$\begin{gathered}\underbrace{\mathbb{E}[\mu] = \vec{\mu}_0}_{\text{prediction}} \quad \underbrace{\mathbb{E}[\boldsymbol{\Sigma}] \propto \frac{\nu}{\nu-n-1}\boldsymbol{L}\boldsymbol{L}^\top}_{\text{aleatoric}} \\ \underbrace{\mathrm{var}[\vec{\mu}] \propto \mathbb{E}[\boldsymbol{\Sigma}]/\nu}_{\text{epistemic}}\end{gathered} \tag{7}$$

To learn the parameters $\boldsymbol{m} = \left(\vec{\mu}_0, \vec{\ell}, \nu\right)$, a NN has to have $n(n+3)/2+1$ outputs $(p_1 \cdots p_m)$. Also, activation functions have to be applied to the outputs of NN to ensure the following:

$$\nu = n(n+5)/2 + \tanh p_\nu \cdot n(n+3)/2 + 1 > n+1 \tag{8}$$

where $p_\nu \in (p_1 \cdots p_m)$. And

$$(\boldsymbol{L})_{jk} = \begin{cases}\exp\{\ell_j\} & \text{if } j = k \\ \ell_{jk} & \text{if } j > k \\ 0 & \text{else.}\end{cases} \tag{9}$$

where $\ell_j, \ell_{jk} \in (p_1 \cdots p_m)$.

Methodology

In this section, we first give a definition of the High Uncertainty Area (HUA). Then, we theoretically analyze the existing limitation of ERN in HUA. Based on our analysis, we propose a novel solution to the problem. Finally, we extend our analysis and propose solutions to variants of ERN with other prior distributions.

High Uncertainty Area (HUA) of ERN

In this section, we show that in the high uncertainty area of ERN, the gradient of ERN will shrink to zero, therefore the outputs of ERN cannot be correctly updated. In this paper, we only study the gradient with respect to α as the gradient with respect to v and β follows a similar fashion.

Definition 1 (**High Uncertainty Area**). *High Uncertainty Area is where α is close to 1, leading to very high uncertainty prediction.*

An effective model ought to possess the capacity to learn from the entire training samples. Unfortunately, this does not hold true in the context of ERN.

Theorem 1. *ERN cannot learn from samples in high uncertainty area.*

Proof. Consider input X and the corresponding label y. We use $\boldsymbol{o} = (o_\gamma, o_v, o_\alpha, o_\beta)$ to denote the output of $f_{\boldsymbol{\theta}}(X)$, therefore:

$$\begin{aligned}\alpha &= \mathrm{SoftPlus}(o_\alpha) + 1 \\ &= \log(\exp(o_\alpha)+1) + 1\end{aligned} \tag{10}$$

Where $\mathrm{SoftPlus}(\cdot)$ denotes $\mathrm{SoftPlus}$ activation (our theorem still holds true when faced with other popular activation functions, such as ReLU and $\exp$, see Appendix A[2] for additional proofs). So the gradient of NLL loss with respect to o_α is given by:

$$\begin{aligned}\frac{\partial\mathcal{L}^{\mathrm{NLL}}}{\partial o_\alpha} &= \frac{\partial\mathcal{L}^{\mathrm{NLL}}}{\partial\alpha}\frac{\partial\alpha}{\partial o_\alpha} \\ &= [\log(1+\frac{\nu(\gamma-y)^2}{2\beta(\nu+1)}) + \psi(\alpha) \\ &\quad - \psi(\alpha+0.5)] \cdot \mathrm{Sigmoid}\left(o_\alpha\right)\end{aligned} \tag{11}$$

where $\psi(\cdot)$ denotes the digamma function.

For a sample in high uncertainty area, we have:

$$\alpha \to 1 \Rightarrow o_\alpha \to -\infty \Rightarrow \mathrm{Sigmoid}\left(o_\alpha\right) \to 0 \tag{12}$$

So, for such training samples:

$$\frac{\partial\mathcal{L}^{\mathrm{NLL}}}{\partial o_\alpha} = 0 \tag{13}$$

And the gradient of $\mathcal{L}^{\mathrm{R}} = |y-\gamma| \cdot (2v+\alpha)$ with respect to o_α is given by:

$$\begin{aligned}\frac{\partial\mathcal{L}^{\mathrm{R}}}{\partial o_\alpha} &= \frac{\partial\mathcal{L}^{\mathrm{R}}}{\partial\alpha}\frac{\partial\alpha}{\partial o_\alpha} \\ &= |y-\gamma| \cdot \mathrm{Sigmoid}\left(o_\alpha\right)\end{aligned} \tag{14}$$

Similarly, we have:

$$\frac{\partial\mathcal{L}^{\mathrm{R}}}{\partial o_\alpha} = 0 \tag{15}$$

And $\mathcal{L}^{\mathrm{ERN}} = \mathcal{L}^{\mathrm{NLL}} + \lambda\mathcal{L}^{\mathrm{R}}$, therefore we have:

$$\begin{aligned}\frac{\partial\mathcal{L}^{\mathrm{ERN}}}{\partial o_\alpha} &= \frac{\partial\mathcal{L}^{\mathrm{NLL}}}{\partial o_\alpha} + \lambda\frac{\partial\mathcal{L}^{\mathrm{R}}}{\partial o_\alpha} \\ &= 0\end{aligned} \tag{16}$$

□

[2]Please find Appendix in arXiv version.

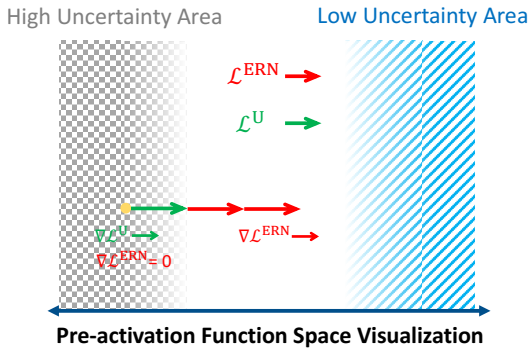


Figure 2: $\mathcal{L}^{\mathrm{ERN}}$ in Equation (5) cannot help the model get out of high uncertainty area while our proposed $\mathcal{L}^{\mathrm{U}}$ can still learn from samples in the grey area.

Since the gradient of the loss function with respect to o_α is zero, there won't be any update on α from such samples. The model fails to learn from samples in high uncertainty area.

Uncertainty Regularization to Bypass HUA

Considering the learning deficiency of ERN, in this paper, we propose an uncertainty regularization to solve the zero gradient problem within HUA:

$$\mathcal{L}^{\mathrm{U}} = -|y-\gamma| \cdot \log(\exp(\alpha-1)-1) \tag{17}$$

In this section, we show that $\mathcal{L}^{\mathrm{U}}$ can address the learning deficiency of ERN.

Theorem 2. *Our proposed uncertainty regularization $\mathcal{L}^{\mathrm{U}}$ can learn from samples within HUA.*

Proof. The gradient of the proposed regularization term $\mathcal{L}^{\mathrm{U}}$ with respect to o_α is given by:

$$\begin{aligned}\frac{\partial\mathcal{L}^{\mathrm{U}}}{\partial o_\alpha} &= \frac{\partial\mathcal{L}^{\mathrm{U}}}{\partial\alpha}\frac{\partial\alpha}{\partial o_\alpha} \\ &= -|y-\gamma| \cdot \frac{\exp(\alpha-1)}{\exp(\alpha-1)-1} \cdot \mathrm{Sigmoid}\,(o_\alpha) \\ &= -|y-\gamma| \cdot [1+\exp(-o_\alpha)] \cdot \frac{1}{1+\exp(-o_\alpha)} \\ &= -|y-\gamma|\end{aligned} \tag{18}$$

□

Uncertainty regularization term $\mathcal{L}^{\mathrm{U}}$ ensures the maintenance of the gradient within the high uncertainty area. Importantly, the value of this gradient scales in accordance with the distance between the predicted value and the ground truth.

Training of Regularized ERN The final training objective for the proposed Uncertainty Regularized ERN (UR-ERN) is formulated as:

$$\mathcal{L} = \mathcal{L}^{\mathrm{ERN}} + \lambda_1\mathcal{L}^{\mathrm{U}} \tag{19}$$

where λ_1 is a settable hyperparameter that balances the regularization and the original ERN loss. $\mathcal{L}^{\mathrm{NLL}}$ is for fitting purpose, $\mathcal{L}^{\mathrm{R}}$ regularizes evidence (Amini et al. 2020). And our proposed $\mathcal{L}^{\mathrm{U}}$ addresses zero gradient problem in the HUA.

Uncertainty Space Visualization Figure 2 visualizes the uncertainty space with x-axis representing o_α. Under ideal conditions, both fitting loss and uncertainty should be low, resulting in samples being mapped to the blue zone. Nevertheless, there exist certain samples predicted with high uncertainty, which may land within the grey region. Within this grey region, $\mathcal{L}^{\mathrm{ERN}}$ fails to update the parameters effectively. Under such circumstances, our proposed uncertainty regularization term $\mathcal{L}^{\mathrm{U}}$ retains the capacity to update the model. This enables the samples to be extracted from the grey area, thus allowing the training to continue.

Uncertainty Regularization for ERN Variants

Based on our theoretical analysis in previous sections, it is quite clear that the zero gradient problem in the HUA of ERN is attributable to certain activation functions that ensure non-negative values. Consequently, this limitation is not confined to ERN but can also extend to other evidential models that utilize similar activation functions. Multivariate ERN (Meinert and Lavin 2021), which we introduced in Section , serves as an illustrative example; it suffers from similar problems to ERN, even when employing different prior distributions. Similar to the previous analysis, we study the parameter ν as an example.

Theorem 3. *Multivariate ERN (Meinert and Lavin 2021) also cannot learn from samples in high uncertainty area.*

Proof. Given the output of a neural network $(p_1 \cdots p_m)$, we have $p_\nu \in (p_1 \cdots p_m)$. And ν is formulated as the following:

$$\nu = n(n+5)/2 + \tanh p_\nu \cdot n(n+3)/2 + 1 \tag{20}$$

Therefore, the gradient of loss function $\mathcal{L}^{\mathrm{MERN}}$ ($\mathcal{L}^{\mathrm{NLL}}$) with respect to p_ν is given by:

$$\begin{aligned}\frac{\partial\mathcal{L}^{\mathrm{MERN}}}{\partial p_\nu} &= \frac{\partial\mathcal{L}^{\mathrm{NLL}}}{\partial\nu}\frac{\partial\nu}{\partial p_\nu} \\ &= \frac{\partial\mathcal{L}^{\mathrm{NLL}}}{\partial\nu} \cdot \frac{n(n+3)}{2} \cdot (1-\tanh^2 p_\nu)\end{aligned} \tag{21}$$

within HUA, we have:

$$p_\nu \to -\infty \Rightarrow (1-\tanh^2 p_\nu) \to 0 \tag{22}$$

Therefore, we have:

$$\frac{\partial\mathcal{L}^{\mathrm{MERN}}}{\partial p_\nu} = 0 \tag{23}$$

The gradient with respect to p_ν is zero, there will be no update to p_ν. Multivariate ERN cannot learn effectively within HUA. □

Similarly, we propose uncertainty regularization term $\mathcal{L}^{\mathrm{U}}$ to help Multivariate ERN learn from samples within HUA. Since Multivariate ERN uses a different activation function,

the proposed $\mathcal{L}^{\mathrm{U}}$ for Multivariate ERN has a different formulation:

$$\mathcal{L}^{\mathrm{U}} = -\frac{1}{2} \cdot |y-\gamma| \cdot \log\left(\frac{n^2+3n}{n^2+4n+1-\nu} - 1\right) \tag{24}$$

We can prove the effectiveness of the proposed $\mathcal{L}^{\mathrm{U}}$ for Multivariate ERN.

Theorem 4. *Our proposed uncertainty regularization $\mathcal{L}^{\mathrm{U}}$ enables Multivariate ERN learn from samples within HUA.*

Proof. The gradient of the proposed $\mathcal{L}^{\mathrm{U}}$ with respect to p_ν is given by:

$$\begin{aligned}
\frac{\partial \mathcal{L}^{\mathrm{U}}}{\partial p_\nu} &= \frac{\partial \mathcal{L}^{\mathrm{U}}}{\partial \nu}\frac{\partial \nu}{\partial p_\nu} \\
&= -|y-\gamma| \cdot \frac{1}{2(n^2+3n)} \\
&\cdot \frac{-(n^2+3n)^2}{(\nu-n-1)(\nu-n^2-4n-1)} \\
&\cdot \frac{n^2+3n}{2} \cdot (1-\tanh^2 p_\nu) \\
&= -|y-\gamma| \cdot \frac{1}{2(n^2+3n)} \\
&\cdot \frac{-(n^2+3n)^2}{(\nu-n-1)(\nu-n^2-4n-1)} \\
&\cdot \frac{n^2+3n}{2} \cdot \frac{-4(\nu-n-1)(\nu-n^2-4n-1)}{(n^2+3n)^2} \\
&= -|y-\gamma|
\end{aligned} \tag{25}$$

□

The proposed regularization term $\mathcal{L}^{\mathrm{U}}$ guarantees a non-zero gradient for Multivariate ERN in the HUA. Therefore, the loss function for uncertainty regularized Multivariate ERN is formulated as:

$$\mathcal{L} = \mathcal{L}^{\mathrm{NLL}} + \lambda_1 \mathcal{L}^{\mathrm{U}} \tag{26}$$

where λ_1 is settable hyperparameter.

While the two terms have different formulations, they share a common intuition. We identify the zero gradient problem arising from the activation function and introduce a term to circumvent zero gradients, simultaneously increasing α. This adjustment guides the training process away from this problematic area. Our mathematical analysis confirms these terms effectively achieve our objective.

The above theoretical analysis reveals that the learning deficiency is not exclusive to ERN (Amini et al. 2020); it also manifests in other evidential models (Meinert and Lavin 2021) that employ different prior distributions.

Experiments

In this section, we first conduct experiments under both synthetic and real-world datasets. For each dataset, we investigate whether the methods fail to learn from samples within and outside HUA. Moreover, we perform additional experiments to demonstrate that even the Multivariate ERN, which employs distinct prior distributions, struggles to learn effectively within HUA. To compare performance, we use baselines including ERN (Amini et al. 2020) ($\mathcal{L}^{\mathrm{NLL}} + \lambda\mathcal{L}^{\mathrm{R}}$), and NLL-ERN ($\mathcal{L}^{\mathrm{NLL}}$). For experiments within HUA, we initialize the model within HUA by setting bias in the activation layer. Please refer to Appendix B for details about experimental setups and experiments about the sensitivity of hyperparameters.

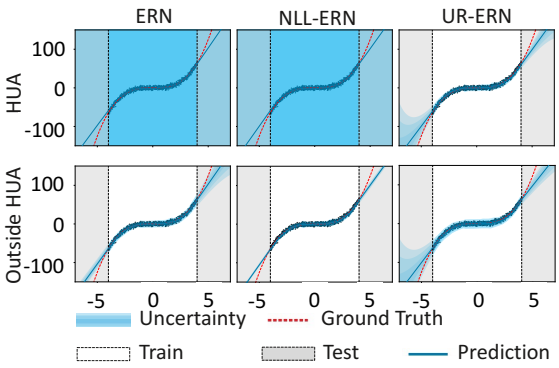


Figure 3: Uncertainty estimation on Cubic Regression. The blue shade represents prediction uncertainty. An effective evidential model would cause the blue shade to cover the distance between the predicted value and the ground truth precisely. Up: Comparison of model performance within HUA. Down: Comparison of model performance outside HUA. UR-ERN can cover the ground truth precisely both within HUA and outside HUA.

Performance on Cubic Regression Dataset

To highlight the limitations of ERN, we compare its performance with our proposed UR-ERN on cubic regression dataset (Amini et al. 2020) within HUA. Following (Amini et al. 2020), we train models on $y = x^3 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 3)$. We conduct training over the interval $x \in [-4, 4]$, and perform testing over $x \in [-6, -4) \cup (4, 6]$.

Evaluation Metrics Our proposed regularization is mainly designed to help the model effectively update the parameter α within HUA. This is essential because, as our theoretical analysis has shown, if the model cannot properly update α, the uncertainty prediction will become unreasonably high. Therefore, we have chosen uncertainty prediction as our evaluation metric. We visualize the experimental results of uncertainty estimation along with ground truth in Figure 3 and the uncertainty is represented by the blue shade. An accurate prediction in uncertainty would lead the blue shade to cover the distance between the predicted value and the ground truth precisely.

Cubic Regression within HUA As illustrated in Figure 3, where the blue shade represents the uncertainty predicted by the models, ERN encounters difficulties in updating parameters in the HUA, resulting in high uncertainty predictions across the dataset. In contrast, the proposed UR-ERN maintains its training efficiency, effectively mitigating this issue.

These observations validate our theoretical analysis, demonstrating the effectiveness of our proposed method.

Cubic Regression outside HUA We extend our investigation to assess the performance of these methods under standard conditions (outside the HUA). Figure 3 illustrates that the inclusion of the term $\mathcal{L}^{\mathrm{R}}$ in $\mathcal{L}^{\mathrm{ERN}}$ contributes to more accurate uncertainty predictions, a result that aligns with the findings of Amini et al.. Moreover, the proposed UR-ERN not only performs robustly in the HUA but also exhibits superior performance compared to the ERN outside the HUA. These observations further demonstrate the effectiveness of our method.

Performance on Monocular Depth Estimation

We further evaluate the performance of our proposed UR-ERN and ERN on more challenging real-world tasks. Monocular depth estimation is a task in computer vision aiming to predict the depth directly from an RGB image. We choose the NYU Depth v2 dataset (Silberman et al. 2012) for experiments. For each pixel, there is a corresponding depth target. Following previous practice (Amini et al. 2020), we train U-Net (Ronneberger, Fischer, and Brox 2015) style neural network as the backbone to learn evidential parameters. Similar to the previous section, we compare the performance of our UR-ERN against ERN within HUA and outside HUA. Limited by space, additional experimental results are detailed in Appendix B.

Evaluation Metrics We first explore whether the models can correctly update parameters within HUA. Similarly, we choose the value of uncertainty as the evaluation metric. Similar to cubic regression, the blue shade in Figure 4(a) depicts predicted uncertainty. The models that cannot learn from samples within HUA will exhibit excessively large blue shade areas, resulting from their high uncertainty prediction across the test set.

Following existing works Amini et al.; Kuleshov, Fenner, and Ermon, we also use cutoff curves and calibration curves to compare the performance of uncertainty estimation. Inspired by previous work (Amini et al. 2020), we further test how the models perform when faced with OOD data. An effective evidential model should predict high uncertainty for OOD data and can distinguish the OOD data. The OOD experimental setup is the same as Amini et al. for comparison.

Monocular Depth Estimation within HUA As illustrated in Figure 4, ERN with $\mathcal{L}^{\mathrm{R}}$ or not, struggles to update parameters effectively within the HUA, leading to suboptimal uncertainty estimation. This constraint forms a significant impediment to effective learning from particular samples. In contrast, the proposed UR-ERN successfully navigates this challenge, demonstrating the capacity to learn from these specific samples and to efficiently estimate uncertainty, mirroring the behavior observed in normal regions.

Figure 4 shows model performances as pixels possessing uncertainty beyond specific thresholds are excluded. The proposed UR-ERN demonstrates robust behavior, characterized by a consistent reduction in error corresponding to increasing levels of confidence. In addition to the performance comparison, Figure 4 provides an assessment of the calibration of our uncertainty estimates. The calibration curves, computed following the methodology described in previous work (Kuleshov, Fenner, and Ermon 2018), should ideally follow $y = x$ for accurate representation. The respective calibration errors for each model are also shown.

Monocular Depth Estimation outside HUA We also look at how the models perform outside the HUA. Figure 5 visualizes the comparison of how the models can estimate uncertainty in depth estimation outside HUA. The proposed UR-ERN has a lower Root Mean Square Error (RMSE) for most confidence levels than the competing models. Also, the calibration curve of our method is closer to the ideal curve than any competing model.

For OOD experiments, the proposed UR-ERN can distinguish OOD data better than the competing models. The above experiments reveal that the proposed regularization can not only be effective at guiding the model to get out of HUA, but it also performs well outside HUA.

Extension to Different ERN Variants

Our theoretical findings reveal that the performance issues within HUA extend beyond ERN. Other evidential models, even those utilizing different prior distributions, similarly exhibit poor performance within this challenging region.

Following the theoretical analysis in the previous section, we compare the performance of models in the context of Multivariate Deep Evidential Regression (Meinert and Lavin 2021). Following the experimental setup in (Meinert and Lavin 2021), we conduct the multivariate experiment and predict $(x, y) \in \mathbb{R}^2$ given $t \in \mathbb{R}$, where x and y being the features of the data sample given input t with the following definition:

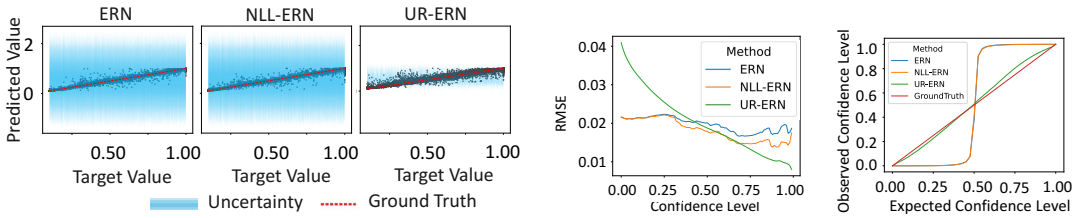
$$x = (1 + \epsilon)\cos t, \quad y = (1 + \epsilon)\sin t, \tag{27}$$

and the distribution of t is formulated as following:

$$t \sim \begin{cases} 1 - \frac{\zeta}{\pi} & \text{if } \zeta \in [0, \pi] \\ \frac{\zeta}{\pi} - 1 & \text{if } \zeta \in (\pi, 2\pi] \\ 0 & \text{else} \end{cases} \tag{28}$$

where $\zeta \in [0, 2\pi]$ is uniformly distributed and $\epsilon \sim \mathcal{N}(0, 0.1)$ is drawn from a normal distribution. Under this setting, the uncertainty is calculated as $\frac{\boldsymbol{L}\boldsymbol{L}^\top}{\nu - 3}$. When $\nu \to 3$, the corresponding uncertainty will be infinite across the dataset. Similar to previous experiments, we initialize the model within HUA by setting bias in the activation layer (See details in Appendix B).

Figure 6(a) shows that the Multivariate ERN struggles to update the parameter ν, resulting in unreasonably high uncertainty estimations (see Appendix B for additional experimental results). Consistent with previous sections, the proposed UR-ERN in Figure 6(b) does not encounter this issue within HUA and provides reasonable uncertainty predictions. This validates our theoretical findings, demonstrating that evidential models, including but not limited to ERN, face challenges in the HUA when utilizing specific activation functions to ensure non-negative values. Our solution effectively overcomes these issues.



(a) Performance within HUA of Depth Estimation (b) RMSE with confidence (c) Uncertainty calibration

Figure 4: Uncertainty prediction of depth estimation within HUA. (a) The blue shade represents prediction uncertainty. A good estimation of uncertainty should cover the gap between prediction and ground truth exactly. (b) Root Mean Square Error (RMSE) at various confidence levels. The evidential model with a larger confidence level should have a lower RMSE. (c) Uncertainty calibration calculated following previous work (Kuleshov, Fenner, and Ermon 2018), the ideal curve is $y = x$. The calibration errors are 0.2261, 0.2250, and 0.0243 for ERN, NLL-ERN and UR-ERN, respectively.

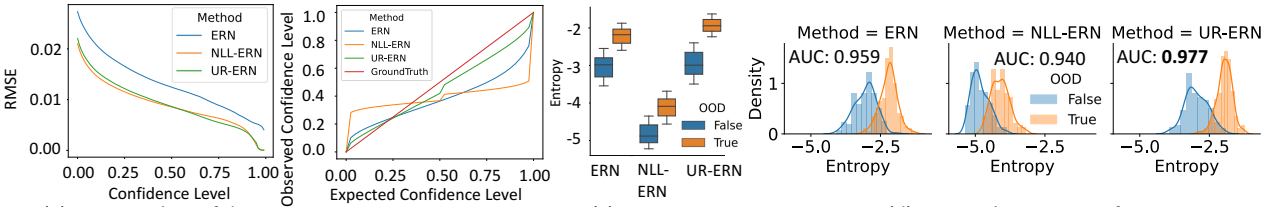
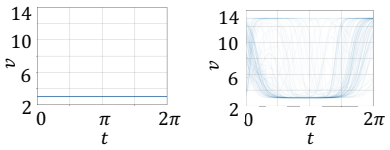


Figure 5: Uncertainty prediction of depth estimation outside HUA. (a) RMSE at various confidence levels. (b) Uncertainty calibration (ideal: $y = x$). The calibration errors are 0.1366, 0.1978, and 0.0289 for ERN, NLL-ERN and UR-ERN, respectively. (c) and (d) show OOD experimental results. (c) Entropy comparisons for different methods. (d) Density histograms of entropy. Entropy is calculated from σ, directly related to uncertainty. A good evidential model should be able to distinguish OOD data.



(a) Multivariate ERN (b) UR-ERN

Figure 6: Prediction of parameter ν in Multivariate ERN and our proposed UR-ERN. Uncertainty ($\frac{\boldsymbol{L}\boldsymbol{L}^\top}{\nu-3}$) will be infinite if ν is close to 3, indicating the evidential model fails to properly estimate the uncertainty of predictions.

Related Works

Uncertainty Estimation in Deep Learning Developing a trustworthy Deep Learning (DL) model requires an accurate estimation of prediction uncertainty. Ensemble methods (Pearce, Leibfried, and Brintrup 2020; Lakshminarayanan, Pritzel, and Blundell 2017) use multiple networks for uncertainty quantification and thus are computationally expensive due to the need for more parameters. Bayesian neural networks (BNNs) (Gal and Ghahramani 2016; Wilson and Izmailov 2020; Blundell et al. 2015), treating neural network weights as random variables, capture weight distribution rather than point estimates. The introduction of Dropout to BNNs during inference (Gal and Ghahramani 2016) approximates Bayesian inference in deep Gaussian processes but also increases computational costs due to sampling.

Evidential Deep Learning Evidential Deep Learning (EDL) (Sensoy, Kaplan, and Kandemir 2018; Amini et al. 2020; Malinin and Gales 2018) is a relatively recent method for uncertainty estimation in deep learning, using a conjugate higher-order evidential prior to estimate uncertainty. These models train the neural network to predict distribution parameters that capture both the target variable and its associated uncertainty. Dirichlet prior is introduced for evidential classification (Sensoy, Kaplan, and Kandemir 2018). And NIG prior is introduced for evidential regression (Amini et al. 2020). Meinert and Lavin (2021) further utilize NIW prior for multivariate regression. Pandey and Yu (2023) first observed convergence issues in evidential models for classification, noting their incapacity to learn from certain samples. To tackle this, they introduced evidence regularization. However, the convergence analysis of ERN for regression tasks remains unexplored.

Conclusion

In this paper, we identify the zero gradient problem for evidential regression models. To combat this issue, we introduce a novel regularization term, and our experiments validate the effectiveness of our solution. Future work could be extending our investigation into more evidential models.

Acknowledgments

This study is partially supported by NSF award (IIS 2045848, IIS 1837956, IIS 2319450, and IIS 2153311).

References

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep Evidential Regression. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 14927–14937. Curran Associates, Inc.

Biloš, M.; Charpentier, B.; and Günnemann, S. 2019. Uncertainty on asynchronous time event prediction. *Advances in Neural Information Processing Systems*, 32.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, 1613–1622. PMLR.

Charpentier, B.; Borchert, O.; Zügner, D.; Geisler, S.; and Günnemann, S. 2021. Natural Posterior Network: Deep Bayesian Predictive Uncertainty for Exponential Family Distributions. In *International Conference on Learning Representations*.

Dai, S.; Ye, K.; Zhao, K.; Cui, G.; Tang, H.; and Zhan, L. 2024. Constrained Multiview Representation for Self-supervised Contrastive Learning. *arXiv preprint arXiv:2402.03456*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feng, Z.; Qi, K.; Shi, B.; Mei, H.; Zheng, Q.; and Wei, H. 2023. Deep evidential learning in diffusion convolutional recurrent neural network. *Electronic Research Archive*, 31(4): 2252–2264.

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

Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 270–279.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

Haußmann, M.; Gerwinn, S.; and Kandemir, M. 2019. Bayesian evidential deep learning with PAC regularization. *arXiv preprint arXiv:1906.00816*.

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

Hernández-Lobato, J. M.; and Adams, R. 2015. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, 1861–1869. PMLR.

Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; and Yang, R. 2018. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 954–960.

Kuleshov, V.; Fenner, N.; and Ermon, S. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, 2796–2804. PMLR.

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

Malinin, A.; Chervontsev, S.; Provilkov, I.; and Gales, M. 2020. Regression prior networks. *arXiv preprint arXiv:2006.11590*.

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

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

Malinin, A.; Mlodozeniec, B.; and Gales, M. 2019. Ensemble Distribution Distillation. In *International Conference on Learning Representations*.

Mei, H.; Li, J.; Liang, Z.; Zheng, G.; Shi, B.; and Wei, H. 2023. Uncertainty-aware Traffic Prediction under Missing Data. *The Proceedings of 2023 IEEE International Conference on Data Mining (ICDM 2023)*.

Meinert, N.; and Lavin, A. 2021. Multivariate deep evidential regression. *arXiv preprint arXiv:2104.06135*.

Oh, D.; and Shin, B. 2022. Improving evidential deep learning via multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7895–7903.

Pandey, D. S.; and Yu, Q. 2023. Learn to Accumulate Evidence from All Training Samples: Theory and Practice. In *International Conference on Machine Learning*, 26963–26989. PMLR.

Pearce, T.; Leibfried, F.; and Brintrup, A. 2020. Uncertainty in neural networks: Approximately bayesian ensembling. In *International conference on artificial intelligence and statistics*, 234–244. PMLR.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

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

Shafer, G. 1976. *A mathematical theory of evidence*, volume 42. Princeton university press.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, 746–760. Springer.

Tang, H.; Ma, G.; Zhang, Y.; Ye, K.; Guo, L.; Liu, G.; Huang, Q.; Wang, Y.; Ajilore, O.; Leow, A. D.; et al. 2023. A Comprehensive Survey of Complex Brain Network Representation. *Meta-Radiology*, 100046.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wilson, A. G.; and Izmailov, P. 2020. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33: 4697–4708.

Ye, K.; Tang, H.; Dai, S.; Guo, L.; Liu, J. Y.; Wang, Y.; Leow, A.; Thompson, P. M.; Huang, H.; and Zhan, L. 2023. Bidirectional Mapping with Contrastive Learning on Multimodal Neuroimaging Data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 138–148. Springer.

Zhao, K.; Yang, B.; Lin, C.; Rong, W.; Villavicencio, A.; and Cui, X. 2023. Evaluating Open-Domain Dialogues in Latent Space with Next Sentence Prediction and Mutual Information. *arXiv preprint arXiv:2305.16967*.