The Evidence Contraction Issue in Deep Evidential Regression: Discussion and Solution

Anonymous submission

Abstract

Deep Evidential Regression (DER) places a prior on the original Gaussian likelihood function and treats learning as an evidence acquisition process to quantify uncertainty by inferring the parameters of the prior. For the validity of the evidence theory, DER requires specialized activation functions to ensure that the prior parameters remain non-negative. However, such constraints can lead to evidence contraction, causing sub-optimal performance. In this paper, we analyse DER theoretically, revealing the intrinsic limitations that lead to sub-optimal performance: the non-negativity constraints on the Normal Inverse-Gamma (NIG) prior parameter lead to the evidence contraction under the specialised activation function, which hampers the optimisation of DER performance. Based on this foundation, we then design a Non-saturating Uncertainty Regularization term, which effectively ensures that the accuracy is further optimised in the right direction, mitigating the above issue. Experiments on the real-world datasets demonstrate that our proposed approach enhances the performance of DER while maximising the ability to preserve uncertainty quantification.

Introduction

Deep Learning has been highly successful in many real-world applications and is encouraged to be applied in various research areas such as Data Mining (Xu et al. 2023), Natural Language Processing (Zhang et al. 2023) and Computer Vision (Kirillov et al. 2023). Despite the attractiveness of Deep Learning models, their deployment in high-risk domains such as Weather Prediction (Bi et al. 2023), Vehicle Control (Choi et al. 2019) and Medical Diagnostics (Seebock et al. 2020) is still limited, which is mainly attributed to the fact that Deep Learning models are subject to uncertainty.

Researchers regard Uncertainty Quantification as one of the foundations for building safe and reliable Deep Learning systems(Guo et al. 2017a; Lakshminarayanan, Pritzel, and Blundell 2017; Gal and Ghahramani 2015), and the causes of uncertainty can be categorised into two groups: uncertainty already contained in the data (data uncertainty) or lack of neural network knowledge (model uncertainty) (Abdar et al. 2020; Gawlikowski et al. 2021). Due to the high expressiveness of deep learning, it might lead to overconfident predictions. Failing to provide reliable uncertainty estimates for decision-making in Deep Learning models would result in catastrophic consequences (Amini et al. 2020; Sensoy, Kaplan, and Kandemir 2018).

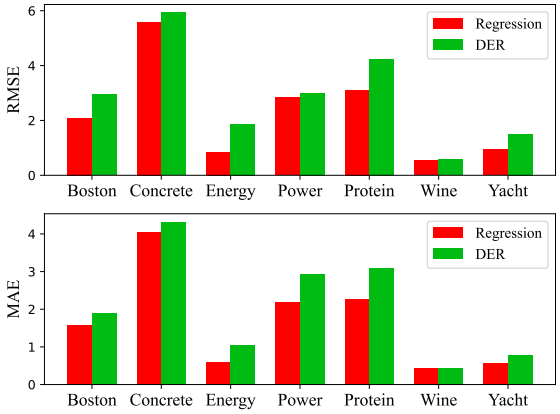

Figure 1: Performance Comparison on seven real world dataset from UCI Regression benchmark. Both RMSE and MAE, smaller is better. It is evident that, in contrast to standard regression methods (red legend), DER (green legend) exhibits noticeably inferior performance across several datasets.

Over the past few years, significant progress has been made in uncertainty quantification, primarily through Bayesian methods (Kendall and Gal 2017; Gal and Ghahramani 2015) and Deep Ensemble methods (Lakshminarayanan, Pritzel, and Blundell 2017; Zaidi et al. 2021). However, these mehtods are limited by the difficulties in approximating posterior computation or the high cost of sampling, and they cannot achieve fine-grained uncertainty quantification (Malinin and Gales 2018; Amini et al. 2020). Indeed, to address these challenges, (Amini et al. 2020) has proposed the concept of Deep Evidential Regression (DER). The DER employs higher-order conjugate priors, Normal Inverse-Gamma (NIG), placed on the likelihood function, formulating learning as an evidence acquisition process. Due to only minor modifications to deterministic neural networks without the sampling and the ability to quantify both epistemic and aleatoric uncertainties in a single forward pass, DER have gained widespread adoption, yielding remarkable

outcomes.(Liu et al. 2021; Soleimany et al. 2021; Cai et al. 2021; Chen, Bromuri, and van Eekelen 2021; Singh et al. 2022; Petek et al. 2022; Li and Liu 2022; Amini et al. 2020; Ma et al. 2021; Malinin et al. 2020; Charpentier et al. 2022; Oh and Shin 2022; Pandey and Yu 2023a). Despite the attractive ability for uncertainty quantification, the DER's error is noticeably bigger than standard regression methods, even in generic scenarios, as shown in Fig. 1. This misalignment with the pursuit of lower error in regression learning poses a challenge for DER's widespread deployment.

In this paper, we provide a theoretical analysis of deep evidence regression to explore the intrinsic hindrances in its performance gap compared to standard regression. Specifically, to ensure the validity of the evidence theory, the DER require specialized activation functions to guarantee the non-negativity of the NIG prior parameters. However, such constraints could potentially result in evidence contraction, i.e., evidence from the data is insufficient to support the prediction. On this foundation, we further elucidate how DER's performance is hindered when the evidence contraction, analysing the role of different NIG parameters. Furthermore, we design a non-saturated regularisation term, which effectively ensures that the gradient is further optimised in the right direction and reduces the error of DER. Experiments on real-world datasets show that our proposed method is effective in the prediction error of DER without compromising the ability to quantify uncertainty.

The main contributions of this paper are as follows:

- We show theoretically that ensuring non-negativity of the NIG prior parameter leads to evidence contraction.
- Next, we show theoretically that evidence contraction leads to sub-optimal performance and is largely attributable to virtual observation ν.
- We design a non-saturated regularisation term, which effectively ensures that the gradient is further optimised in the right direction and improves the performance of DER.
- Experiments on real-world datasets demonstrate the effectiveness of our approach.

Related Wok

Uncertainty Quantification in Deep Learning. Effective uncertainty quantification capabilities are essential for developing reliable Deep Learning systems. Bayesian Neural Network (BNN) place priors on model weights, explicitly modeling network parameters as random variables. As a result, BNN naturally quantify uncertainty by learning the posterior over parameters (Abdar et al. 2021). Indeed, for Bayesian networks with a large number of parameters, it's posterior probability of Bayesian networks is intractable. As a result, several Bayesian approximation methods have been proposed to address above issue. For instance, Markov Chain Monte Carlo (MCMC) (Karras et al. 2022) and Stochastic Gradient MCMC (SG-MCMC) (Welling and Teh 2011) are such exemplar. But those methods heavily rely on sampling from the posterior distribution, which leads to increased computational costs. Another well-known Bayesian approximation method is Monte Carlo Dropout (MC Dropout) (Gal and Ghahramani 2016). It treats dropout layers as Bernoulli-distributed random variables, and training the network with Dropout layer can be interpreted as an approximation to variational inference. However, these methods require significant modifications to the training process and come with high computational costs. Additionally, they are unable to distinguish between epistemic and aleatoric uncertainty. Unlike the Bayesian perspective, frequentist researchers have a unique insight to uncertainty quantification and have proposed deep ensemble techniques (Pearce, Leibfried, and Brintrup 2020; Lakshminarayanan, Pritzel, and Blundell 2017). This method builds an ensemble of neural networks and uses the consistency/inconsistency among ensemble members to quantify uncertainty. However, ensemble-based approaches significantly increase the number of model parameters, resulting in inevitable computational overhead.

Evidential Neural Network. The Evidential Neural Network (ENN) are based on the Dempster-Shafer evidence theory (DST) (Sentz and Ferson 2002), which formulates the learning process as an evidence acquisition process. The ENN acquires support for the evidence distribution from the data. According to the task settings, Deep Evidence Network can be classified into two categories: Dirichlet-based Evidence Network (Dirichlet-based EN) for classification (Sensoy, Kaplan, and Kandemir 2018; Bao, Yu, and Kong 2021; Zhao et al. 2020), and Normal-Inverse Gamma-based Evidence Network (NIG-based EN) for regression (Amini et al. 2020; Pandey and Yu 2022). The Dirichlet-based EN introduces Dirichlet priors on the evidence classification multinomial likelihood to construct the Evidence Deep Learning (EDL) model. This model enables the estimation of both aleatoric and epistemic uncertainties without the need for out-of-distribution (OOD) auxiliary data. Deep Evidence Regression (Amini et al. 2020) is a exemplar for NIG-bsed EN, which introduces the Normal-Inverse Gamma (NIG) evidence prior on the original Gaussian likelihood function to quantify uncertainty of the regression tasks. The NIG evidence prior is considered as a higher-order evidence distribution over unknown lower-order likelihood distributions, from which observed results can be inferred.

Theoretical Analysis of Evidential Models. Despite the popularity of evidence , some studies have raised theoretical shortcomings. According to (Bengs, Hüllermeier, and Waegeman 2022), they argue that classical Evidence Deep Learning (EDL) fails to incentivize learners to faithfully predict their epistemic uncertainty due to its sensitivity to regularization parameters. Addressing above issue, (Bengs, Hüllermeier, and Waegeman 2023) introduces second-order scoring rules to assess the credibility of the cognitive uncertainty in evidence models. Similar issues also exist in evidence regression models, as highlighted by (Meinert, Gawlikowski, and Lavin 2023), which investigates the problem of excessive parameterization in uncertainty representation and explores its unreasonable effectiveness. Meanwhile, regarding the trade-off between model accuracy and uncertainty quantification, (Pandey and Yu 2023b) suggests that the non-negativity constraint on Dirichlet prior param-

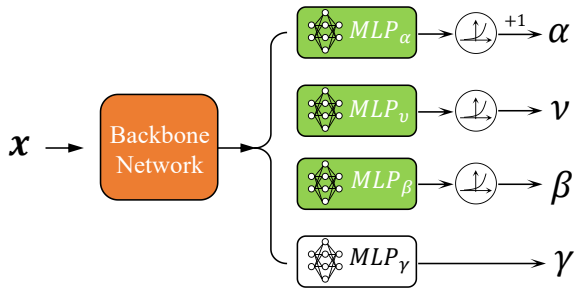

Figure 2: Deep Evidential Regression. Among them, A represents the four parameters of the NIG distribution, three of which, α, β, ν, need to pass through the activation function to ensure the reasonableness of the NIG. It is imperative: $\gamma \in \mathbb{R}, \nu > 0, \alpha > 1, \beta > 0$

eters may lead to poor predictive performance and proposes the concept of "zero-evidence regions" to explain this phenomenon. Unlike the Dirichlet prior, the NIG prior has four parameters, and the impact of non-negativity constraints on performance is more complex, which is also one of the challenges of this paper. On the other hand, (Oh and Shin 2021) proposes that high uncertainty can cause high errors and attempts to alleviate this issue from a multi-task learning perspective. However, it falls short in providing additional insights from the evidence model's perspective.

In this paper, we focus on Deep Evidential Regression (DER) and investigate how the non-negativity prior constraint in NIG hinders model prediction (γ) optimization. Building on theoretical analysis, we propose a novel regularization term to facilitate the broader applicability of Deep Evidential Regression in real-world practical scenarios.

Preliminary

Problem Definition

In this paper, we shall look into supervised regression learning: given a dataset, $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$, we aim to learn a model f with a set of weights, $\boldsymbol{\theta}$, that can be formalised as follows:

$$\arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_i(\boldsymbol{\theta}) \tag{1}$$

where $\mathcal{L}_i(\cdot)$ denotes a loss function, N denotes the dataset size. In this paper, we aim to learn a model to infer $\boldsymbol{\theta}$ that maximize the likelihood of observing our targets value, y, given by $p(y_i|\boldsymbol{\theta})$.

Deep Evidential Regression

The research foundation of this paper builds on Deep Evidential Regression (DER) (Amini et al. 2020). We assume that the target values, y_i, is drawn from a Gaussian distribution and obeys i.i.d, but its variance (σ^2) and mean (μ) are unknown. Our intention is to quantify uncertainty by estimating the variance and mean of the target value. DER model this by placing a prior distribution on (μ, σ^2). Thanks to existing statistical knowledge, the Gaussian prior can be employed as a conjugate prior for the unknown mean, while the Inverse-Gamma prior is for the unknown variance:

$$\begin{aligned} (y_1, \ldots, y_N) &\sim \mathcal{N}(\mu, \sigma^2) \\ \mu \sim \mathcal{N}(\gamma, \sigma^2\nu^{-1}) \qquad & \sigma^2 \sim \Gamma^{-1}(\alpha, \beta). \end{aligned} \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function.

Our intention is to estimate a posterior distribution of variance (σ^2) and mean (μ): $q(\mu, \sigma^2) = p(\mu, \sigma^2|y_1, \ldots, y_N)$. The Normal Inverse-Gamma (NIG) prior can be obtained:

$$\begin{aligned} &p(\underbrace{\mu, \sigma^2}_{\boldsymbol{\theta}} \,|\, \underbrace{\gamma, \nu, \alpha, \beta}_{\boldsymbol{m}}) = \\ &\frac{\beta^\alpha\sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left\{-\frac{2\beta + \nu(\gamma-\mu)^2}{2\sigma^2}\right\} \end{aligned} \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function, note $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta)$, and satisfy $\gamma \in \mathbb{R}, \nu > 0, \alpha > 1, \beta > 0$.

Prediction and Uncertainty Estimation Aleatoric uncertainty, also referred to as data uncertainty, arises due to the complexity inherent in the data itself, such as label noise. Epistemic uncertainty, also known as model uncertainty, emerges as a result of the model's lack of knowledge (Gawlikowski et al. 2021). DER can output four parameters of NIG, $\boldsymbol{m} = (\gamma, \nu, \alpha, \beta)$. Utilizing these parameters, we can compute the prediction, aleatoric, and epistemic uncertainty as

$$\underbrace{E[\mu] = \gamma}_{\text{prediction}}, \quad \underbrace{E[\sigma^2] = \frac{\beta}{\alpha-1}}_{\text{aleatoric}}, \quad \underbrace{Var[\mu] = \frac{E[\sigma^2]}{\nu}}_{\text{epistemic}}. \tag{4}$$

Evidence and Virtual Observation (Amini et al. 2020) define the total evidence: $\boldsymbol{\Phi} = 2\nu + \alpha$, which is based on a heuristic Bayesian interpretation to the NIG prior parameters. (Amini et al. 2020; Jordan 2009; Meinert, Gawlikowski, and Lavin 2023) interprets the parameters of the NIG distribution as count of virtual observation that provide support for the given attributes. For instance, NIG's mean can be intuitively understood as an estimation derived from $\boldsymbol{\nu}$ virtual observation samples, where the sample mean of these virtual observations is $\boldsymbol{\gamma}$. The more such virtual observations are available, the more reliable the estimation of the NIG mean becomes. Following from this interpretation, evidence is composed of virtual observations, and the quantity of virtual observations directly determines the magnitude of the evidence. As a result, the total evidence, $\Phi = 2\nu + \alpha$, holds a physical interpretation, representing the sum of all virtual observation counts.

Learning the evidential distribution

From Bayesian probability theory, Deep Evidential Regression estimates the variance σ^2 and mean μ of the target value y by learning a higher-order evidence distribution, $\boldsymbol{m} = \{\gamma,\ \alpha,\ \nu,\ \beta\}$, which can be expressed as the marginal likelihood:

$$p(y|\mathbf{m}) = \int_{\sigma^2=0}^{\sigma^2=\infty}\int_{\mu=-\infty}^{\mu=\infty} p(y|\mu, \sigma^2)p(\mu, \sigma^2|\mathbf{m})d\mu d\sigma^2 \tag{5}$$

An analytical solution exists for this marginal likelihood:

$$p(y_i|\mathbf{m}) = \mathrm{St}\left(y_i; \gamma, \frac{\beta(1+\upsilon)}{\upsilon\,\alpha}, 2\alpha\right). \tag{6}$$

where $\mathrm{St}\,(y; l, s, n)$ is the Student-t distribution evaluated at y with location l, scale s, and n degrees of freedom. Deep Evidential Regression denote the loss, $L_i^{NLL}(w)$, as the negative logarithm of model evidence:

$$\begin{aligned} L_{NLL}(y, \mathbf{m}) &= \tfrac{1}{2}\log(\tfrac{\pi}{\nu}) - \alpha\log\Lambda \\ &+ (\alpha+\tfrac{1}{2})\log((y-\gamma)^2\nu+\Lambda) + \log(\tfrac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}) \end{aligned} \tag{7}$$

where $\Omega = 2\beta(1+\nu)$. This loss objective can drive the model to output the parameters of the NIG by maximising the evidence to fit the observations.

Theoretical Analysis of Learning Deficiency in Deep Evidential Regression

In this section, we theoretically analyse Deep Evidential Regression, revealing its inherent limitations only sub-optimal performance: with the special activation function, the NIG prior parameter would be **zero**, which triggers Evidence Contraction. And the Evidence Contraction leads to a zero gradient of the NLL loss to prediction and stops the optimisation.

Ensuring Non-negativity of NIG Parameters Triggers Evidence Contraction

Deep Evidential Regression places a higher-order evidence prior, the Normal-Inverse Gaussian (NIG), on the initial likelihood function of the network. Due to strict mathematical definitions, the three parameters of NIG need to satisfy non-negativity: $\{\nu > 0,\, \alpha > 1,\, \beta > 0\}$. In this subsection, we reveal that such non-negativity constraints on the parameters lead to Evidence Contraction.

Definition 1. Evidence Contraction For Deep Evidential Regression, total evidence is comprised of **virtual observations**. When the virtual observations decreases, it causes the total evidence to get smaller. Smaller total evidence implies that the model derives less support for the prediction from the data. We define this phenomenon as **Evidence Contraction**.

Theorem 1. *Given a training sample x, the logits of the Deep Evidential Regression is denoted as $\boldsymbol{o} = \{o_\gamma, o_\alpha, o_\nu, o_\beta\}$, and $\boldsymbol{m} = \{\gamma,\ \alpha,\ \nu,\ \beta\}$ is the final output after activation function. The virtual observation counts are denoted as α, ν, and together they form the total evidence. If the evidence network outputs zero virtual observations and the gradient of the NLL loss with respect to virtual observations is zero, it indicates that Evidence Contraction is occurring.*

Proof. Considering inputs $\boldsymbol{x}$ with target y. Let $\boldsymbol{o} = \{o_\gamma, o_\alpha, o_\nu, o_\beta\}$ represent original logits before activate function, and $\boldsymbol{m} = \{\gamma,\ \alpha,\ \nu,\ \beta\}$ is the final output after activation function, and α, ν represent the virtual observation counts, and $\boldsymbol{\Phi} = 2\nu + \alpha$ denotes total evidence supporting the prediction.

$$\alpha = Act(o_\alpha) + 1,\ \nu = Act(o_\nu),\ \beta = Act(o_\beta) \tag{8}$$

In Deep Evidential Regression, the loss objective is given by Eq. 36. Now, compute gradients of the NLL loss with respect to both α and ν:

$$\frac{\partial L_{NLL}}{\partial\alpha} = \log(1 + \frac{(y-\gamma)^2\nu}{2\beta(\nu+1)}) + \Psi(\alpha) - \Psi(\alpha+\tfrac{1}{2}) \tag{9}$$

$$\frac{\partial L_{NLL}}{\partial\nu} = -\tfrac{1}{2\nu} - \tfrac{\alpha}{\nu+1} + (\alpha+\tfrac{1}{2})\frac{(y-\gamma)^2 + 2\beta}{(y-\gamma)^2\nu + 2\beta(1+\nu)} \tag{10}$$

where $\Psi(\cdot)$ represents the digamma function, $\Psi(x) = \frac{\mathrm{d}\ln\Gamma(x)}{\mathrm{d}x} = \frac{\Gamma'(x)}{\Gamma(x)}$. Next, we take gradients with respect to the original output for o_α, o_ν:

$$\frac{\partial L_{NLL}}{\partial o_\alpha} = \frac{\partial L_{NLL}}{\partial\alpha}\frac{\partial\alpha}{\partial o_\alpha} \tag{11}$$

$$\frac{\partial L_{NLL}}{\partial o_\nu} = \frac{\partial L_{NLL}}{\partial\nu}\frac{\partial\nu}{\partial o_\nu} \tag{12}$$

Consider the activation function to be the softplus function.

$$\alpha = log\,(1 + e^{o_\alpha}) \Rightarrow \frac{\partial\alpha}{\partial o_\alpha} = \frac{1}{1+e^{-o_\alpha}} \tag{13}$$

$$\nu = log\,(1 + e^{o_\nu}) \Rightarrow \frac{\partial\nu}{\partial o_\nu} = \frac{1}{1+e^{-o_\nu}} \tag{14}$$

When $o_\nu (or\ o_\alpha) \to -\infty$, the virtual observation $\nu\ (or\ \alpha)$ are also zero, and the gradient of virtual observation $\nu\ (or\ \alpha)$ becomes zero:

$$\lim_{o_\nu\to-\infty} log(1 + e^{o_\nu}) = 0 \tag{15}$$

$$\lim_{o_\nu\to-\infty} \frac{1}{1+e^{-o_\nu}} = 0 \tag{16}$$

The same result applies to α, and further elaboration is unnecessary. The conclusion is the same when considering the activation function as an exponential function. More details are provided in the Appendix.

Mark: Does zero virtual observation exist ? For a well-designed evidence model, when encountering unbalanced or hard negative samples, the model may fail to acquire sufficient virtual observations from such samples. The model would make unreliable predictions, i.e. the virtual observations is so small. The bound of the case may be unseen samples, and the model is unable to absorb any virtual observations from such samples, i.e. zero virtual observations.

Evidence Contraction Hinders Optimal Performance

Based on the Bayesian interpretation of virtual observation and the theoretical foundation presented earlier, we believe that the virtual observation ν is related to the mean estimation of the NIG prior. Therefore, evidence contraction caused by ν will primarily affect the precision prediction, leading to sub-optimal performance.

Theorem 2. *For Deep Evidential Regression, when evidence contraction occurs, it leads to sub-optimal performance. That is primarily attributed to the virtual observation ν associated with the NIG mean (μ).*

Proof. Calculate the gradient of the NLL loss with respect to ν:

$$\frac{\partial L_{NLL}}{\partial \gamma} = -(2\alpha+1)\frac{(y-\gamma)\nu}{(y-\gamma)^2+2\beta(1+\nu)} \tag{17}$$

Next, we analyze the impact of virtual observations α and ν on the aforementioned gradients:

$$\lim_{\nu\to 0^+} \frac{\partial L_{NLL}}{\partial \gamma} = 0 \tag{18}$$

$$\lim_{\alpha\to 1^+} \frac{\partial L_{NLL}}{\partial \gamma} = -3\cdot\frac{(y-\gamma)\nu}{(y-\gamma)^2+2\beta(1+\nu)} \tag{19}$$

Eq. 18 indicates that as ν tends to zero, the limit of the gradient of the NLL loss with respect to γ degenerates to zero. At this point, evidence model will stop optimization with regard to γ, even for sub-optimal performance. Unlike the behaviour of ν, Eq. 19 shows that there is a non-degenerate relationship between α and the gradient of the NLL loss with respect to γ. Therefore, we can conclude that evidence contraction leads to sub-optimal performance, which is primarily attributed to the virtual observed $\boldsymbol{\nu}$ related to estimation of NIG mean (μ).

Continuing Optimization through Non-saturating Uncertainty Regularization

Due to the intrinsic properties of the NIG prior, it is necessary to satisfy non-negativity, $\{\nu > 0,\ \alpha > 1\ ,\ \beta > 0\}$, when running Deep Evidential Regression. However, under certain conditions, the parameters passed through the activate function tend to approach zero, causing the virtual observed to approach zero, resulting in Evidence Contraction. This implies that the model is no longer deriving knowledge from the data. An inevitable consequence is that the model is underfitting, reducing performance. Furthermore, we reveal that the impact of the virtual observed ν on performance is more severe. Therefore, mitigating the influence of evidence contraction on performance can be done in two ways: preventing evidence contraction and ensuring gradients during evidence contraction.

Theorem 3. *Non-saturating Uncertainty Regularization ensures that there is a gradient to the prediction everywhere in the domain of definition, thus improving performance.*

Proof. We now consider an evidence model with the exponential activation function to transform logits into NIG parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\nu}$. We propose a novel Non-saturating Uncertainty Regularization term:

$$L_U = (y-\gamma)^2\frac{\nu(\alpha-1)}{\beta(\nu+1)} \tag{20}$$

Where $\frac{\nu(\alpha-1)}{\beta(\nu+1)}$ is the inverse of total uncertainty. We calculate the gradient of L_U with respect to γ as:

$$\frac{\partial L_R}{\partial \gamma} = \begin{cases} -\frac{\nu(\alpha-1)}{\beta(\nu+1)} & \text{if } x = y > \gamma \\ \frac{\nu(\alpha-1)}{\beta(\nu+1)} & \text{if } x = y < \gamma \end{cases} \tag{21}$$

L_U freezes the gradient of the total uncertainty, while α, β, ν are still optimized through the NLL to ensure uncertainty quantification capability. At the same time, we prevent degradation to zero by heuristically setting a lower bound on the inverse of the total uncertainty. Therefore, during the training process, the Non-saturated Uncertainty Regularisation can ensure that there is a gradient to the prediction everywhere and optimised in the direction of the correct gradient.

We formulate an overall objective used to train Deep Evidential Regression. In our proposed methodology, the evidential model is trained to maximize the correct evidence and avoid the evidence contraction during training. The overall loss is:

$$L(x,y) = L_{NLL}(x,y) + \eta_1 L_R + \eta_2 L_U \tag{22}$$

Where L_{NLL} is defined by Eq. 36, L_R is the evidence misdirection regularization term proposed by (Amini et al. 2020), and L_U is the non-saturating uncertainty regularization term introduced in this paper to prevent evidence contraction.

Experiments

We first demonstrate the limitations of existing Deep Evidential Regression to confirm our theoretical findings. Then, we evaluated the proposed non-saturating uncertainty regularisation term to show its effectiveness. Finally, we conducted additional empirical analyses to provide more insights about our methodology.

Dataset and Setup We consider the regression problem with UCI regression benchmark [1], Drug-target affinity regression (Shin et al. 2019) and Sentiment Analysis task. Specifically, we use two classical datasets: Davis (Davis et al. 2011) and Kiba (Tang et al. 2014). For the UCI regression baseline dataset, our model setup is kept consistent with existing works (Amini et al. 2020; Oh and Shin 2022). For the Drug-target affinity regression, our experimental setup remained consistent with DeepDTA (Öztürk, Özgür, and Ozkirimli 2018). For the Sentiment Analysis task, we use Stanford Sentiment Treebank (SST-5)[2] and we choose the BERT-based sentiment regression (Munikar, Shakya, and Shrestha 2019) as the backbone.

Evaluation Metric For UCI Benchmark regression datasets, our evaluation metrics include RMSE (Root Mean Squared Error), NLL (Negative Log-Likelihood). For Drug-target affinity regression dataset and Sentiment Analysis task dataset, the MSE (Mean-Square Error), NLL (Negative Log-Likelihood), ECE (Expected Calibration Error) (Guo et al. 2017b), and CI (Concordance Index) (Yu et al. 2011) are

[1] UCI: https://archive.ics.uci.edu/

[2] SST-5: https://nlp.stanford.edu/sentiment/

	RMSE↓					NLL↓				
Dataset	MC dropout	Deep Ensembles	Vanilla DER	MT-DER	Ours	MC dropout	Deep Ensembles	Vanilla DER	MT-DER	Ours
Boston	**2.97 ± 0.19**	3.28 ± 1.00	3.06 ± 0.16	3.04±0.21	**2.67±0.17**	2.46 ± 0.06	2.41 ± 0.25	2.35 ± 0.06	2.31±0.04	**2.31±0.06**
Concrete	**5.23 ± 0.12**	6.03 ± 0.58	5.85 ± 0.15	**5.60±0.17**	5.82±0.21	3.04 ± 0.02	3.06 ± 0.18	**3.01 ± 0.02**	**2.97±0.02**	3.11±0.03
Energy	**1.66 ± 0.04**	2.09 ± 0.29	2.06 ± 0.10	2.04±0.07	**1.83±0.06**	1.99 ± 0.02	1.38 ± 0.22	1.39 ± 0.06	**1.17±0.05**	**1.36±0.03**
Kin8nm	0.10 ± 0.00	0.09 ± 0.00	0.09 ± 0.00	0.08±0.00	**0.06±0.00**	-0.95 ± 0.01	-1.20 ± 0.02	-1.24 ± 0.01	-1.19±0.01	**-1.27±0.02**
Naval	0.01 ± 0.00	**0.00 ± 0.00**	**0.00 ± 0.00**	**0.00±0.00**	**0.00±0.00**	-3.80 ± 0.01	-5.63 ± 0.05	-5.73 ± 0.07	**-5.96±0.03**	**-5.87±0.04**
Power	**4.02 ± 0.04**	4.11 ± 0.17	4.23 ± 0.09	4.03±0.07	**3.02±0.00**	2.80 ± 0.01	2.79 ± 0.04	2.81 ± 0.07	**2.75±0.01**	**2.58±0.01**
Protein	**4.36 ± 0.01**	4.71 ± 0.06	4.64 ± 0.03	4.73±0.07	**4.18±0.02**	2.89 ± 0.00	2.83 ± 0.02	**2.63 ± 0.00**	**2.64±0.01**	2.69±0.05
Wine	**0.62 ± 0.01**	0.64 ± 0.04	0.61 ± 0.02	0.63±0.01	**0.56±0.01**	0.93 ± 0.01	0.94 ± 0.12	**0.89 ± 0.05**	**0.86±0.02**	**0.89±0.08**
Yacht	**1.11 ± 0.09**	1.58 ± 0.48	1.57 ± 0.56	**1.03±0.08**	1.49±0.13	1.55 ± 0.03	1.18 ± 0.21	1.03 ± 0.19	**0.78±0.06**	**0.94±0.15**

Table 1: **UCI Benchmark regression datasets.** The performance on RMSE and NLL. We bold the top two best results, $n = 20$ for sampling baselines. '↓' denotes the lower the better.

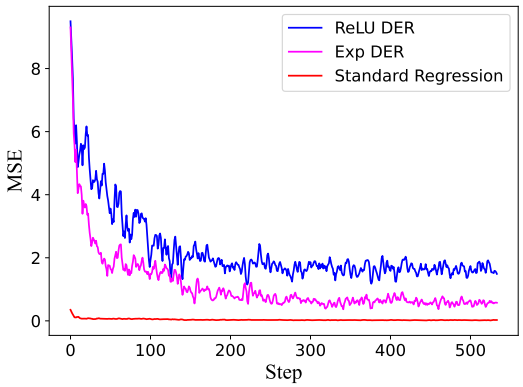

Figure 3: We compare three methods: the ReLU-DER, the Exp-DER and the Standard Regression.The ReLU-DER is the DER that uses ReLU as the activation function, as does the Exp-DER.

adopted, which is aligned with existing paper (Amini et al. 2020; Oh and Shin 2022).

Learning Deficiency of Evidential Models

We conducted an empirical study on a real-world Sentiment Analysis dataset (SST-5), instead of building a toy dataset. Consider the baseline model is (Munikar, Shakya, and Shrestha 2019), and the training loss is the same as the Vanilla DER. As shown in Fig. 3, we compare three methods: ReLU-DER, Exp-DER, and Standard Regression. For the ReLU-DER, when the model's logits is negative, it is compressed directly to zero through ReLU activation, indicating the severe evidence contraction. Although in Exp-DER, evidence contraction also occurs, but compared with ReLU, the function image of Exp is more gentle and evidence contraction is less severe. From Fig. 3, it can be observed that standard regression achieves the best MSE score, followed by Exp-DER, while Exp-DER fares the worst. We ascertain that more serious evidence contraction leads to poorer performance, thus corroborating our theoretical assertion that evidence contraction impedes model performance.

Effectiveness of the Our methods

UCI Regression Benchmark As shown in Table 1, we perform a comparison with Mc dropout (Gal and Ghahramani 2016), Deep Ensembles (Lakshminarayanan, Pritzel, and Blundell 2017), Deep Evidential Regression (Vanilla DER) (Amini et al. 2020) and Multi-task Deep Evidential Regression (MT-DER) (Oh and Shin 2022) on UCI regression benchmark datasets. The experimental setup remains consistent with (Amini et al. 2020; Oh and Shin 2022). Our approach attains the best or competitive RMSE across all datasets and achieves the best NLL on several datasets, thus demonstrating the effectiveness of our proposed method. Compared with Vanilla DER, our method achieves superior RMSE values across all datasets, and better or competitive NLL on all datasets. Even for MT-DER, our method achieves better RMSE and NLL on several datasets. This shows that our method improves the prediction performance while maintaining uncertainty quantification.

	Davis			
	MSE ↓	CI ↑	ECE ↓	NLL ↓
MC dropout	**0.248(0.01)**	**0.884(0.00)**	0.217(0.01)	0.633(0.02)
Vanilla DER	0.275(0.00)	0.856(0.02)	0.184(0.02)	-2.344(0.42)
MT-DER	0.273(0.01)	0.864(0.01)	**0.156(0.03)**	**-2.424(0.07)**
Ours	**0.262(0.01)**	**0.869(0.03)**	**0.141(0.10)**	**-2.373(0.08)**
	Kiba			
	MSE ↓	CI ↑	ECE ↓	NLL ↓
MC dropout	**0.178(0.00)**	0.872(0.00)	0.162(0.01)	0.465(0.01)
Vanilla DER	0.190(0.00)	0.885(0.00)	0.077(0.03)	**-1.544(0.05)**
MT-DER	0.181(0.00)	**0.887(0.00)**	**0.066(0.01)**	-1.433(0.07)
Ours	**0.178(0.01)**	**0.887(0.03)**	**0.054(0.02)**	**-1.448(0.05)**

Table 2: The performance evaluation results on the DTA benchmark datasets.'↑' denotes the higher the better, '↓' denotes the lower the better. We bold the top two best results.

Drug-target affinity regression As shown in Table 2, we evaluated the performance of our method on two datasets from DTA: Davis and Kiba. For Davis, our method achieves better performance on all four metrics compared to Vanilla DER. This again demonstrates that our method can improve the predictive performance of the model while maintaining the original uncertainty quantification capability. And, our method achieved better MSR, CI and ECE scores compared to MT-DER. Despite the decrease in NLL scores, our method is superior as far as model calibration capability (depends on CI) is concerned. For Kiba, we outperform Vanilla DER and MT DER on MSE, CI, and ECE. The NLL scores outperform MT-DER but underperform Vanilla DER, which we attribute to insufficient evidence due to the sparsity of the Kiba dataset.

	SST-5			
	MSE ↓	CI ↑	ECE ↓	NLL ↓
Vanilla DER	**0.5385(0.00)**	**0.7850(0.01)**	**0.1157(0.04)**	**1.1086(0.00)**
MT-DER	0.5466(0.01)	0.7816(0.02)	0.1346(0.01)	1.1519(0.05)
Ours	**0.5281(0.02)**	**0.7877(0.00)**	**0.0738(0.03)**	**1.0612(0.05)**

Table 3: The performance evaluation results on the SST-5 benchmark datasets.'↑' denotes the higher the better, '↓' denotes the lower the better. We bold the top two best results.

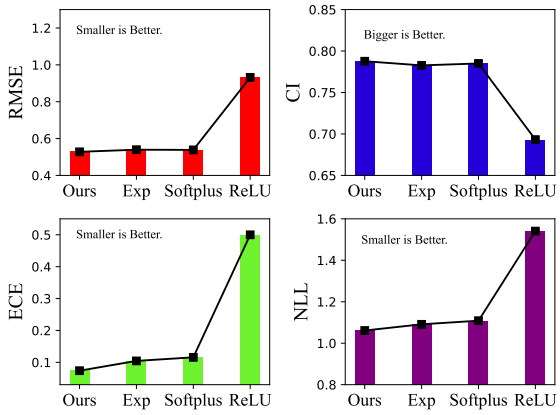

Figure 4: Compared the Vanilla DER with different activation functions (Exp, Softplus and ReLU) and our method to ablate the variables of the activation function. SST-5 is used here.

Sentiment Regression Dataset As shown in Table 3, we conducted experiments on the Sentiment Analysis dataset. The results show that our method outperforms Vanilla DER and MT-DER on all four metrics. It is worth noting that our method improves the predictive performance of the model while simultaneously improving the model's probabilistic calibration ability and uncertainty quantification ability.

Empirical Analyses

Different activation functions. As shown in Fig. 4, we performed ablation experiments on the activation function. The comparison with our method when Vanilla is paired with different activation functions (Exp, Softplus and ReLU). As can be seen from the figure, the Exp activation shows excellent performance because it has lighter evidence contraction, on top of which our method achieves further enhancement.

Visualisation of total evidence and total uncertainty. As shown in Fig. 5, we visualise the trends in total evidence and total uncertainty on the Davis dataset. Total evidence is defined as:

$$\Phi = 2\nu + \alpha \tag{23}$$

And the total uncertainty is defined as:

$$Total\ Evi = Var[\mu] + E[\sigma^2] = \frac{\beta(\nu+1)}{\nu(\alpha-1)} \tag{24}$$

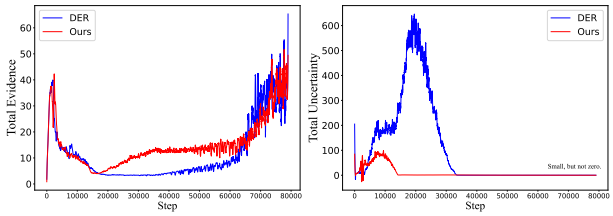

Figure 5: Trends in total evidence and total uncertainty.

The left sub-figure in Fig. 5 shows that during training, our method acquires evidence greater than or equal to Vanilla DER, and at the end of training the two are comparable. This shows that our method increases the model's prediction accuracy without compromising the model's ability to obtain evidence from the data. This conclusion is also demonstrated on the trend of total uncertainty, as shown in the right panel in Fig. 5. As the model iterates, the uncertainty derived by our method drops quickly to a smaller value and stabilises, whereas the Vanilla DER only drops to the same level after almost 10,000 iterations, although the two are comparable at the end of the training, suggesting that our method learns enough evidence faster.

Conclusion

In this paper, we delve into the issue of evidence contraction in deep evidential regression: the non-negativity constraint on normal inverse gamma (NIG) prior parameter leads to evidence contraction under specialized activation functions, thereby impeding the optimization of DER prediction. Building upon this, we have devised an non-saturating uncertainty regularisation term that effectively ensures the progression of accuracy in the right direction, consequently enhancing predictive precision. We conducted extensive experiments on real-world datasets: first, we confirmed the limitations of the existing deep evidence regression; second, we evaluated the proposed non-saturated regularisation term to show its effectiveness; and finally, we conducted additional empirical analyses to reveal more properties of our method. In future research studies, we will design a normalised inverse gamma loss with normalisation properties to ensure the non-saturation property of the evidence regression model.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76: 243–297.

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P. W.; Cao, X.; Khosravi, A.; Acharya, U. R.; Makarenkov, V.; and Nahavandi, S. 2020. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Inf. Fusion*, 76: 243–297.

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33: 14927–14937.

Bao, W.; Yu, Q.; and Kong, Y. 2021. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13349–13358.

Bengs, V.; Hüllermeier, E.; and Waegeman, W. 2022. Pitfalls of epistemic uncertainty quantification through loss minimisation. *Advances in Neural Information Processing Systems*, 35: 29205–29216.

Bengs, V.; Hüllermeier, E.; and Waegeman, W. 2023. On Second-Order Scoring Rules for Epistemic Uncertainty Quantification. In *International Conference on Machine Learning*.

Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; and Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 1–6.

Cai, P.; Wang, H.; Huang, H.; Liu, Y.; and Liu, M. 2021. Vision-Based Autonomous Car Racing Using Deep Imitative Reinforcement Learning. *IEEE Robotics and Automation Letters*, 6(4).

Charpentier, B.; Borchert, O.; Zugner, D.; Geisler, S.; and Günnemann, S. 2022. Natural Posterior Network: Deep Bayesian Predictive Uncertainty for Exponential Family Distributions. In *International Conference on Learning Representations*.

Chen, X.; Bromuri, S.; and van Eekelen, M. 2021. *Neural Machine Translation for Harmonized System Codes Prediction*. Association for Computing Machinery. ISBN 978-1-450-38940-2.

Choi, J.; Chun, D.; Kim, H.; and Lee, H.-J. 2019. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, 502–511.

Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; and Zarrinkar, P. P. 2011. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11): 1046–1051.

Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S. M.; and Weischedel, R. M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, 837–840. Lisbon.

Gal, Y.; and Ghahramani, Z. 2015. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*.

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.

Gawlikowski, J.; Tassi, C. R. N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A. M.; Triebel, R.; Jung, P.; Roscher, R.; Shahzad, M.; Yang, W.; Bamler, R.; and Zhu, X. 2021. A Survey of Uncertainty in Deep Neural Networks. *ArXiv*, abs/2107.03342.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017a. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*.

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

Jordan, M. I. 2009. The exponential family: Conjugate priors.

Karras, C.; Karras, A.; Avlonitis, M.; and Sioutas, S. 2022. An overview of mcmc methods: From theory to applications. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 319–332. Springer.

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

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. arXiv:2304.02643.

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

Li, H.; and Liu, J. 2022. 3D High-Quality Magnetic Resonance Image Restoration in Clinics Using Deep Learning.

Liu, Z.; Amini, A.; Zhu, S.; Karaman, S.; Han, S.; and Rus, D. 2021. Efficient and Robust LiDAR-Based End-to-End Navigation. *IEEE International Conference on Robotics and Automation*.

Ma, H.; Han, Z.; Zhang, C.; Fu, H.; Zhou, J. T.; and Hu, Q. 2021. Trustworthy Multimodal Regression with Mixture of Normal-inverse Gamma Distributions. In *Neural Information Processing Systems*.

Malinin, A.; Chervontsev, S.; Provilkov, I.; and Gales, M. J. F. 2020. Regression Prior Networks. *CoRR*, abs/2006.11590.

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

Meinert, N.; Gawlikowski, J.; and Lavin, A. 2023. The unreasonable effectiveness of deep evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8, 9134–9142.

Munikar, M.; Shakya, S.; and Shrestha, A. 2019. Fine-grained sentiment classification using BERT. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, 1–5. IEEE.

Oh, D.; and Shin, B. 2021. Improving evidential deep learning via multi-task learning. In *AAAI Conference on Artificial Intelligence*.

Oh, D.; and Shin, B. 2022. Improving Evidential Deep Learning via Multi-Task Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7): 7895–7903.

Öztürk, H.; Özgür, A.; and Ozkirimli, E. 2018. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17): i821–i829.

Pandey, D. S.; and Yu, Q. 2022. Evidential Conditional Neural Processes. *arXiv preprint arXiv:2212.00131*.

Pandey, D. S.; and Yu, Q. 2023a. Evidential conditional neural processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 9389–9397.

Pandey, D. S.; and Yu, Q. 2023b. 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.

Petek, K.; Sirohi, K.; Büscher, D.; and Burgard, W. 2022. Robust Monocular Localization in Sparse HD Maps Leveraging Multi-Task Uncertainty Estimation.

Seebock, P.; Orlando, J. I.; Schlegl, T.; Waldstein, S. M.; Bogunovic, H.; Klimscha, S.; Langs, G.; and Schmidt-Erfurth, U. 2020. Exploiting Epistemic Uncertainty of Anatomy Segmentation for Anomaly Detection in Retinal OCT. *IEEE Transactions on Medical Imaging*, 39: 87–98.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 3183–3193.

Sentz, K.; and Ferson, S. 2002. Combination of evidence in Dempster-Shafer theory. *US Department of Energy*.

Shin, B.; Park, S.; Kang, K.; and Ho, J. C. 2019. Self-attention based molecule representation for predicting drug-target interaction. In *Machine Learning for Healthcare Conference*, 230–248. PMLR.

Singh, S. K.; Fowdur, J. S.; Gawlikowski, J.; and Medina, D. 2022. Leveraging Evidential Deep Learning Uncertainties with Graph-based Clustering to Detect Anomalies. *IEEE Transactions on Intelligent Transportation Systems*, 25.

Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; and Coley, C. W. 2021. Evidential Deep Learning for Guided Molecular Property Prediction and Discovery. *ACS Central Science*, 7(8).

Tang, J.; Szwajda, A.; Shakyawar, S.; Xu, T.; Hintsanen, P.; Wennerberg, K.; and Aittokallio, T. 2014. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3): 735–743.

Welling, M.; and Teh, Y. W. 2011. Bayesian learning via stochastic gradient langevin dynamics. *International Conference on Machine Learning*, 28.

Xu, Y.; Shi, B.; Ma, T.; Dong, B.; Zhou, H.; and Zheng, Q. 2023. CLDG: Contrastive Learning on Dynamic Graphs. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 696–707. IEEE.

Yu, C.-N.; Greiner, R.; Lin, H.-C.; and Baracos, V. 2011. Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*, 24: 1845–1853.

Zaidi, S.; Zela, A.; Elsken, T.; Holmes, C. C.; Hutter, F.; and Teh, Y. 2021. Neural ensemble search for uncertainty estimation and dataset shift. *Advances in Neural Information Processing Systems*, 34: 7898–7911.

Zhang, Y.; Li, P.; Sun, M.; and Liu, Y. 2023. Continual Knowledge Distillation for Neural Machine Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7978–7996. Toronto, Canada: Association for Computational Linguistics.

Zhao, X.; Chen, F.; Hu, S.; and Cho, J.-H. 2020. Uncertainty aware semi-supervised learning on graph data. *Advances in Neural Information Processing Systems*, 33: 12827–12836.

Zhou, W.; Liu, F.; and Chen, M. 2021. Contrastive Out-of-Distribution Detection for Pretrained Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1100–1111. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Appendix

More proof details

Evidence Contraction under different Activation Function

Theorem 1 still holds when considering Exponential or ReLU functions. Consider the activation function to be the Exponential function:

$$\alpha = e^{o_\alpha} \Rightarrow \frac{\partial \alpha}{\partial o_\alpha} = e^{o_\alpha} \tag{25}$$

$$\nu = e^{o_\nu} \Rightarrow \frac{\partial \nu}{\partial o_\nu} = e^{o_\nu} \tag{26}$$

When $o_\nu (or\ o_\alpha) \to -\infty$, the virtual observation $\nu\ (or\ \alpha)$ are also zero, and the gradient of virtual observation $\nu\ (or\ \alpha)$ becomes zero:

$$\lim_{o_\nu \to -\infty} e^{o_\nu} = 0 \tag{27}$$

$$\lim_{o_\nu \to -\infty} e^{o_\nu} = 0 \tag{28}$$

Consider the activation function to be the ReLU function:

$$\alpha = ReLU(o_\alpha) \Rightarrow \frac{\partial \alpha}{\partial o_\alpha} = \begin{cases} 1 & if \quad o_\alpha > 0 \\ 0 & otherwise \end{cases} \tag{29}$$

$$\nu = ReLU(o_\nu) \Rightarrow \frac{\partial \nu}{\partial o_\nu} = \begin{cases} 1 & if \quad o_\nu > 0 \\ 0 & otherwise \end{cases} \tag{30}$$

When $o_\nu (or\ o_\alpha) < 0$, the virtual observation $\nu\ (or\ \alpha)$ are also zero, and the gradient of virtual observation $\nu\ (or\ \alpha)$ becomes zero.

When the activation function is considered to be Exponential or ReLU, evidence contraction still occurs, i.e., Theorem 1 still holds.

Does the limit exist in Eq. 18?

This part proves that the limit in Eq. 18 exists. Here proof refers to (Oh and Shin 2022). The gradient of the NLL loss with respect to ν:

$$\frac{\partial L_{NLL}}{\partial \gamma} = -(2\alpha + 1)\frac{(y-\gamma)\nu}{(y-\gamma)^2 + 2\beta(1+\nu)} \tag{31}$$

Let $\lambda = \gamma - y$:

$$\frac{\partial L_{NLL}}{\partial \gamma} = -(2\alpha + 1)\frac{\lambda\nu}{\lambda^2 + 2\beta(1+\nu)} \tag{32}$$

(Oh and Shin 2022) show that if for every $\epsilon > 0$, there exists a $\delta > 0$ such that, for all $\nu > 0$ if $0 < \nu < \delta$, then $|(2\alpha + 1)\frac{\lambda\nu}{\lambda^2\nu + 2\beta(1+\nu)}| < \epsilon$.

$$|\frac{\lambda\nu}{\lambda^2\nu + 2\beta(1+\nu)}| < |(2\alpha+1)\frac{\lambda\nu}{\lambda^2\nu + 2\beta(1+\nu)}| < \epsilon$$

$$(\text{since } \lambda^2\nu + 2\beta(1+\nu) > 0)$$

$$|\lambda|\nu = |\lambda\nu| < \epsilon(\lambda^2\nu + 2\beta(1+\nu))$$

$$(|\lambda| - \epsilon\lambda^2 - 2\beta\epsilon)\nu < 2\epsilon\beta$$

If $(|\lambda| - \epsilon\lambda^2 - 2\beta\epsilon) < 0$, the proposition is always true regardless of δ since $\beta > 0$ in NIG. Else:

$$\nu < \frac{2\epsilon\beta}{(|\lambda| - \epsilon\lambda^2 - 2\beta\epsilon)}$$

Therefore, if one set $\delta \leq \frac{2\epsilon\beta}{(|\lambda| - \epsilon\lambda^2 - 2\beta\epsilon)}$, then there exists a limit in the neighbourhood $[\nu - \delta, \nu + \delta]$ of ν as:

$$\lim_{\nu \to 0^+} \frac{\partial L_{NLL}}{\partial \gamma} = 0 \tag{33}$$

Comparison of smoothness between Exp and ReLU

In Fig.3, we compare three methods: the ReLU-DER, the Exp-DER and the Standard Regression. The ReLU-DER is the DER that uses ReLU as the activation function, as does the Exp-DER. We find that more severe evidence contraction leads to worse performance from a causal inference perspective, thus confirming our theoretical assertion that evidence contraction hinders model performance. However, the above conclusion is based on the assumption that, compared with ReLU, the function curve of Exponential function is more gentle and evidence contraction is less severe. We will demonstrate that Exponential function is more gentle than ReLU function.

Consider the loss $\mathcal{L}$ is used to train the DER (Deep Evidential Regression), let the logits of the DER is denoted as $\boldsymbol{o} = \{o_\gamma, o_\alpha, o_\nu, o_\beta\}$ applying the activation $\mathcal{A}$, and let $\{\alpha,\ \nu\}$ as evidential output.

For **ReLU**:

$$\frac{\partial \mathcal{L}_1}{\partial w} = \frac{\partial \mathcal{L}_1}{\partial e}\frac{\partial e}{\partial o}\frac{\partial o}{\partial w} = 0 \tag{34}$$

For **Exp**:

$$\begin{aligned} &\frac{\partial \mathcal{L}_2}{\partial w} = \frac{\partial \mathcal{L}_2}{\partial e}\frac{\partial o}{\partial w}\exp(o) = \\ &\frac{\partial \mathcal{L}_2}{\partial e}\frac{\partial o}{\partial w}\{[1 + \exp(o)]\text{Sigmoid}(o)\} \end{aligned} \tag{35}$$

Where the w denotes as model weight.

We can conclude: $\frac{\partial \mathcal{L}_2}{\partial w} \geq \frac{\partial \mathcal{L}_1}{\partial w}$. Compared to the ReLU activation function, this suggests that the Exponential activation function declines more slowly and more gently, with less evidence of contraction.

Details of Experiment

Dateset

We consider the regression problem with UCI regression benchmark, Drug-target affinity regression (Shin et al. 2019) and Sentiment Analysis task. Those datasets are all public datasets and can be obtained publicly.

UCI regression benchmark Datasets for regression sourced from the UCI machine learning repository have been curated for benchmarking purposes, complete with test-train divisions. For the UCI regression baseline dataset, our model setup is kept consistent with existing works (Amini et al. 2020; Oh and Shin 2022), so we report their scores.

Drug-target affinity regression For the Drug-target affinity regression, our experimental setup remained consistent with DeepDTA (Öztürk, Özgür, and Ozkirimli 2018). Specifically, we use two classical datasets: Davis (Davis et al. 2011) and Kiba (Tang et al. 2014). The Davis dataset consists of clinical kinase inhibitor ligands and relevant dissociation constant (Kd) values(Affinity values), and it dataset contains 68 compounds and 442 proteins, for a total of 30,056 affinity instance. The Kiba is a similar dataset, containing 2111 compounds, 229 in proteins, and a total of 118,254 in affinity examples. We kept our experimental setup consistent with (Oh and Shin 2022) and therefore reported the scores from their paper.

Stanford Sentiment Treebank This dataset consists of 11,855 single sentences extracted from movie reviews. Each phrase is labeled as negative, somewhat negative, neutral, somewhat positive or positive. The corpus with all 5 labels is called SST-5, and the SST-5 used in this paper. In this paper, we choose the BERT-based sentiment regression (Munikar, Shakya, and Shrestha 2019) as the backbone. We ran Vinalla DER and MT-DER on SST-5, and the scores reported are the average of 5 experiments.

Details of Evaluation Metrics

Our evaluation metrics are the RMSE, MSE, NLL, ECE (Expected Calibration Error) and CI (Concordance Index). Among them, RMSE and MSE are well known.

Negative Log Likelihood This metrics is the negative logarithm of the likelihood function and can be used to measure the model's ability to quantify uncertainty. In the DER model, the NLL score is defined as follows:

$$\begin{aligned} L_{NLL}(y, \mathbf{m}) &= \tfrac{1}{2}\log(\tfrac{\pi}{\nu}) - \alpha \log \Lambda \\ &+ (\alpha + \tfrac{1}{2})\log((y-\gamma)^2\nu + \Lambda) + \log(\tfrac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}) \end{aligned} \tag{36}$$

Concordance Index CI can be used to compare the performance of these models. A higher C-index value means that the model is more accurate in sample ordering and prediction(Yu et al. 2011). The CI is defined as:

$$CI = \frac{1}{N}\sum_{y_i > y_j} h(\hat{y_i} > \hat{y_j}), \quad h(x) = \begin{cases} 1, & x > 0 \\ 0.5, & x = 0 \\ 0, & \text{else} \end{cases} \tag{37}$$

where N is the total number of data pairs; y_i is target value, and $\hat{y_i}$ is a predicted value.

Expected Calibration Error ECE is a metric used to assess the calibration of classification models. Calibration is the degree of agreement between the confidence (probability value) of the model's predictions and the actual observed accuracy. The ECE can be calculated as:

$$ECE = \frac{1}{N}\sum_{i=1}^{N} |acc(P_i) - P_i| \tag{38}$$

Method	AUROC ↑	AUPRC ↑	FAR95 ↓
Vanilla DER	**0.7896**	0.5513	**0.4457**
Ours	0.7839	**0.5554**	0.4615

Table 4: OOD detection performance on on SST (as ID) and ACE (as OOD).

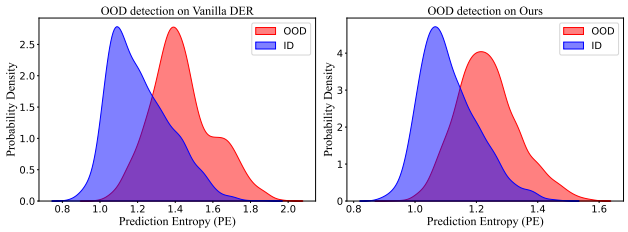

Figure 6: Visualization of OOD detection.

Where $acc(P_i)$ is accuracy using P_i confidence interval (e.g. $P_i = 0.95$) and N is the number of intervals. We use $P_{[0:N]} = \{0.01, 0.02 \ldots, 1.00\}, N = 100$.

Supplementary Experiments

Ou-of-distribution detection We conducted OOD detection experiments on SST (as ID) and ACE (Doddington et al. 2004) (as OOD), as shown in Table 4. We consider OOD detection as a binary classification problem where OOD samples are labelled as true positive samples. The uncertainty prediction entropy (*PE*) is adopted as the scoring function, see in Eq.39. Consistent with existing work (Zhou, Liu, and Chen 2021), evaluation metrics are: Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC) and False Acceptance Rate at 95% specificit (FAR95).

$$PE(\sigma^2) = \frac{1}{2}\log\left(2\pi e^{\sigma^2}\right) \tag{39}$$

Compared to vanilla DER, our results are competitive on AUROC (78.39% vs 78.96%) and AUPRC (55.54% vs 55.13%), where the former is ours. The results show that our method improves DER's accuracy while preserving its original ability to detect OOD sample. In addition, we provide visualisation of OOD detection performance in Fig. 6. It can be seen that the effects of vanilla DER and Ours are comparable, which is consistent with the Table 4.