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

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

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

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:

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

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).

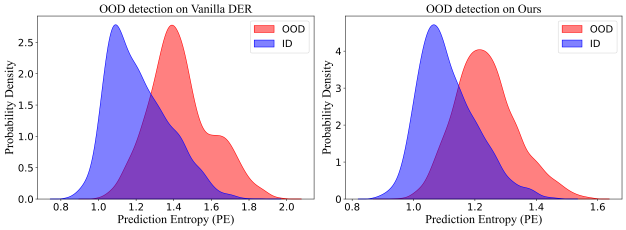

Figure 6: Compared the Vanilla DER with different activation functions (Exp, Softplus and ReLU) and our method to ablate the variables of the activation function. SST-5 is used here.

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.